SYSTEM IDENTIFICATION

IEEE Press
445 Hoes Lane, P.O. Box 1331
Piscataway, NJ 08855-1331

Kenneth Moore, *Director of IEEE Press*
Catherine Faduska, *Senior Acquisitions Editor*
John Griffin, *Acquisitions Editor*
Robert H. Bedford, *Assistant Acquisitions Editor*
Anthony VenGraitis, *Project Editor*

Cover design: William T. Donnelly, *WT Design*

Technical Reviewers

Keith Godfrey, *University of Warwick, Coventry, UK*
Brett Ninness, *University of Newcastle, Australia*

SYSTEM IDENTIFICATION

A Frequency Domain Approach

Rik Pintelon
Department of Electrical Engineering
Vrije Universiteit Brussel
BELGIUM

Johan Schoukens
Department of Electrical Engineering
Vrije Universiteit Brussel
BELGIUM

IEEE
PRESS

The Institute of Electrical and Electronics Engineers, Inc. New York

This book and other books may be purchased at a discount from the publisher when ordered in bulk quantities. Contact:

IEEE Press Marketing
Attn: Special Sales
445 Hoes Lane
P.O. Box 1331
Piscataway, NJ 08855-1331
Fax: +1 732 981 9334

For more information about IEEE Press products, visit the IEEE Online Catalog & Store: http://www.ieee.org/store.

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

ISBN 0-7803-6000-1
IEEE Order No. PC5832

Library of Congress Cataloging-in-Publication Data

Pintelon, R. (Rik)
System identification: a frequency domain approach / Rik Pintelon, Johan Schoukens.
p. cm.
ISBN 0-7803-6000-1
1. System identification. I. Schoukens, J. (Johan) II. Title.

QA402.P56 2001
003'.85–dc21

2001016782

To all people of goodwill

Rik

To Annick, Ine, Maarten, and Sanne

Johan

Contents

CHAPTER 6 An Intuitive Introduction to Frequency Domain Identification 177

CHAPTER 7 Estimation with Known Noise Model 183

CHAPTER 17 Identification of Semilinear Models 535

Preface

Identification is a powerful technique for building accurate models of complex systems from noisy data. It consists of three basic steps, which are interrelated: (1) the design of an experiment; (2) the construction of a model, black box or from physical laws; and (3) the estimation of the model parameters from the measurements. The art of modeling lies in proper use of the skills and specialized knowledge of experts in the field of study, who decide what approximations can be made, suggest how to manipulate the system, reveal the important aspects, and so on. Consequently, modeling should preferably be executed by these experts themselves. Naturally, they require relevant tools for extracting information of interest. However, most experts will not be familiar with identification theory and will struggle in each new situation with the same difficulties while developing their own identification techniques, losing time over problems already solved in the literature of identification.

This book presents a thorough description of methods to model linear dynamic time-invariant systems by their transfer function. The relations between the transfer function and the physical parameters of the system are very dependent upon the specific problem. Because transfer function models are generally valid, we have restricted the scope of the book to these alone, so as to develop and study general purpose identification techniques. This should not be unnecessarily restricting for readers who are more interested in the physical parameters of a system: the transfer function still contains all the information that is available in the measurements, and it can be considered to be an intermediate model between the measurements and the physical parameters. Also, the transfer function model is very suitable for those readers looking for a black box description of the input-output relations of a system. And, of course, the model is directly applicable to predict the output of the system.

In this book, we use mainly frequency domain representations of the data. In combination with periodic excitations, this opens many possibilities to identify continuous-time (Laplace-domain) or discrete-time (z-domain) models, if necessary extended with an arbitrary and unknown delay. Although we strongly advocate using periodic excitations, we also extend the methods and models to deal with arbitrary excitations. The "classical" time-domain identification methods that are specifically directed toward these signals are briefly covered and encapsulated in the identification framework that we offer to the reader.

This book provides answers to questions at different levels, such as: What is identification and why do I need it? How to measure the frequency response function of a linear dynamic system? How to identify a dynamic system? All these are very basic questions, directly focused on the interests of the practitioner. Especially for these readers, we have added guidelines to many chapters for the user, giving explicit and clear advice on what are good choices in order to attain a sound solution. Another important part of the material is intended for readers who want to study identification techniques at a more profound level. Questions on how to analyze and prove the properties of an identification scheme are addressed in this part. This study is not restricted to the identification of linear dynamic systems; it is valid for a very wide class of weighted, nonlinear least squares estimators. As such, this book provides a great deal of information for readers who want to set up their own identification scheme to solve their specific problem.

The structure of the book can be split into four parts: (1) collection of raw data or nonparametric identification; (2) parametric identification; (3) comparison with existing frameworks, guidelines, and illustrations; (4) profound development of theoretical tools.

In the first part, after the introductory chapter on identification, we discuss the collection of the raw data: How to measure a frequency response function of a system. What is the impact of nonlinear distortions? How to recognize, qualify, and quantify nonlinear distortions. How to select the excitation signals in order to get the best measurements. This nonparametric approach to identification is discussed in detail in Chapters 2, 3, and 4.

In the second part, we focus on the identification of parametric models. Signal and system models are presented, using a frequency and a time domain representation. The equivalence and impact of leakage effects and initial conditions are shown. Nonparametric and parametric noise models are introduced. The estimation of the parameters in these models is studied in detail. Weighted (nonlinear) least squares methods, maximum likelihood, and subspace methods are discussed and analyzed. First, we assume that the disturbing noise model is known; next, the methods are extended to the more realistic situation of unknown noise models that have to be extracted from the data, together with the system model. Special attention is paid to the numerical conditioning of the sets of equations to be solved. Taking some precautions, very high order systems, with 100 poles and zeros or even more, can be identified. Finally, validation tools to verify the quality of the models are explained. The presence of unmodeled dynamics or nonlinear distortions is detected, and simple rules to guide even the inexperienced user to a good solution are given. This material is presented in Chapters 5 to 9.

The third part begins with an extensive comparison of what is classically called *time and frequency domain identification*. It is shown that, basically, both approaches are equivalent, but some questions are more naturally answered in one domain instead of the other. The most important question is periodic excitations versus nonperiodic or arbitrary excitations. Next, we provide the practitioner with detailed guidelines to help avoid pitfalls from the very beginning of the process (collecting the raw data), over the selection of appropriate identification methods until the model validation. Finally, we illustrate many of the developed ideas in a wide variety of examples from different fields. This part covers Chapters 10, 11, and 12.

The last part of the book is intended for readers who want to acquire a thorough understanding of the material or those who want to develop their own identification scheme. Not only do we give an introduction to the stochastic concepts we use, but we also show, in a structured approach, how to prove the properties of an estimator. This avoids the need for each freshman in this field to find out, time and again, the basic steps to solve such a problem. Starting from this background, a general but detailed framework is set up to analyze the properties of nonlinear least squares estimators with deterministic and stochastic weighting.

For the special and quite important class of semilinear models, it is possible to make this analysis in much more detail. This material is covered in Chapters 13 to 18.

It is possible to extract a number of undergraduate courses from this book. In most of the chapters that can be used in these courses, we added exercises that introduce the students to the typical problems that appear when applying the methods to solve practical problems.

A first, quite general undergraduate course subject is the measurement of frequency response functions of dynamic systems, as discussed in Chapters 2 to 4.

A second possibility is a first introduction to the identification of linear dynamic systems. Such an undergraduate course should include Chapter 1 and some selected parts of Chapters 5, 6, 7, 8, and 9.

A last course, at the graduate level, is an advanced course on identification based on the methods that are explained in Chapters 15, 16, 17, and 18. This gives an excellent introduction for students who want to develop their own algorithms.

A complete MATLAB® toolbox, which includes the techniques developed in this book, is available. It can be used with a graphical user interface, avoiding most problems and nasty questions for the inexperienced user. At the basic level, this toolbox produces almost autonomously a good model. At the intermediate or advanced level, the user obtains access to some of the parameters in order to optimize the operation of the toolbox to solve dedicated modeling problem. Finally, for those who want to use it as a research tool, there is also a command level that gives full access to all the parameters that can be set to optimize and influence the behavior of the algorithms. More information on this package can be obtained by sending an E-mail to one of the authors: rik.pintelon@vub.ac.be or johan.schoukens@vub.ac.be

Rik Pintelon
Department of Electrical Engineering
Vrije Universiteit Brussel
BELGIUM

Johan Schoukens
Department of Electrical Engineering
Vrije Universiteit Brussel
BELGIUM

Acknowledgments

This book is the culmination of 20 years of research in the identification group of the Department of Fundamental Electricity and Instrumentation (ELEC) of the Vrije Universiteit Brussel (VUB). It is the result of close and harmonious cooperation between many of the workers in the department, and we would like to thank all of them for exchanges of ideas and important discussions that have taken place over the gestation period of this, the end result. We are greatly indebted to ELEC, which is directed by Professor Alain Barel, to the R&D Department of the VUB, to the FWO-Vlaanderen (the Fund for Scientific Research in Flanders), and to the Flemish and the Federal Government (GOA and IUAP research programs). Without their sustained support, this work would never have seen the light of day.

We are grateful to many colleagues and coworkers in the field who imparted new ideas and methods to us, took time for discussions, and were prepared to listen to some of our less conventional ideas that did not fit in the mainstream of the "classical" identification. Their critical remarks and constructive suggestions contributed significantly to our view on the field.

Last, but not least, we want to thank Yves Rolain for contributing to our work as a friend, a colleague, and a coauthor of many of our papers. Without his sustained support, we would never have been able to access the advanced measurement equipment we used for the many experiments that are reported in this book.

Rik Pintelon

Johan Schoukens

List of Operators and Notational Conventions

$\mathbb{A}$	outline uppercase font denotes a set, for example, $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$ and $\mathbb{C}$ are, respectively, the natural, the integer, the rational, the real, and the complex numbers
$\otimes$	the Kronecker matrix product
$*$	convolution operator
Re( )	real part of
Im( )	imaginary part of
$\arg\min\limits_{x} f(x)$	the minimizing argument of $f(x)$
$O(x)$	an arbitrary function with the property $\lim\limits_{x \to 0} \lvert O(x)/x \rvert < \infty$
$o(x)$	an arbitrary function with the property $\lim\limits_{x \to 0} \lvert o(x)/x \rvert = 0$
$\hat{\theta}$	estimated value of θ
$\bar{x}$	complex conjugate of x
subscript 0	true value
subscript Re	$A_{\mathrm{Re}} = \begin{bmatrix} \mathrm{Re}(A) & -\mathrm{Im}(A) \\ \mathrm{Im}(A) & \mathrm{Re}(A) \end{bmatrix}$
subscript re	$A_{\mathrm{re}} = \begin{bmatrix} \mathrm{Re}(A) \\ \mathrm{Im}(A) \end{bmatrix}$
subscript u	with respect to the input of the system
subscript y	with respect to the output of the system
subscript $*$	limiting estimate
superscript T	matrix transpose

superscript $-T$	transpose of the inverse matrix
superscript H	Hermitian transpose: complex conjugate transpose of a matrix
superscript $-H$	Hermitian transpose of the inverse matrix
superscript +	Moore-Penrose pseudoinverse
superscript $\perp$	orthogonal complement of a subspace or a matrix
$\angle x$	phase (argument) of the complex number x
$X_{[i]}(s)$	ith entry of the vector function $X(s)$
$A_{[i,j]}(s)$	i, jth entry of the matrix function $A(s)$
$A_{[:,j]}$	jth column of A
$A_{[i,:]}$	ith row of A
$X^{[k]}(s)$	kth realization of a random process $X(s)$
$\lambda(A)$	eigenvalue of a square matrix A
$\sigma(A)$	singular value of an $n \times m$ matrix A
$\kappa(A) = (\max_i \sigma_i(A))/(\min_i \sigma_i(A))$	condition number of an $n \times m$ matrix A
$\lvert x \rvert = \sqrt{(\mathrm{Re}(x))^2 + (\mathrm{Im}(x))^2}$	magnitude of a complex number x
$\lVert A \rVert_1 = \max_{1 \le j \le m} \sum_{i=1}^{n} \lvert A_{[i,j]} \rvert$	1-norm of an $n \times m$ matrix A
$\lVert A \rVert_2 = \max_{1 \le i \le m} \sigma_i(A)$	2-norm of an $n \times m$ $(n \ge m)$ matrix A
$\lVert X \rVert_2 = \sqrt{X^H X}$	2-norm of the column vector X
$\lVert A \rVert_\infty = \max_{1 \le i \le n} \sum_{j=1}^{m} \lvert A_{[i,j]} \rvert$	∞-norm of an $n \times m$ matrix A
$\lVert A \rVert_F = \sqrt{\mathrm{tr}(A^H A)}$	Frobenius norm of an $n \times m$ matrix A
$\mathrm{diag}(A_1, A_2, \ldots, A_K)$	block diagonal matrix with blocks A_k, $k = 1, 2, \ldots, K$
$\mathrm{herm}(A) = (A + A^H)/2$	Hermitian symmetric part of an $n \times m$ matrix A
$\mathrm{null}(A)$	null space of the $n \times m$ matrix A, linear subspace of $\mathbb{C}^m$ defined by $Ax = 0$
$\mathrm{range}(A)$	range of the $n \times m$ matrix A, linear subspace of $\mathbb{C}^n$ that is reachable by making linear combinations of the columns of A $(\mathrm{range}(A) = (\mathrm{null}(A^T))^\perp)$
$\mathrm{rank}(A)$	rank of the $n \times m$ matrix A, maximum number of linear independent rows (columns) of A

$\mathrm{span}\{a_1, a_2, \ldots, a_m\}$	the span of the vectors $a_1, a_2, \ldots, a_m$ is the linear subspace obtained by making all possible linear combinations of $a_1, a_2, \ldots, a_m$
$\mathrm{tr}(A) = \sum_{i=1}^{n} A_{[i,i]}$	trace of an $n \times n$ matrix A
$\mathrm{vec}(A)$	a column vector formed by stacking the columns of the matrix A on top of each other
a.s.lim	almost sure limit, limit with probability one
l.i.m.	limit in mean square
plim	limit in probability
Lim	limit in distribution
$\mathcal{E}\{\ \}$	mathematical expectation
Prob( )	probability
$b_X = X - \mathcal{E}\{X\}$	bias of the estimate X
$\mathrm{Cov}(X, Y) = \mathcal{E}\{(X - \mathcal{E}\{X\})(Y - \mathcal{E}\{Y\})^H\}$	cross-covariance matrix of X and Y
$\mathrm{covar}(x, y) = \mathcal{E}\{(x - \mathcal{E}\{x\})\overline{(y - \mathcal{E}\{y\})}\}$	covariance of x and y
cum( )	cumulant
$\mathrm{var}(x) = \mathcal{E}\{\lvert x - \mathcal{E}\{x\}\rvert^2\}$	variance of x
$C_X = \mathrm{Cov}(X) = \mathrm{Cov}(X, X)$	covariance matrix of X
$\hat{C}_X = \frac{1}{M-1}\sum_{m=1}^{M}(X^{[m]} - \hat{X})(X^{[m]} - \hat{X})^H$	sample covariance matrix of M realizations of X
$C_{XY} = \mathrm{Cov}(X, Y)$	cross-covariance matrix of X and Y
$\hat{C}_{XY} = \frac{1}{M-1}\sum_{m=1}^{M}(X^{[m]} - \hat{X})(Y^{[m]} - \hat{Y})^H$	sample cross-covariance matrix of M realizations of X and Y
$CR(X)$	Cramér-Rao lower bound on X
$\mathrm{DFT}(x(t))$	discrete Fourier transform of the samples $x(t),\ t = 0, 1, \ldots, N-1$
$Fi(X)$	Fisher information matrix with respect to the parameters X
I_m	$m \times m$ identity matrix
q	backward shift operator: $qu(kT_s) = u((k-1)T_s)$
$\mathrm{MSE}(X) = \mathcal{E}\{(X - X_0)(X - X_0)^H\}$	mean square error of the estimate X
$R_{xx}(\tau) = \mathcal{E}\{x(t)x^H(t-\tau)\}$	autocorrelation of $x(t)$
$R_{xy}(\tau) = \mathcal{E}\{x(t)y^H(t-\tau)\}$	cross-correlation of $x(t)$ and $y(t)$
$S_{XX}(j\omega)$	Fourier transform of $R_{xx}(\tau)$ (autopower spectrum of $x(t)$)
$S_{XY}(j\omega)$	Fourier transform of $R_{xy}(\tau)$ (cross-power spectrum of $x(t)$ and $y(t)$)

$\hat{X} = \frac{1}{M}\sum_{m=1}^{M} X^{[m]}$	sample mean of M realizations (experiments) of X
$\mu_x = \mathcal{E}\{x\}$	mean value of x
$\sigma_x^2 = \mathrm{var}(x)$	variance of the x
$\hat{\sigma}_x^2 = \frac{1}{M-1}\sum_{m=1}^{M} \lvert x^{[m]} - \hat{x}\rvert^2$	sample variance of M realizations of x
$\sigma_{xy}^2 = \mathrm{covar}(x, y)$	covariance of x and y
$\hat{\sigma}_{xy}^2 = \frac{1}{M-1}\sum_{m=1}^{M} (x^{[m]} - \hat{x})(\overline{y^{[m]} - \hat{y}})$	sample covariance of M realizations of x and y

List of Symbols

$A(\Omega, \theta) = \sum_{r=0}^{n_p} a_r p_r(\Omega)$	denominator polynomial plant model expanded in the polynomial basis $p_r(\Omega)$
$A(\Omega, \theta) = \sum_{r=0}^{n_a} a_r \Omega^r$	denominator polynomial plant model
$B(\Omega, \theta) = \sum_{r=0}^{n_q} b_r q_r(\Omega)$	numerator plant model expanded in the polynomial basis $q_r(\Omega)$
$B(\Omega, \theta) = \sum_{r=0}^{n_b} b_r \Omega^r$	numerator polynomial plant model
$C(z^{-1}, \theta) = \sum_{r=0}^{n_c} c_r z^{-r}$	numerator polynomial noise model
$D(z^{-1}, \theta) = \sum_{k=0}^{n_d} d_k z^{-k}$	denominator polynomial noise model
$e(t)$	white noise at time t
$E(k)$	discrete Fourier transform of the samples $e(tT_s),\ t = 0, 1, \ldots, N-1$
f	frequency
F	number of frequency domain data samples
f_s	sampling frequency
$G(j\omega)$	frequency response function
$G_R(j\omega)$	best linear approximation of a nonlinear plant
$G(\Omega, \theta) = B(\Omega, \theta)/A(\Omega, \theta)$	parametric plant model
$H(z^{-1}, \theta) = C(z^{-1}, \theta)/D(z^{-1}, \theta)$	parametric noise model
$I(\Omega, \theta) = \sum_{r=0}^{n_i} i_r \Omega^r$	polynomial of the initial and the final conditions of the plant model $B(\Omega, \theta)/A(\Omega, \theta)$
j	$j^2 = -1$

$J(z^{-1}, \theta) = \sum_{r=0}^{n_d} j_r z^{-r}$	polynomial of the initial and the final conditions of the noise model $C(z^{-1}, \theta)/D(z^{-1}, \theta)$
M	number of (repeated) experiments
N	number of time domain data samples
n_a, n_b, n_c, n_d, n_i and n_j	order of the polynomials $A(\Omega, \theta)$, $B(\Omega, \theta)$, $C(z^{-1}, \theta)$, $D(z^{-1}, \theta)$, $I(\Omega, \theta)$, and $J(z^{-1}, \theta)$
n_θ	dimension of the parameter vector θ
$n_u(t)$, $n_y(t)$	disturbing time domain noise on the input $u(t)$ and output $y(t)$ signals, respectively
$N_U(k)$, $N_Y(k)$	discrete Fourier transform of the samples $n_u(tT_s)$ and $n_y(tT_s)$, $k = 0, 1, \ldots, N-1$, respectively
s	Laplace transform variable
s_k	Laplace transform variable evaluated along the imaginary axis at DFT frequency k: $s_k = j\omega_k$
$T(\Omega, \theta) = I(\Omega, \theta)/A(\Omega, \theta)$	parametric transient model of the plant $B(\Omega, \theta)/A(\Omega, \theta)$
t	continuous or discrete time variable
T_s	sampling period
$U(e^{j\omega T_s})$, $Y(e^{j\omega T_s})$	Fourier transform of $u(tT_s)$ and $y(tT_s)$
$U(k)$, $Y(k)$	discrete Fourier transform of the samples $u(tT_s)$ and $y(tT_s)$, $t = 0, 1, \ldots, N-1$
U_k, Y_k	Fourier coefficients of the periodic signals $u(t)$, $y(t)$
$U(j\omega)$, $Y(j\omega)$	Fourier transform of $u(t)$ and $y(t)$
$U(s)$, $Y(s)$	one-sided Laplace transform of $u(t)$ and $y(t)$
$u(t)$, $y(t)$	input and output time signals
$U(z)$, $Y(z)$	one-sided Z-transform of $u(tT_s)$, $y(tT_s)$
$V_*(\theta)$	asymptotic ($F \to \infty$) cost function
$V_F(\theta, z)$	cost function based on F measurements
$V_F'(\theta, z)$	derivative cost function w.r.t. θ (dimension $1 \times n_\theta$)
$V_F''(\theta, z)$	second-order derivative (Hessian) cost function w.r.t. θ (dimension $n_\theta \times n_\theta$)
$Z(k) = [Y(k)U(k)]^T$	data vector containing the measured input and output (DFT) spectra at (DFT) frequency k
$Z = [Z^T(1)Z^T(2)\ldots Z^T(F)]^T$	data vector containing the measured input and output DFT spectra (dimension $2F$)

z	Z-transform variable
z_k	Z-transform variable evaluated along the unit circle at DFT frequency k: $z_k = e^{j\omega_k T_s} = e^{j2\pi k/N}$
$\varepsilon(\theta, Z)$	column vector of the (weighted) model residuals (dimension F)
θ	column vector of the model parameters
$\tilde{\theta}(Z_0)$	minimizing argument of the cost function $V_F(\theta)$
$\hat{\theta}(Z)$	estimated model parameters, minimizing argument of the cost function $V_F(\theta, Z)$
$\hat{\underline{\theta}}(Z)$	truncated estimator
$\sigma_U^2(k) = \mathrm{var}(U(k))$	variance of the measured input DFT spectrum
$\sigma_Y^2(k) = \mathrm{var}(Y(k))$	variance of the measured output DFT spectrum
$\sigma_{YU}^2(k) = \mathrm{covar}(Y(k), U(k))$	covariance of the measured output and input DFT spectra
τ	time delay (normalized with the sampling period for discrete time systems)
$J(\theta, Z) = \partial\varepsilon(\theta, Z)/\partial\theta$	gradient of residuals $\varepsilon(\theta, Z)$ w.r.t. the parameters θ (dimension $F \times n_\theta$)
$\omega = 2\pi f$	angular frequency
Ω	generalized transform variable: Laplace domain $\Omega = s$, Z-domain $\Omega = z^{-1}$, Richardson domain $\Omega = \tanh(\tau_R s)$, and diffusion phenomena $\Omega = \sqrt{s}$
Ω_k	generalized transform variable evaluated at DFT frequency k: Laplace domain $\Omega_k = j\omega_k$, Z-domain $\Omega_k = e^{-j\omega_k T_s}$, Richardson domain $\Omega_k = \tanh(\tau_R j\omega_k)$, and diffusion phenomena $\Omega_k = \sqrt{j\omega_k}$, with $\omega_k = 2\pi k/N$

List of Abbreviations

ARMA	AutoRegressive Moving Average
ARMAX	AutoRegressive Moving Average with eXternal input
ARX	AutoRegressive with eXternal input
BJ	Box-Jenkins (model structure)
BTLS	Bootstrapped Total Least Squares
CRB	Cramér-Rao bound for biased estimators
DFT	Discrete Fourier Transform
DUT	Device Under Test
EV	Errors-in-Variables
FFT	Fast Fourier Transform
FRF	Frequency Response Function
GSVD	Generalized Singular Value Decomposition
GTLS	Generalized Total Least Squares
iid	independent identically distributed
IV	Instrumental Variables
IWLS	Iterative weighted linear least squares
IQML	Iterative Quadratic Maximum Likelihood
LS	Least Squares
ML	Maximum Likelihood
NLS	Nonlinear Least Squares
NLS-FRF	Nonlinear Least Squares based on FRF measurements
NLS-IO	Nonlinear Least Squares based on Input-Output measurements

OE	Output Error (model structure)
pdf	probability density function
PE	Prediction Error
rms	root mean square value
SBTLS	sample BTLS
SGTLS	sample GTLS
SISO	Single Input, Single Output
SML	sample ML
SNR	Signal-to-Noise Ratio
SSUB	sample SUB
SUB	subspace
SVD	Singular Value Decomposition
TLS	Total Least Squares
UCRB	Cramér-Rao Bound for Unbiased estimators
w.p.1	with probability one
WGTLS	Weighted Generalized Total Least Squares
WLS	Weighted Least Squares

1

An Introduction to Identification

Abstract: In this chapter a brief, intuitive introduction to the identification theory is given. By means of a simple example the reader is made aware of a number of pitfalls associated with a model built from noisy measurements. Starting from this example, the advantages of an identification approach for measuring and modeling are shown, and finally a family of estimators is introduced. A comprehensive introduction to identification can be found, among others, in Beck and Arnold (1977), Goodwin and Payne (1977), Norton (1986), Sörenson (1980), and also in Kendall and Stuart (1979). Basic concepts of statistics such as the expected value, the covariance matrix, and probability density functions are assumed to be known.

1.1 WHAT IS IDENTIFICATION?

From the beginning of our lives, as we grew up, we interacted with our environment. Intuitively, we learned to control our actions by predicting their effect. These predictions are based on an inborn model fitted to reality, using our past experiences. Starting from very simple actions (if I push a ball, it rolls), we soon became very able to deal with much more complicated challenges (walking, running, biking, playing Ping-Pong). Finally, this process culminates in the design of very complicated systems such as radios, airplanes, and mobile phones to satisfy our needs. We even build models just to get a better understanding of our observations of the universe: what does the life cycle of the sun look like? Can we predict the weather of this afternoon, tomorrow, next week, next month? From all these examples it is seen that we never deal with the whole of nature at once: we always focus on the aspects we are interested in and do not try to describe all of reality using one coherent model. The job is split up, and efforts are concentrated on just one part of reality at a time. This part is called the system, the rest of nature being referred to as the environment of the system. Interactions between the system and its environment are described by input and output ports. For a very long time in the history of mankind the models were qualitative, and even today we describe most real-life situations using this "simple" approach: for example, a ball will roll downhill; temperature will rise if the heating has been switched on; it seems it will rain because the sky looks very dark. In the last centuries this qualitative approach was complemented with quantitative models based on advanced mathematics, and until the last decade this seemed to be

System Identification: A Frequency Domain Approach. Rik Pintelon and Johan Schoukens
Copyright © 2001 Institute of Electrical and Electronics Engineers. ISBN: 0-780-36000-1

the most successful approach in many fields of science. Most physical laws are quantitative models describing some part of our impression of reality. However, it also became clear, very soon, that it can be very difficult to match a mathematical model to the available observations and experiences. Consequently, qualitative logical methods typified by fuzzy modeling became more popular, once more. In this book we deal with the mathematical, quantitative modeling approach. Fitting these models to our observations creates new problems. We look at the world through "dirty" glasses: when we measure a length, the weight of a mass, the current or voltage, and so on we always make errors because the instruments we use are not perfect. Also, the models are imperfect; reality is far more complex than the rules we apply. Many systems are not deterministic. They also show a stochastic behavior that makes it impossible to predict exactly their output. Noise in a radio receiver, Brownian motion of small particles, variation of the wind speed in a thunderstorm are all illustrations of this nature. Usually we split the model into a deterministic part and a stochastic part. The deterministic aspects are captured by the mathematical system model, while the stochastic behavior is modeled as a noise distortion. The aim of identification theory is to provide a systematic approach to fit the mathematical model, as well as possible, to the deterministic part, eliminating the noise distortions as much as possible.

Later in this book the meaning of terms such as "system" and "goodness of fit" will be precisely described. Before formalizing the discussion we want to motivate the reader by analyzing a very simple example, illustrating many of the aspects and problems that appear in identification theory.

1.2 IDENTIFICATION: A SIMPLE EXAMPLE

1.2.1 Estimation of the Value of a Resistor

Two groups of students had to measure a resistance. Their measurement setup is shown in Figure 1-1. They passed a constant but unknown current through the resistor. The voltage u_0 across the resistor and the current i_0 through it were measured using a voltmeter and an ampere meter. The input impedance of the voltmeter is very large compared with the unknown resistor so that all the measured current is assumed to pass through the resistor. A set of voltage and current measurements, respectively, $u(k)$, $i(k)$ with $k = 1, 2, ..., N$ is made. The measurement results of each group are shown in Figure 1-2. Because the measurements were very noisy, the groups decided to average their results. Following a lengthy discussion, three estimators for the resistance were proposed:

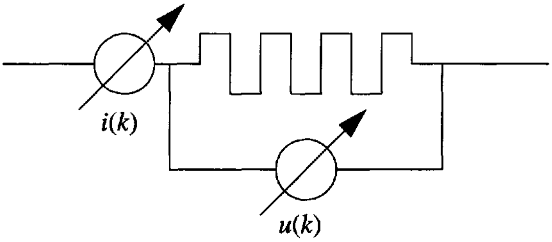


Figure 1-1. Measurement of a resistor.

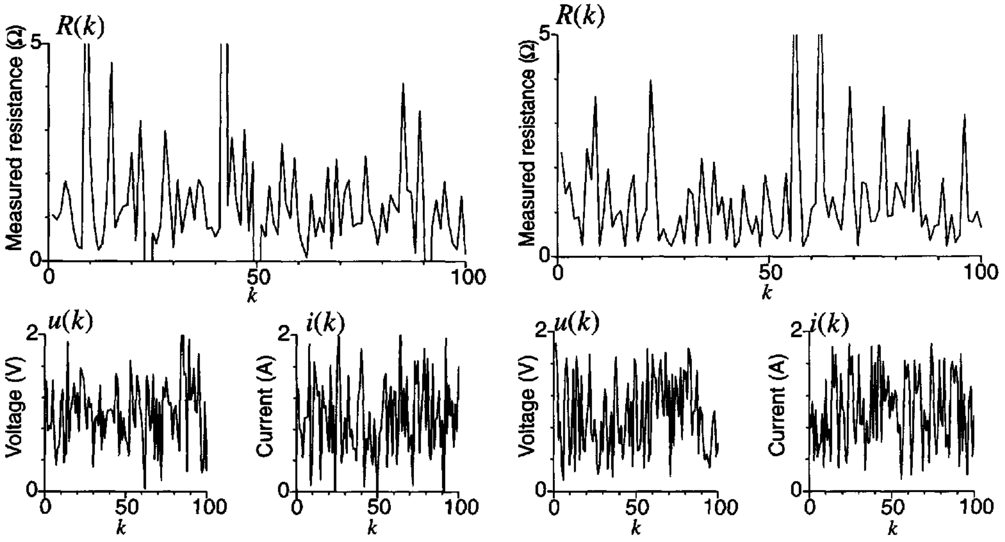
$$\hat{R}_{SA}(N) = \frac{1}{N}\sum_{k=1}^{N}\frac{u(k)}{i(k)} \tag{1-1}$$

$$\hat{R}_{\mathrm{LS}}(N) = \frac{\frac{1}{N}\sum_{k=1}^{N} u(k)i(k)}{\frac{1}{N}\sum_{k=1}^{N} i^2(k)} \tag{1-2}$$

$$\hat{R}_{\mathrm{EV}}(N) = \frac{\frac{1}{N}\sum_{k=1}^{N} u(k)}{\frac{1}{N}\sum_{k=1}^{N} i(k)} \tag{1-3}$$

The index N indicates that the estimate is based on N observations. Note that the three estimators result in the same estimate on noiseless data. Both groups processed their measurements, and their results are given in Figure 1-3. From this figure a number of interesting observations can be made:

- All estimators have large variations for small values of N and seem to converge to an asymptotic value for large values of N, except $\hat{R}_{\mathrm{SA}}(N)$ of group A. This corresponds to the intuitively expected behavior: if a large number of data points are processed we should be able to eliminate the noise influence by the averaging effect.
- The asymptotic values of the estimators depend on the kind of averaging technique that is used. This shows that there is a serious problem: at least two out of the three methods converge to a wrong value. It is not even certain that any one of the estimators is doing well. This is quite catastrophic: even an infinite amount of measurements does not guarantee that the exact value is found.
- The $\hat{R}_{\mathrm{SA}}(N)$ of group A behaves very strangely. Instead of converging to a fixed value, it jumps irregularly up and down before convergence is reached.



Group A Group B

Figure 1-2. Measurement results $u(k)$, $i(k)$ for groups A and B. The plotted value $R(k)$ is obtained by direct division of the voltage by the current: $R(k) = u(k)/i(k)$.

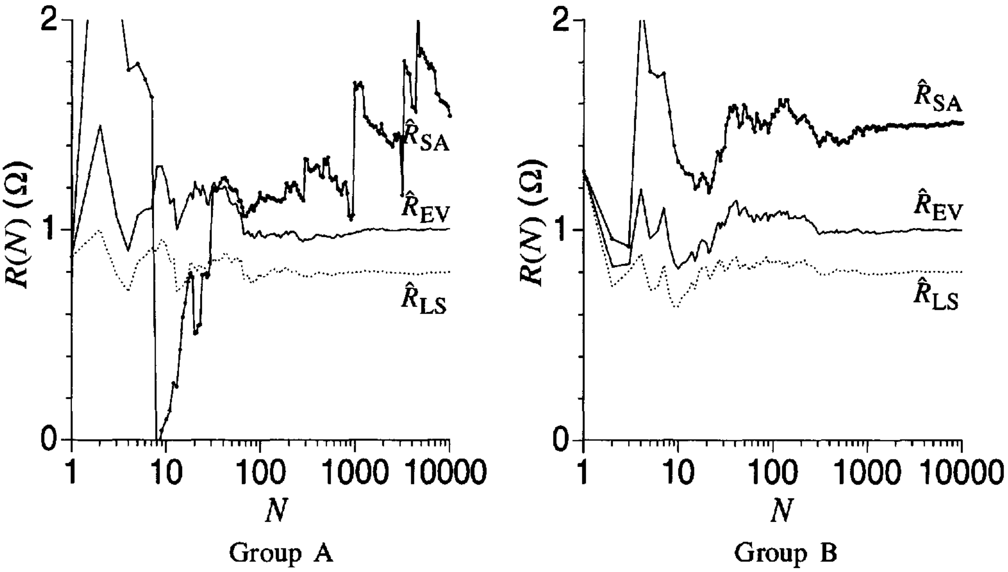


Figure 1-3. Estimated resistance values $\hat{R}(N)$ for both groups as a function of the number of processed data N; full dotted line: $\hat{R}_{SA}$, dotted line: $\hat{R}_{LS}$, full line: $\hat{R}_{EV}$.

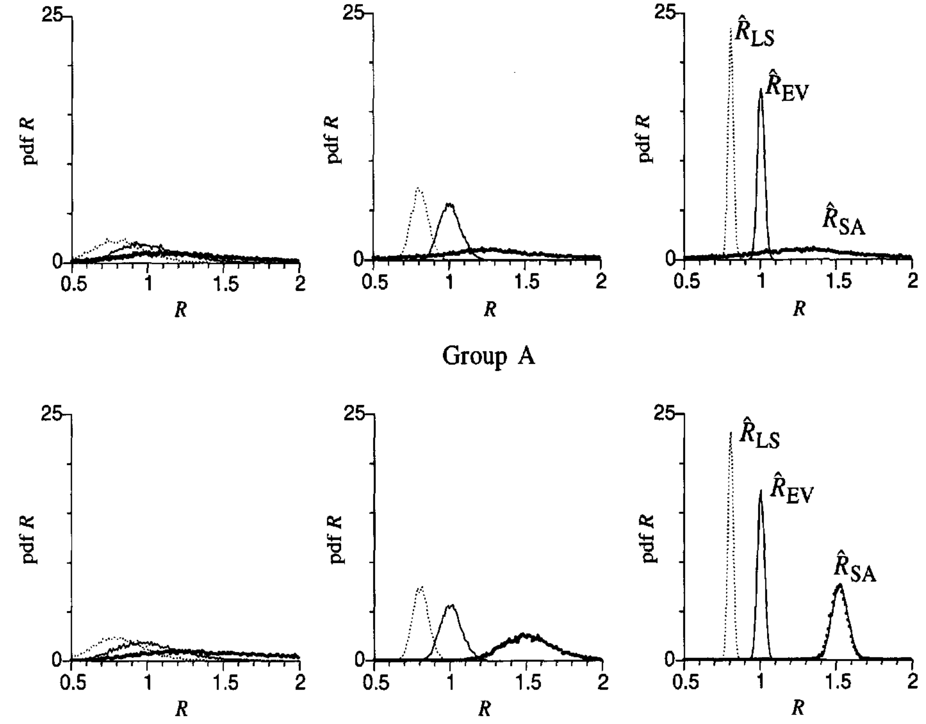
These observations prove very clearly that a good theory is needed to explain and understand the behavior of candidate estimators. This will allow us to make a sound selection out of many possibilities and to indicate in advance, before running expensive experiments, whether the selected method is prone to serious shortcomings.

In order to get a better understanding of their results, the students repeated their experiments many times and looked to the histogram of $\hat{R}(N)$ for $N = 10, 100,$ and 1000. Normalizing these histograms gives an estimate of the pdf (probability density function) of $\hat{R}(N)$ as shown in Figure 1-4. Again, the students could learn a lot from these figures:

- For small values of N the estimates are widely scattered. As the number of processed measurements increases, the pdf becomes more concentrated.
- The estimates $\hat{R}_{LS}(N)$ are less scattered than $\hat{R}_{EV}(N)$, while for $\hat{R}_{SA}(N)$ the odd behavior in the results of group A appears again. The distribution of this estimate does not contract for growing values of N for group A, while it does for group B.
- Again it is clearly visible that the distributions are concentrated around different values.

At this point in the exercise, the students still could not decide which estimator is the best. Moreover, there seems to be a serious problem with the measurements of group A because $\hat{R}_{SA}(N)$ behaves very oddly. First they decided to focus on the scattering of the different estimators, trying to get more insight into the dependence on N. In order to quantify the scattering of the estimates, their standard deviation is calculated and plotted as a function of N in Figure 1-5.

- The standard deviation of $\hat{R}(N)$ decreases monotonically with N except for the pathological case, $\hat{R}_{SA}(N)$, of group A. Moreover, it can be concluded by comparing with the broken line that the standard deviation is proportional to $1/\sqrt{N}$. This is



Group A

Group B

Figure 1-4. Observed pdf of $\hat{R}(N)$ for both groups, from left to right $N = 10,\ 100$, and 1000; full dotted line: $\hat{R}_{SA}(N)$, dotted line: $\hat{R}_{LS}(N)$, full line: $\hat{R}_{EV}(N)$.

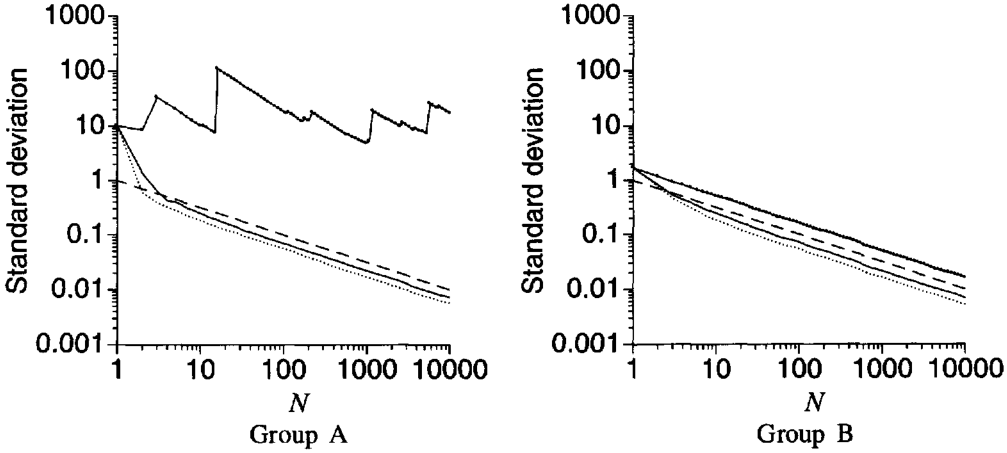


Figure 1-5. Standard deviation of $\hat{R}(N)$ for the different estimators and comparison with $1/\sqrt{N}$; full dotted line: $\hat{R}_{SA}(N)$; dotted line: $\hat{R}_{LS}(N)$, full line: $\hat{R}_{EV}(N)$, dashed line $1/\sqrt{N}$.

in agreement with the rule of thumb which states that the uncertainty on an averaged quantity obtained from independent measurements decreases as $1/\sqrt{N}$.

- The uncertainty in this experiment depends on the estimator. Moreover, the proportionality to $1/\sqrt{N}$ is obtained only for sufficiently large values of N for $\hat{R}_{LS}(N)$ and $\hat{R}_{EV}(N)$.

Because both groups of students used the same programs to process their measurements, they concluded that the strange behavior of $\hat{R}_{SA}(N)$ in group A should be due to a difference in the raw data. For that reason they took a closer look at the time records given in Figure 1-2. Here it can be seen that the measurements of group A are a bit more scattered than those of group B. Moreover, group A measured some negative values for the current while group B did not. In order to get a better understanding, they made a histogram of the raw current data as shown in Figure 1-6.

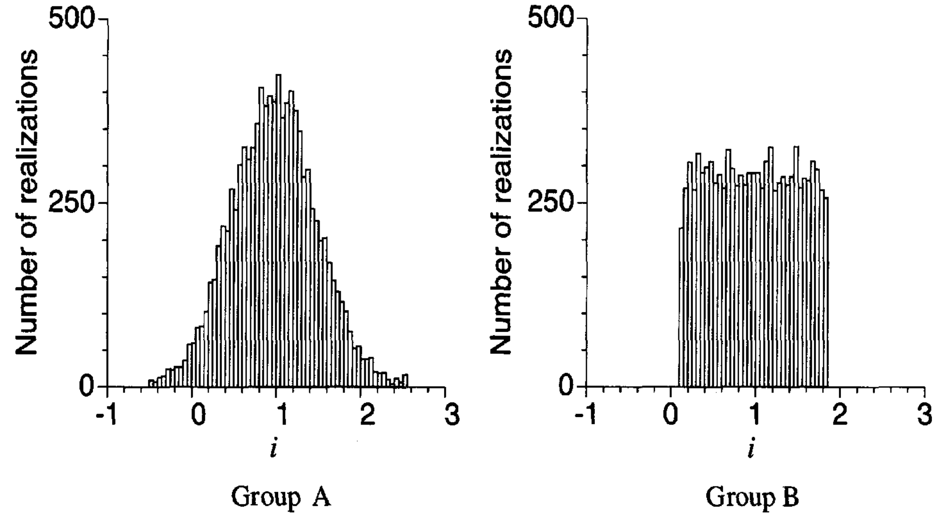


Figure 1-6. Histogram of the current measurements.

These histograms clarify the strange behavior of $\hat{R}_{SA}$ of group A. The noise on the measurements of group A looks completely different from that of group B. Because of the noise on the current measurements, there is a significant risk of getting current values that are very close to zero for group A, whereas this is not so for group B. These small current measurements blow up the estimate $\hat{R}(k) = u(k)/i(k)$ for some k, so that the running average $\hat{R}_{SA}$ cannot converge, or more precisely, the expected value $\mathcal{E}\{u(k)/i(k)\}$ does not exist. This will be discussed in more detail later in this chapter. This example shows very clearly that there is a strong need for methods that can generate and select between different estimators. Before setting up a general framework, the resistance problem is further elaborated.

It is also remarkable to note that although the noise on the measurements is completely differently distributed, the distribution of the estimated resistance values $\hat{R}_{LS}$ and $\hat{R}_{EV}$ seems to be the same in Figure 1-4 for both groups.

1.2.2 Simplified Analysis of the Estimators

With the knowledge obtained from the previous series of experiments, the students eliminated $\hat{R}_{\text{SA}}$, but they were still not able to decide whether $\hat{R}_{\text{LS}}$ or $\hat{R}_{\text{EV}}$ was the best. More advanced analysis techniques are needed to solve this problem. As the estimates are based on a combination of a finite number of noisy measurements, there are bound to be stochastic variables. Therefore, an analysis of the stochastic behavior is needed to select between both estimators. This is done by calculating the limiting values and making series expansions of the estimators. In order to keep the example simple, we will use some of the limit concepts quite loosely. Precise definitions are postponed to Section 14.6. Three observed problems are analyzed in the following:

- Why do the asymptotic values depend on the estimator?
- Can we explain the behavior of the variance?
- Why does the $\hat{R}_{\text{SA}}$ estimator behave strangely for group A?

To do this it is necessary to specify the stochastic framework: how are the measurements disturbed with the noise (multiplicative, additive), and how is the noise distributed? For simplicity, we assume that the current and voltage measurements are disturbed by additive zero mean, independently and identically distributed noise, formally formulated as:

$$i(k) = i_0 + n_i(k) \qquad u(k) = u_0 + n_u(k) \tag{1-4}$$

where i_0 and u_0 are the exact but unknown values of the current and the voltage, $n_i(k)$ and $n_u(k)$, are the noise on the measurements.

Assumption 1.1 (disturbing noise): $n_i(k)$ and $n_u(k)$ are mutually independent, zero mean, independent and identically distributed (iid) random variables with a symmetric distribution and with variance σ_u^2 and σ_i^2.

1.2.2.1 Asymptotic Value of the Estimators. In this section the limiting value of the estimates for $N \to \infty$ is calculated. The calculations are based on the observation that the sample mean of iid random variables $x(k)$, $k = 1, \ldots, N$ converges to its expected value (see Section 14.9), $\mathcal{E}\{x\}$

$$\lim_{N \to \infty} \frac{1}{N} \sum_{k=1}^{N} x(k) = \mathcal{E}\{x\} \tag{1-5}$$

Moreover, if $x(k)$ and $y(k)$ obey Assumption 1.1, then

$$\lim_{N \to \infty} \frac{1}{N} \sum_{k=1}^{N} x(k) y(k) = 0 \tag{1-6}$$

Because we are dealing here with stochastic variables, the meaning of this statement should be defined more precisely, but in this section we will just use this formal notation and make the calculations straightforwardly (see Chapter 14.6 for a formal definition).

The first estimator we analyze is $\hat{R}_{\mathrm{LS}}(N)$. Taking the limit of (1-2) gives

$$\lim_{N\to\infty}\hat{R}_{\mathrm{LS}}(N) = \lim_{N\to\infty}\frac{\sum_{k=1}^{N}u(k)i(k)}{\sum_{k=1}^{N}i^2(k)} = \frac{\lim_{N\to\infty}\sum_{k=1}^{N}(u_0+n_u(k))(i_0+n_i(k))}{\lim_{N\to\infty}\sum_{k=1}^{N}(i_0+n_i(k))^2} \tag{1-7}$$

Or, after dividing the numerator and denominator by N,

$$\lim_{N\to\infty}\hat{R}_{\mathrm{LS}}(N) = \frac{\lim_{N\to\infty}\left[u_0i_0+\frac{u_0}{N}\sum_{k=1}^{N}n_i(k)+\frac{i_0}{N}\sum_{k=1}^{N}n_u(k)+\frac{1}{N}\sum_{k=1}^{N}n_u(k)n_i(k)\right]}{\lim_{N\to\infty}\left[i_0^2+\frac{1}{N}\sum_{k=1}^{N}n_i^2(k)+\frac{2i_0}{N}\sum_{k=1}^{N}n_i(k)\right]}$$

Because n_i and n_u are zero mean iid, it follows from (1-5) and (1-6) that

$$\lim_{N\to\infty}\frac{1}{N}\sum_{k=1}^{N}n_u(k) = 0,\ \lim_{N\to\infty}\frac{1}{N}\sum_{k=1}^{N}n_i(k) = 0, \text{ and } \lim_{N\to\infty}\frac{1}{N}\sum_{k=1}^{N}n_u(k)n_i(k) = 0$$

However, the sum of the squared current noise distributions does not converge to zero but converges to a constant value different from zero

$$\lim_{N\to\infty}\frac{1}{N}\sum_{k=1}^{N}n_i^2(k) = \sigma_i^2$$

so that the asymptotic value becomes:

$$\lim_{N\to\infty}\hat{R}_{\mathrm{LS}}(N) = \frac{u_0i_0}{i_0^2+\sigma_i^2} = R_0\frac{1}{1+\sigma_i^2/i_0^2} \tag{1-8}$$

This simple analysis gives a lot of insight into the behavior of the $\hat{R}_{\mathrm{LS}}(N)$ estimator. Asymptotically, this estimator underestimates the value of the resistance due to quadratic noise contributions in the denominator. Although the noise disappears in the averaging process of the numerator, it contributes systematically in the denominator. This results in a systematic error (called bias) that depends on the signal-to-noise ratio (SNR) of the current measurements: i_0/σ_i.

The analysis of the second estimator $\hat{R}_{\mathrm{EV}}(N)$ is completely similar. Using (1-3), we get

$$\lim_{N\to\infty}\hat{R}_{\mathrm{EV}}(N) = \lim_{N\to\infty}\frac{\sum_{k=1}^{N}u(k)}{\sum_{k=1}^{N}i(k)} = \frac{\lim_{N\to\infty}\frac{1}{N}\sum_{k=1}^{N}(u_0+n_u(k))}{\lim_{N\to\infty}\frac{1}{N}\sum_{k=1}^{N}(i_0+n_i(k))} \tag{1-9}$$

or

$$\lim_{N\to\infty}\hat{R}_{\mathrm{EV}}(N)=\frac{u_0+\lim_{N\to\infty}\frac{1}{N}\sum_{k=1}^{N}n_u(k)}{i_0+\lim_{N\to\infty}\frac{1}{N}\sum_{k=1}^{N}n_i(k)}=\frac{u_0}{i_0}=R_0 \tag{1-10}$$

so that we can conclude now that $\hat{R}_{\mathrm{EV}}(N)$ converges to the true value and should be preferred over $\hat{R}_{\mathrm{LS}}(N)$. These conclusions are also confirmed by the students' results in Figure 1-3, where it is seen that the asymptotic value of $\hat{R}_{\mathrm{LS}}(N)$ is much smaller than that of $\hat{R}_{\mathrm{EV}}(N)$.

1.2.2.2 Strange Behavior of the "Simple Approach." Finally, we have to analyze $\hat{R}_{\mathrm{SA}}(N)$ in order to understand its strange behavior. Can't we repeat the previous analysis here? Consider

$$\hat{R}_{\mathrm{SA}}(N)=\frac{1}{N}\sum_{k=0}^{N}\frac{u(k)}{i(k)}=\frac{1}{N}\sum_{k=0}^{N}\frac{u_0+n_u(k)}{i_0+n_i(k)} \tag{1-11}$$

A major difference from the previous estimators is the order of summing and dividing: here the measurements are first divided and then summed together, whereas for the other estimators we first summed the measurements together before making the division. In other words, for $\hat{R}_{\mathrm{LS}}(N)$ and $\hat{R}_{\mathrm{EV}}(N)$ we first applied an averaging process (summing over the measurements) before making the division. This makes an important difference.

$$\hat{R}_{\mathrm{SA}}(N)=\frac{1}{N}\frac{u_0}{i_0}\sum_{k=0}^{N}\frac{1+n_u(k)/u_0}{1+n_i(k)/i_0} \tag{1-12}$$

In order to process $\hat{R}_{\mathrm{SA}}(N)$ along the same lines as the other estimators, we should get rid of the division, for example, by making a Taylor series expansion:

$$\frac{1}{1+x}=\sum_{l=0}^{\infty}(-1)^l x^l \text{ for } |x|<1 \tag{1-13}$$

with $x=n_i(k)/i_0$. Because the terms $n_i^{2l+1}(k)$ and $n_u^l(k)n_i^l(k)$ disappear in the averaging process (the pdfs are symmetric), the limiting value becomes

$$\lim_{N\to\infty}\hat{R}_{\mathrm{SA}}(N)=R_0\left(1+\frac{1}{N}\sum_{k=1}^{N}(n_i(k)/i_0)^2+\frac{1}{N}\sum_{k=1}^{N}(n_i(k)/i_0)^4+\cdots\right) \tag{1-14}$$

with $|n_i(k)/i_0|<1$. If we neglect all terms of order 4 or more, the final result becomes

$$\lim_{N\to\infty}\hat{R}_{\mathrm{SA}}(N)=R_0(1+\sigma_i^2/i_0^2) \tag{1-15}$$

if $|n_i(k)/i_0|<1,\ \forall k$.

From this analysis we can draw two important conclusions:

- The asymptotic value exists only if the following condition on the measurements is met: the series expansion must exist otherwise Eq. (1-15) is NOT valid. The measurements of group A violate the condition that is given in (1-14), while those of group B obey it (see Figure 1-6). A more detailed analysis shows that this condition is too rigorous. In practice it is enough that the expected value $\mathcal{E}\{\hat{R}_{\mathrm{SA}}(N)\}$ exists (see Chapter 15). Because this value depends on the pdf of the noise, a more detailed analysis of the measurement noise would be required. For some noise distributions the expected value exists even if the Taylor expansion does not!
- If the asymptotic value exists, Eq. (1-15) shows that it will be too large. This is also seen in the results of group B in Figure 1-3. We know already that $\hat{R}_{\mathrm{EV}}(N)$ converges to the exact value, and $\hat{R}_{\mathrm{SA}}(N)$ is clearly significantly larger.

1.2.2.3 Variance Analysis. In order to get a better understanding of the sensitivity of the different estimators to the measurement noise, the students made a variance analysis using first-order Taylor series approximations.

Again they began with the $\hat{R}_{\mathrm{LS}}(N)$. Starting from Eq. (1-7) and neglecting all second-order contributions such as $n_u(k)n_i(k)$ or $n_i^2(k)$, it is found that

$$\hat{R}_{\mathrm{LS}}(N) \approx R_0\left(1 + \frac{1}{N}\sum\nolimits_{k=1}^{N}(n_u(k)/u_0 - n_i(k)/i_0)\right) = R_0 + \Delta R \tag{1-16}$$

The approximated variance $\mathrm{var}(\hat{R}_{\mathrm{LS}}(N))$ is (using Assumption 1.1)

$$\mathrm{var}(\hat{R}_{\mathrm{LS}}(N)) = \mathcal{E}\{(\Delta R)^2\} = \frac{R_0^2}{N}\left(\frac{\sigma_u^2}{u_0^2} + \frac{\sigma_i^2}{i_0^2}\right) \tag{1-17}$$

with $\mathcal{E}\{\ \}$ the expected value. Note that during the calculation of the variance, the shift of the mean value of $\hat{R}_{\mathrm{LS}}(N)$ is not considered because it is a second-order contribution.

For the other two estimators, exactly the same results are found:

$$\mathrm{var}(\hat{R}_{\mathrm{EV}}(N)) = \mathrm{var}(\hat{R}_{\mathrm{SA}}(N)) = \frac{R_0^2}{N}\left(\frac{\sigma_u^2}{u_0^2} + \frac{\sigma_i^2}{i_0^2}\right) \tag{1-18}$$

The result $\mathrm{var}(\hat{R}_{\mathrm{SA}}(N))$ is valid only if the expected values exist.

Again, a number of interesting conclusions can be drawn from this result

- The standard deviation is proportional to $1/\sqrt{N}$ as was found before in Figure 1-5.
- Although it is possible to reduce the variance by averaging over repeated measurements, this is no excuse for sloppy experiments because the uncertainty is inversely proportional to the SNR of the measurements. Increasing the SNR requires many more measurements in order to get the same final uncertainty on the estimates.
- The variances of the three estimators should be the same. This seems to conflict with the results of Figure 1-5. However, the theoretical expressions are based on first-order approximations. If the SNR drops to values that are too small, the second-order moments are no longer negligible. In order to check this, the students set up a simulation and tuned the noise parameters so that they got the same behavior as they

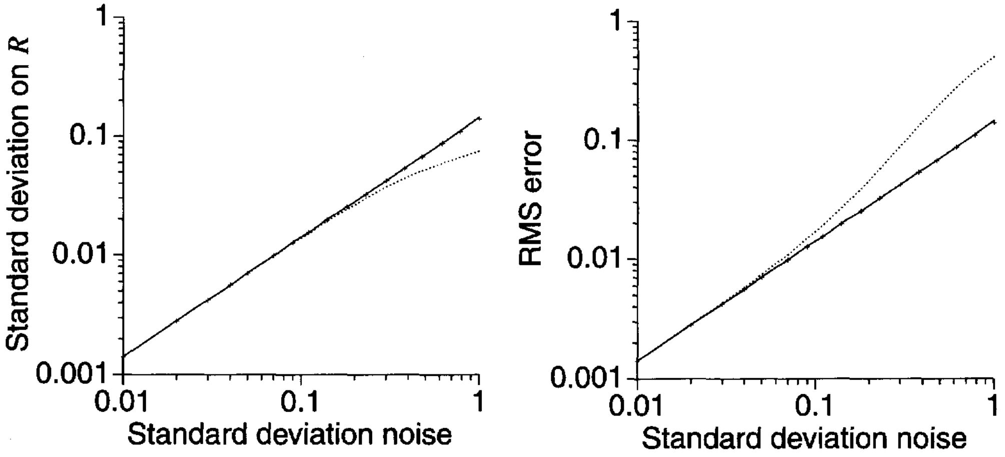


Figure 1-7. Evolution of the standard deviation and the rms error on the estimated resistance value as a function of the standard deviation of the noise ($\sigma_u = \sigma_i$). ____ : $\hat{R}_{EV}(N)$, ... : $\hat{R}_{LS}(N)$, +++ theoretical value σ_R.

had observed in their measurements. These values were: $i_0 = 1\text{A}$, $u_0 = 1\text{V}$, $\sigma_i = 1\text{A}$, $\sigma_u = 1\text{V}$. The noise of group A is normally distributed and uniformly distributed for group B. Next they varied the standard deviations and plotted the results in Figure 1-7 for $\hat{R}_{EV}(N)$ and $\hat{R}_{LS}(N)$. Here it is clear that for higher SNR the uncertainties coincide, whereas they differ significantly for the lower SNR. To give closed form mathematical expressions for this behavior, it is not enough any more to specify the first- and second-order moments of the noise (mean, variance); the higher order moments or the pdf of the noise is also required (see Section 14.15).

- Although $\hat{R}_{LS}(N)$ has a smaller variance than $\hat{R}_{EV}(N)$ for low SNR, its total root mean square (rms) error (difference with respect to the true value) is significantly larger because of its systematic error. The following is quite a typical observation: many estimators reduce the stochastic error at the cost of systematic errors. For the $\hat{R}_{EV}$ the rms error is completely due to the variability of the estimator because the rms error coincides completely with the theoretical curve of the standard deviation.

1.2.3 Interpretation of the Estimators: A Cost Function–Based Approach

The previous section showed that there is not just one single estimator for each problem. Moreover, the properties of the estimators can vary quite a lot. This raises two questions: how can we generate good estimators and how can we evaluate their properties? The answers are given in this and the following sections. In order to recognize good estimators it is necessary to specify what a good estimator is. This is done in the next section. First we will deal with the question of how estimators are generated. Again, there exist different approaches. A first group of methods starts from a deterministic approach. A typical example is the observation that the noiseless data should obey some model equations. The system parameters are then extracted by intelligent manipulation of these equations, usually inspired by numerical or algebraic techniques. Next, the same procedure is used on noisy data. The major disadvantage of this approach is that it does not guarantee at all that the resulting estimator has good noise behavior. The estimates can be extremely sensitive to disturbing noise. The alternative is to embed the problem in a stochastic framework. A typical question to be answered is: where does the disturbing noise sneak into my problem and how does it behave? To answer this question, it is necessary to make a careful analysis of the measurement setup. Next, the best parameters are selected using statistical considerations. In most cases these methods lead to a cost function interpretation and the estimates are found as the arguments that minimize the cost function. The estimates of the previous section can be found as the minimizers of the following cost functions:

$\hat{R}_{\text{SA}}(N)$: Consider the successive resistance estimates $R(k) = u(k)/i(k)$. The overall estimate after N measurements is then the argument minimizing the following cost function:

$$\hat{R}_{\text{SA}}(N) = \arg\min_R V_{\text{SA}}(R, N) \text{ with } V_{\text{SA}}(R, N) = \sum_{k=1}^{N} (R(k) - R)^2 \tag{1-19}$$

This is the most simple approach ("SA" stands for simple approach) of the estimation problem. As seen before, it has very poor properties.

$\hat{R}_{\text{LS}}(N)$: A second possibility is to minimize the equation errors in the model equation $u(k) - Ri(k) = e(k, R)$ in least squares (LS) sense. For noiseless measurements $e(k, R_0) = 0$, with R_0 the true resistance value,

$$\hat{R}_{\text{LS}}(N) = \arg\min_R V_{\text{LS}}(R, N) \text{ with } V_{\text{LS}}(R, N) = \sum_{k=1}^{N} e^2(k, R) \tag{1-20}$$

$\hat{R}_{\text{EV}}(N)$: The basic idea of the last approach is to express that the current as well as the voltage measurements are disturbed by noise. This is called the errors-in-variables (EV) approach. The idea is to estimate the exact current and voltage (i_0, u_0), parameterized as $(i_{\text{p}}, u_{\text{p}})$ keeping in mind the model equation $u_0 = Ri_0$.

$$\begin{aligned} \hat{R}_{\text{EV}}(N) &= \arg\min_{R,\, i_{\text{p}},\, u_{\text{p}}} V_{\text{EV}}(R, i_{\text{p}}, u_{\text{p}}, N) \text{ subject to } u_{\text{p}} = Ri_{\text{p}} \\ V_{\text{EV}}(R, i_{\text{p}}, u_{\text{p}}, N) &= \sum_{k=1}^{N} (u(k) - u_{\text{p}})^2 + \sum_{k=1}^{N} (i(k) - i_{\text{p}})^2 \end{aligned} \tag{1-21}$$

This wide variety of possible solutions and motivations illustrates very well the need for a more systematic approach. In this book we put the emphasis on a stochastic embedding approach, selecting a cost function on the basis of a noise analysis of the general measurement setup that is used.

All the cost functions that we presented are of the "least squares" type. Again there exist many other possibilities, for example, the sum of the absolute values. There are two reasons for choosing a quadratic cost: first, it is easier to minimize than other functions, and second, we will show that normally distributed disturbing noise leads to a quadratic criterion. This does not imply that it is the best choice from all points of view. If it is known that some outliers in the measurements can appear (due to exceptionally large errors, a temporary sensor failure, a transmission error, etc.), it can be better to select a least absolute values cost function (sum of the absolute values) because these outliers are strongly emphasized in a least squares concept (Huber, 1981; Van den Bos, 1985). Sometimes a mixed criterion is used; for example, the small errors are quadratically weighted while the large errors only appear linear in the cost to reduce the impact of outliers (Ljung, 1995).

1.3 DESCRIPTION OF THE STOCHASTIC BEHAVIOR OF ESTIMATORS

Because the estimates are obtained as a function of a finite number of noisy measurements, they are stochastic variables as well. Their pdf is needed in order to characterize them completely. However, in practice it is usually very hard to derive it, so that the behavior of the estimates is described by a few numbers only, such as their mean value (as a description of the location) and the covariance matrix (to describe the dispersion). Both aspects are discussed in the following. A detailed discussion is given in Chapter 14.

1.3.1 Location Properties: Unbiased and Consistent Estimates

The choice for the mean value is not obvious at all from a theoretical point of view. Other location parameters such as the median or the mode (Stuart and Ord, 1987) could be used too, but the latter are much more difficult to analyze in most cases. As it can be shown that many estimates are asymptotically normally distributed under weak conditions, this choice is not so important because in the normal case, these location parameters coincide. It seems very natural to require that the mean value equals the true value, but it turns out to be impractical. What are the true parameters of a system? We can speak about true parameters only if an exact model exists. It is clear that this is a purely imaginary situation because in practice we always stumble on model errors so that only excitation-dependent approximations can be made. For theoretical reasons it still makes sense to consider the concept of "true parameters," but it is clear at this point that we have to generalize to more realistic situations. One possible generalization is to consider the estimator evaluated in the noiseless situation as the "best" approximation. These parameters are then used as a reference value to compare the results obtained from noisy measurements. The goal is then to remove the influence of the disturbing noise so that the estimator converges to this reference value.

Definition 1.2 (Unbiasedness): An estimator $\hat{\theta}$ of the parameters θ_0 is unbiased if $\mathcal{E}\{\hat{\theta}\} = \theta_0$, for all true parameters θ_0. Otherwise it is a biased estimator.

If the expected value equals the true value only for an infinite number of measurements, then the estimator is called asymptotically unbiased. In practice, it turns out that (asymptotic) unbiasedness is a hard requirement to deal with.

Example 1.3 (Unbiased and Biased Estimators): At the end of their experiments the students wanted to estimate the value of the voltage over the resistor. Starting from the measurements (1-4), they first carry out a noise analysis of their measurements by calculating the sample mean value and the sample variance:

$$\hat{u}(N) = \frac{1}{N}\sum_{k=1}^{N} u(k) \quad \text{and} \quad \hat{\sigma}_u^2(N) = \frac{1}{N}\sum_{k=1}^{N} (u(k) - \hat{u}(N))^2 \tag{1-22}$$

Applying the previous definition, it is readily seen that

$$\mathcal{E}\{\hat{u}(N)\} = \frac{1}{N}\sum_{k=1}^{N} \mathcal{E}\{u(k)\} = \frac{1}{N}\sum_{k=1}^{N} u_0 = u_0 \tag{1-23}$$

because the noise is zero mean, so that their voltage estimate is unbiased. The same can be done for the variance estimate:

$$\mathscr{E}\{\hat{\sigma}_u^2(N)\} = \frac{N-1}{N}\sigma_u^2 \tag{1-24}$$

This estimator shows a systematic error of σ_u^2/N and is thus biased. However, as $N \to \infty$ the bias disappears, and following the definitions it is asymptotically unbiased. It is clear that a better estimate would be $\frac{1}{N-1}\sum_{k=1}^{N}(u(k) - \hat{u}(N))^2$, which is the expression that is found in the handbooks on statistics. □

For many estimators, it is very difficult or even impossible to find the expected value analytically. Sometimes it does not even exist as was the case for $\hat{R}_{\mathrm{SA}}(N)$ of group A. Moreover, unbiased estimators can still have a bad distribution; for example, the pdf of the estimator is symmetrically distributed around its mean value, with a minimum at the mean value. Consequently, a more handy tool (e.g., consistency) is needed.

Definition 1.4 (Consistency): An estimator $\hat{\theta}(N)$ of the parameters θ_0 is weakly consistent if it converges in probability to θ_0: $\operatorname{plim}_{N\to\infty}\hat{\theta}(N) = \theta_0$ and strongly consistent if it converges with probability one (almost surely) to θ_0: $\operatorname{a.s.lim}_{N\to\infty}\hat{\theta}(N) = \theta_0$.

The precise explanation of these probability limits is given in Section 14.6. Loosely explained, it means that the pdf of $\hat{\theta}(N)$ contracts around the true value θ_0, or $\lim_{N\to\infty}\operatorname{Prob}(|\hat{\theta}(N) - \theta_0| > \delta > 0) = 0$. The major advantage of the consistency concept is purely mathematical: it is much easier to prove consistency than unbiasedness using probabilistic theories starting from the cost function interpretation. A general outline on how to prove consistency is given in Section 15.3. Another nice property of the plim is that it can be interchanged with a continuous function: $\operatorname{plim} f(a) = f(\operatorname{plim}(a))$ if both limits exist (see Section 14.8). In fact, it was this property that we applied during the calculations of the limit values of $\hat{R}_{\mathrm{LS}}$ and $\hat{R}_{\mathrm{EV}}$, for example,

$$\operatorname*{plim}_{N\to\infty}\hat{R}_{\mathrm{EV}}(N) = \operatorname*{plim}_{N\to\infty}\frac{\frac{1}{N}\sum_{k=1}^{N}u(k)}{\frac{1}{N}\sum_{k=1}^{N}i(k)} = \frac{\operatorname*{plim}_{N\to\infty}\frac{1}{N}\sum_{k=1}^{N}u(k)}{\operatorname*{plim}_{N\to\infty}\frac{1}{N}\sum_{k=1}^{N}i(k)} = \frac{u_0}{i_0} = R_0 \tag{1-25}$$

Consequently, $\hat{R}_{\mathrm{EV}}(N)$ is a weakly consistent estimator. Calculating the expected value is much more involved in this case due to the division. Therefore, consistency is a better suited concept than (asymptotic) unbiasedness to study it.

1.3.2 Dispersion Properties: Efficient Estimators

In this book the covariance matrix is used to measure the dispersion of an estimator, that is, to ascertain how much the actual estimator is scattered around its limiting value. Again this choice, among other possibilities (for example, percentiles), is highly motivated from a mathematical point of view. Within the stochastic framework used, it will be quite easy to calculate the covariance matrix whereas it is much more involved to obtain the other measures. For normal distributions, all dispersion measures are obtainable from the covariance matrix so that for most estimators this choice is not too restrictive because their distribution converges to a normal one.

As users, we are highly interested in estimators with minimal errors. However, because we can collect only a finite number of noisy measurements, it is clear that there are limits on the accuracy and precision we can reach. This is precisely quantified in the Cramér-Rao inequality. This inequality provides a lower bound on the covariance matrix of a(n) (un)biased estimator starting from the likelihood function. First we introduce the likelihood function; next we present the Cramér-Rao lower bound.

Consider the measurements $z \in \mathbb{R}^N$ obtained from a system described by a hypothetical, exact model that is parameterized in θ. These measurements are disturbed by noise and are, hence, stochastic variables that are characterized by a probability density function $f(z|\theta_0)$ that depends on the exact model parameters θ_0 with $\int_{z \in \mathbb{R}^N} f(z|\theta_0)dz = 1$. Next we can interpret this relation conversely, namely, how likely is it that a specific set of measurements $z = z_m$ are generated by a system with parameters θ? In other words, we now consider a given set of measurements and view the model parameters as the free variables:

$$L(z_m|\theta) = f(z = z_m|\theta) \tag{1-26}$$

with θ the free variables. $L(z_m|\theta)$ is called the likelihood function. In many calculations the log likelihood function $l(z|\theta) = \ln(L(z|\theta))$ is used. In (1-26) we used z_m to indicate explicitly that we use the numerical values of the measurements that were obtained from the experiments. From here on, we just use z as a symbol because it will be clear from the context what interpretation should be given to z. The reader should be aware that $L(z|\theta)$ is not a probability density function with respect to θ because $\int_{\theta} L(z|\theta)d\theta \neq 1$. Notice the subtle difference in terminology; that is, probability is replaced by likeliness.

The Cramér-Rao lower bound gives a lower limit on the covariance matrix of parameters. Under quite general conditions, this limit is universal and independent of the selected estimator: no estimator that violates this bound can be found. It is given by (see Section 14.12)

$$\begin{aligned} CR(\theta_0) &= \left(I_{n_\theta} + \frac{\partial b_\theta}{\partial \theta}\right)^T Fi^{-1}(\theta_0)\left(I_{n_\theta} + \frac{\partial b_\theta}{\partial \theta}\right) \\ Fi(\theta_0) &= \mathcal{E}\left\{\left(\frac{\partial l(z|\theta)}{\partial \theta}\right)^T\left(\frac{\partial l(z|\theta)}{\partial \theta}\right)\right\} = -\mathcal{E}\left\{\frac{\partial^2 l(z|\theta)}{\partial \theta^2}\right\} \end{aligned} \tag{1-27}$$

The derivatives are calculated in $\theta = \theta_0$, and $b_\theta = \mathcal{E}\{\hat{\theta}\} - \theta_0$ is the bias on the estimator. Note that for biased estimators ($\partial b_\theta / \partial \theta \neq 0$) the lower bound (1-27) can be be zero: $CR(\theta_0) = 0$ (see Example 14.20 on page 458). For unbiased estimators (1-27) reduces to $CR(\theta_0) = Fi^{-1}(\theta_0)$.

$Fi(\theta)$ is called the Fisher information matrix: it is a measure of the information in an experiment: the larger the matrix, the more information there is. In (1-27) it is assumed that the first and second derivatives of the log likelihood function exist with respect to θ.

Example 1.5 (Influence of the Number of Parameters on the Cramér-Rao Lower Bound): A group of students wanted to determine the flow of tap water by measuring the height $h_0(t)$ of the water in a measuring jug as a function of time t. However, their work was not precise and in the end they were not sure about the exact starting time of their experiment.

They included it in the model as an additional parameter: $h_0(t) = a(t - t_{\text{start}}) = at + b$, and $\theta = [a, b]^T$. Assume that the noise $n_h(k)$ on the height measurements is iid zero mean normally distributed $N(0, \sigma^2)$, and the noise on the time instances is negligible $h(k) = at_k + b + n_h(k)$; then the following stochastic model can be used:

$$\text{Prob}(h(k), t_k) = \text{Prob}(h(k) - (at_k + b)) = \text{Prob}(n_h(k))$$

where $\text{Prob}(h(k), t_k)$ is the probability of making the measurements $h(k)$ at t_k. The likelihood function for the set of measurements $h = \{(h(1), t_1), \ldots, (h(N), t_N)\}$ is

$$L(h|a, b) = \frac{1}{(2\pi\sigma^2)^{N/2}} e^{-\frac{1}{2\sigma^2}\sum_{k=1}^{N}(h(k) - at_k - b)^2} \tag{1-28}$$

and the log likelihood function becomes

$$l(h|a, b) = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{k=1}^{N}(h(k) - at_k - b)^2 \tag{1-29}$$

The Fisher information matrix and the Cramér-Rao lower bound are found using (1-27):

$$Fi(a, b) = \frac{N}{\sigma^2}\begin{bmatrix} s^2 & \mu \\ \mu & 1 \end{bmatrix} \rightarrow CR(a, b) = Fi^{-1}(a, b) = \frac{\sigma^2}{N(s^2 - \mu^2)}\begin{bmatrix} 1 & -\mu \\ -\mu & s^2 \end{bmatrix} \tag{1-30}$$

with $\mu = \frac{1}{N}\sum_{k=1}^{N} t_k$ and $s^2 = \frac{1}{N}\sum_{k=1}^{N} t_k^2$. These expressions are very informative. First of all, we can note that the attainable uncertainty is proportional to the standard deviation of the noise. This means that inaccurate measurements result in poor estimates, or identification is no excuse for sloppy measurements. The uncertainty decreases as $\sqrt{N}$, which can be used as a rule of thumb whenever independent measurements are processed. Finally, it can also be noted that the uncertainty depends on the actual time instances used in the experiment. In other words, by making a proper design of the experiment, it is possible to influence the uncertainty on the estimates. This idea will be exploited fully in Chapter 4. Another question we can answer now is what price is paid to include the additional model parameter b to account for the unknown starting time. By comparing $Fi^{-1}(a, b)$ with $Fi^{-1}(a)$ (assuming that b is known), it is found that

$$\sigma_a^2(a, b) = \frac{\sigma^2}{N(s^2 - \mu^2)} \geq \frac{\sigma^2}{Ns^2} = \sigma_a^2(a) \tag{1-31}$$

where $\sigma_a^2(a, b)$ is the lower bound on the variance of a if both parameters are estimated, else $\sigma_a^2(a)$ is the lower bound if only a is estimated. This shows that adding additional parameters to a model increases the minimum attainable uncertainty on it. Of course, these parameters may be needed to remove systematic errors so that a balance between stochastic errors and systematic errors is achieved. This is further elaborated in Chapter 9. □

The Cramér-Rao lower bound is a conservative estimate of the smallest possible covariance matrix that is not always attainable (the values may be too small). Tighter bounds exist (Abel, 1993) but these are more involved to calculate. Consequently, the Cramér-Rao bound is the criterion most used to verify the efficiency of an estimator.

Definition 1.6 (Efficiency): An unbiased estimator is called efficient if its covariance matrix is smaller than that of any other unbiased estimator.

An unbiased estimator that reaches the Cramér-Rao lower bound is also an efficient estimator. For biased estimators, a generalized expression should be used (see Section 14.12).

1.4 BASIC STEPS IN THE IDENTIFICATION PROCESS

Each identification session consists of a series of basic steps. Some of them may be hidden or selected without the user being aware of his choice. Clearly, this can result in poor or suboptimal results. In each session the following actions should be taken:

- Collect information about the system.
- Select a model structure to represent the system.
- Choose the model parameters to fit the model as well as possible to the measurements: selection of a "goodness of fit" criterion.
- Validate the selected model.

Each of these points is discussed in more detail next.

1.4.1 Collect Information about the System

If we want to build a model for a system, we should get information about it. This can be done by just watching the natural fluctuations (e.g., vibration analysis of a bridge that is excited by normal traffic), but most often it is more efficient to set up dedicated experiments that actively excite the system (e.g., controlled excitation of a mechanical structure using a shaker). In the latter case, the user has to select an excitation that optimizes his own goal (for example, minimum cost, minimum time, or minimum power consumption for a given measurement accuracy) within the operator constraints (e.g., the excitation should remain below a maximum allowable level). The quality of the final result can depend heavily on the choices that are made. Later in this book we will spend a lot of time on the selection of the excitation signals.

1.4.2 Select a Model Structure to Represent the System

A choice should be made within all the possible mathematical models that can be used to represent the system. Again a wide variety of possibilities exist, such as

- Parametric versus nonparametric models

  In a parametric model, the system is described using a limited number of characteristic quantities called the parameters of the model, whereas in a nonparametric model the system is characterized by measurements of a system function at a large number of points. Examples of parametric models are the transfer function of a filter described by its poles and zeros and the motion equations of a piston. An example of a nonparametric model is the description of a filter by its impulse response at a large number of points.

Usually it is simpler to create a nonparametric model than a parametric one because the modeler needs less knowledge about the system itself in the former case. However, physical insight and concentration of information are more substantial for parametric models than for nonparametric ones. In this book we will concentrate on transfer function models (parametric models), but the problem of frequency response function measurements (nonparametric model) will also be elaborated.

- White box models versus black box models

  In the construction of a model, physical laws whose availability and applicability depend on the insight and skills of the experimenter can be used (Kirchhoff's laws, Newton's laws, etc.). Specialized knowledge related to different scientific fields may be brought into this phase of the identification process. The modeling of a loudspeaker, for example, requires extensive understanding of mechanical, electrical, and acoustical phenomena. The result may be a physical model, based on comprehensive knowledge of the internal functioning of the system. Such a model is called a white box model.

  Another approach is to extract a black box model from the data. Instead of making a detailed study and developing a model based upon physical insight and knowledge, a mathematical model is proposed that allows sufficient description of any observed input and output measurements. This reduces the modeling effort significantly. For example, instead of modeling the loudspeaker using physical laws, an input-output relation, taking the form of a high-order transfer function, could be proposed.

  The choice between the different methods depends on the aim of the study: the white box approach is better for gaining insight into the working principles of a system, but a black box model may be sufficient if the model will be used only for prediction of the output.

  Although, as a rule of thumb, it is advisable to include as much prior knowledge as possible during the modeling process, it is not always easy to do it. If we know, for example, that a system is stable, it is not simple to express this information if the polynomial coefficients are used as parameters.

- Linear models versus nonlinear models

  In real life, almost every system is nonlinear. Because the theory of nonlinear systems is very involved, these are mostly approximated by linear models, assuming that in the operation region the behavior can be linearized. This kind of approximation makes it possible to use simple models without jeopardizing properties that are of importance to the modeler. This choice depends strongly on the intended use of the model. For example, a nonlinear model is needed to describe the distortion of an amplifier, but a linear model will be sufficient to represent its transfer characteristics if the linear behavior is dominant and is the only interest.

- Linear-in-the-parameters versus nonlinear-in-the-parameters

  A model is called linear-in-the-parameters if there exists a linear relation between these parameters and the error that is minimized. This does not imply that the system itself is linear. For example, $\varepsilon = y - (a_1 u + a_2 u^2)$ is linear in the parameters a_1 and a_2 but describes a nonlinear system. On the other hand,

$$\varepsilon(j\omega) = Y(j\omega) - \frac{a_0 + a_1 j\omega}{b_0 + b_1 j\omega} U(j\omega)$$

describes a linear system but the model is nonlinear in the b_1 and b_2 parameters. Linearity in the parameters is a very important aspect of models because it has a strong impact on the complexity of the estimators if a (weighted) least squares cost function is used. In that case, the problem can be solved analytically for models that are linear in the parameters so that an iterative optimization problem is avoided. This is illustrated in Section 1.5.1.

1.4.3 Match the Selected Model Structure to the Measurements

Once a model structure is chosen (e.g., a parametric transfer function model), it should be matched as well as possible with the available information about the system. Mostly, this is done by minimizing a criterion that measures a goodness of the fit. The choice of this criterion is extremely important because it determines the stochastic properties of the final estimator. As seen from the resistance example, many choices are possible and each of them can lead to a different estimator with its own properties. Usually, the cost function defines a distance between the experimental data and the model. The cost function can be chosen on an ad hoc basis using intuitive insight, but there also exists a more systematic approach based on stochastic arguments as explained in Section 1.5. Simple tests on the cost function exist (necessary conditions) to check even before deriving the estimator whether it can be consistent (see Chapter 7, Section 7.5).

1.4.4 Validate the Selected Model

Finally, the validity of the selected model should be tested: does this model describe the available data properly or are there still indications that some of the data are not well modeled, indicating remaining model errors? In practice, the best model (= the smallest errors) is not always preferred. Often a simpler model that describes the system within user-specified error bounds is preferred. Tools will be provided that guide the user through this process by separating the remaining errors into different classes, for example, unmodeled linear dynamics and nonlinear distortions. From this information, further improvements of the model can be proposed, if necessary.

During the validation tests it is always important to keep the application in mind. The model should be tested under the same conditions as it will be used later. Extrapolation should be avoided as much as possible. The application also determines what properties are critical.

1.4.5 Conclusion

This brief overview of the identification process shows that it is a complex task with a number of interacting choices. It is important to pay attention to all aspects of this procedure, from the experiment design to the model validation, in order to get the best results. The reader should be aware that besides this list of actions other aspects are also important. A short inspection of the measurement setup can reveal important shortcomings that can jeopardize a lot of information. Good understanding of the intended applications helps to set up good experiments, and is very important to make the proper simplifications during the model-building process. Many times, choices are made that are not based on complicated theories but are dictated by the practical circumstances. In these cases a good theoretical understanding of the applied methods will help the user to be aware of the sensitive aspects of his techniques. This will enable him to put all his effort on the most critical decisions. Moreover, he will become aware of the weak points of the final model.

1.5 A STATISTICAL APPROACH TO THE ESTIMATION PROBLEM

In the previous sections it was shown that an intuitive approach to a parameter estimation problem can cause serious errors without even being noticed. To avoid severe mistakes, a theoretical framework is needed. Here a statistical development of the parameter estimation theory is made. Four related estimators are studied: the least squares (LS) estimator, weighted least squares (WLS) estimator, maximum likelihood (ML) estimator, and, finally, the Bayes estimator. It should be clear that, as mentioned before, it is still possible to use other estimators, such as the least absolute values. However, a comprehensive overview of all possible techniques is beyond the scope of this book.

To use the Bayes estimator, the a priori probability density function (pdf) of the unknown parameters and the pdf of the noise on the measurements are required. Although it seems, at first, quite strange that the parameters have a pdf, we will illustrate in the next section that we use this concept regularly in daily life. The ML estimator requires only knowledge of the pdf of the noise on the measurements, and the WLS estimator can be applied optimally if the covariance matrix of the noise is known. Even if this information is lacking, the LS method is usable. Each of these estimators will be explained in more detail and illustrated in the following sections.

1.5.1 Least Squares Estimation

One of the simplest estimation techniques is the least squares estimator. In this case, the match between the model and the measurements is quantified by a least squares cost function. As this is an arbitrary choice, initially, it is clear that the result is not necessarily optimal. By choosing other cost functions such as the sum of the least absolute values, it is possible to find other estimators, with different properties, that perform better in specific situations. Some of these are studied explicitly in the literature. In this book we concentrate on least squares, a choice strongly motivated by numerical aspects: minimizing a least squares cost function is usually less involved than the alternative cost functions. Later on, this choice will also be shown to be motivated from the stochastic point of view. Normally distributed noise leads, naturally, to least squares estimation. As seen in the resistance example, even within the class of least squares estimators, there are different possibilities resulting in completely different estimators. A full treatment of the problem is beyond the scope of this book, hence, we focus only on the aspects that are of direct importance to our major goal.

Consider a multiple input, single output system modeled by $y_0(k) = g(u_0(k), \theta_0)$ with k the measurement index, $y(k) \in \mathbb{R}$, $u_0(k) \in \mathbb{R}^{1 \times n_u}$, and $\theta_0 \in \mathbb{R}^{n_\theta}$ the true parameter vector. The aim is to estimate the parameters from noisy observations at the output of the system: $y(k) = y_0(k) + n_y(k)$. This is done by minimizing the sum of the squared errors $e(k, \theta) = y(k) - y(k, \theta)$, with $y(k, \theta)$ the modeled output:

$$\hat{\theta}_{\mathrm{NLS}}(N) = \arg\min_{\theta} V_{\mathrm{NLS}}(\theta, N), \text{ with } V_{\mathrm{NLS}}(\theta, N) = \sum_{k=1}^{N} e^2(k, \theta) \tag{1-32}$$

In general, the analytical solution of the nonlinear least squares problem (1-32) is not known, so numerical methods must be used. A whole bunch of techniques are described in the literature (Fletcher, 1991), and many of them are available in mathematical packages that are commercially available. They vary from very simple techniques such as simplex methods that require no derivatives at all, through gradient or steepest descent methods (based on first-order derivatives), to Newton methods that make use of second-order derivatives. The optimal choice strongly depends on the specific problem. However, the Gauss-Newton method is very well suited to deal with the least squares minimization problem because it makes explicit use of the structure of the cost function. The second derivatives of the cost function (the Hessian matrix) are approximated in this method by the first-order derivatives of $e(\theta)$. Define the Jacobian matrix $J(\theta) \in \mathbb{R}^{N \times n_\theta}$: $J(\theta) = \partial e(\theta)/\partial \theta$ and consider the Hessian matrix:

$$\frac{\partial^2 V_{\mathrm{NLS}}(\theta, N)}{\partial \theta^2} = J^T(\theta) J(\theta) - \frac{1}{N} \sum_{k=1}^{N} e(k, \theta) \frac{\partial^2 g(u_0(k), \theta)}{\partial \theta^2} \tag{1-33}$$

If the second term in Eq. (1-33) is small (for example, $\|e(\theta)\|_2$ is "small") with respect to the first one, then $J^T(\theta)J(\theta)$ will be a good approximation for the second-order derivatives of the cost function. The numerical solution is then found by applying the following iterative process:

$$\theta^{(i+1)} = \theta^{(i)} + \Delta\theta^{(i+1)} \text{ with } J^T(\theta^{(i)}) J(\theta^{(i)}) \Delta\theta^{(i+1)} = -J^T(\theta^{(i)}) e(\theta^{(i)}) \tag{1-34}$$

Equation (1-34) reveals two important advantages. First, only the gradient needs to be calculated, and not the Hessian, thus reducing the calculation time. Moreover, very often, the condition number of the Hessian matrix is the square of that of the Jacobian. This leads us to the second advantage: using, for example, singular value decomposition (SVD) or QR decomposition techniques, Eq. (1-34) can be solved without forming the product $J^T(\theta^{(i)})J(\theta^{(i)})$ so that more complex problems can be solved, because the numerical errors are significantly reduced (see Exercise 1.11). If (1-34) converges to the global minimum of (1-32), then $\hat{\theta}_{\mathrm{NLS}}(N) = \theta^{(\infty)}$.

Because there are no explicit expressions available for the estimator as a function of the measurements, it is not straightforward to study its properties. For this reason, special theories are developed to analyze the properties of the estimator by analyzing the cost function. These techniques are covered in detail in Section 17.4 . Under quite general assumptions on the noise (for example, iid noise with finite second- and fourth-order moments), some regularity conditions on the model $g(u_0(k), \theta)$, and the excitation (choice of $u_0(k)$), consistency of the least squares estimator is proved. Also, an approximate expression for the covariance matrix $\mathrm{Cov}(\hat{\theta}_{\mathrm{NLS}}(N))$ is available:

$$\mathrm{Cov}(\hat{\theta}_{\mathrm{NLS}}(N)) \approx (J^T(\theta)J(\theta))^{-1} J^T(\theta) \mathrm{Cov}(n_y) J(\theta) (J^T(\theta)J(\theta))^{-1} \Big|_{\theta = \hat{\theta}_{\mathrm{LS}}(N)} \tag{1-35}$$

with $\mathrm{Cov}(n_y) = \mathcal{E}\{n_y n_y^T\}$. Note that this approximation is still a stochastic variable because it depends on $\hat{\theta}_{\mathrm{NLS}}(N)$, while the exact expression should be in θ_0. If the model is linear-in-the-parameters, $y_0 = K(u_0)\theta_0$, and $e(\theta) = y - K(u_0)\theta$), then (1-32) reduces to a linear least squares cost function, and explicit expressions are available for the estimator (note that $K = -\partial e(\theta)/\partial \theta = -J(\theta)$ is parameter independent in this case). In order to keep the expressions compact, we do not include the arguments of K in the following.

$$\hat{\theta}_{\mathrm{LS}}(N) = (K^T K)^{-1} K^T y \tag{1-36}$$

The covariance matrix still equals (1-35) with $J(\hat{\theta}_{\mathrm{LS}}(N))$ replaced by $-K$, but now it is an exact expression and no longer an approximation. Moreover, it is possible to prove that the estimator is unbiased for zero mean noise:

$$\mathcal{E}\{\hat{\theta}_{\mathrm{LS}}(N)\} = (K^TK)^{-1}K^T\mathcal{E}\{y\} = (K^TK)^{-1}K^Ty_0 = (K^TK)^{-1}K^TK\theta_0 = \theta_0 \tag{1-37}$$

This result is valid only if K is not disturbed by noise. If the inputs u are also disturbed by noise, it is no longer possible to bring $(K^TK)^{-1}K^T$ outside the expectation. In this case, additional quadratic noise contributions appear in K^TK so that $\hat{\theta}_{\mathrm{LS}}(N)$ underestimates the true values. This was visible in the estimation of the resistance ($K_{[k]} = i(k)$, $y(k) = u(k)$, $\theta = R$) where Eq. (1-8) shows the impact of the quadratic contributions of the input noise.

Example 1.7 (Weighing a Loaf of Bread): John is asked to estimate the weight of a loaf of bread from N noisy measurements $y(k) = \theta_0 + n_y(k)$ with θ_0 the true but unknown weight, $y(k)$ the weight measurement, and $n_y(k)$ the measurement noise. From a prior analysis, making repeated measurements, it turns out that $n_y(k)$ is zero mean iid with variance σ_y^2. The model becomes $y = K\theta + n_y$ with $K = (1, 1, \ldots, 1)^T$. Using (1-36), the estimate is

$$\hat{\theta}_{\mathrm{LS}}(N) = (K^TK)^{-1}K^Ty = \frac{1}{N}\sum\nolimits_{k=1}^{N} y(k) \tag{1-38}$$

with variance

$$\mathrm{var}(\hat{\theta}_{\mathrm{LS}}(N)) = (K^TK)^{-1}K^T(\sigma_y^2 I_N)K(K^TK)^{-1} = \sigma_y^2/N \tag{1-39}$$

This example shows that it is much easier to get the solution when it is possible to formulate the problem under the standard conditions. □

This short analysis shows that the least squares estimator is applicable to a very wide range of problems. No prior information is required to use it, which explains its success. However, its specific properties depend on the actual situation. General statements can be made only if some noise characteristics are known. In that case it is also possible to improve the quality of the estimates by using this knowledge in the estimator. If, for example, the covariance matrix of the noise is known, a weighted least squares can be used.

1.5.2 Weighted Least Squares Estimation

In Eq. (1-32) all measurements are equally weighted. In many problems it is desirable to put more emphasis on one measurement with respect to the other. This can be done to make the difference between measurements and model smaller in some regions, but it can also be motivated by stochastic arguments. If the covariance matrix of the noise is known, then it seems logical to suppress measurements with high uncertainty and to emphasize those with low uncertainty. In practice, it is not always clear what weighting should be used. If it is, for example, known that model errors are present, then the user may prefer to put in a dedicated weighting in order to keep the model errors small in some specific operation regions instead of using the weighting dictated by the covariance matrix.

In general, the weighted nonlinear least squares estimate $\hat{\theta}_{\text{WNLS}}(N)$ is

$$\hat{\theta}_{\text{WNLS}}(N) = \arg\min_{\theta} V_{\text{WNLS}}(\theta, N) \text{ with } V_{\text{WNLS}}(\theta, N) = e^T(\theta)We(\theta) \tag{1-40}$$

where $W \in \mathbb{R}^{N \times N}$ is a symmetric positive definite weighting matrix (the asymmetric part does not contribute to a quadratic form). The evaluation of this cost function requires $O(N^2)$ operations, which are very time consuming. Consequently, (block) diagonal weighting matrices are preferred in many problems, reducing the number of operations to $O(N)$. All the remarks on the numerical aspects of the least squares estimator are also valid for the weighted least squares. This can be understood easily by applying the following transformation: $\varepsilon(\theta) = Se(\theta)$ with $S^TS = W$ so that $V_{\text{WNLS}}(\theta, N) = \varepsilon^T(\theta)\varepsilon(\theta)$, which is a least squares estimator in the transformed variables. This also leads to the following Gauss-Newton algorithm to minimize the cost function

$$\theta^{(i+1)} = \theta^{(i)} + \Delta\theta^{(i+1)} \text{ with } J^T(\theta^{(i)})WJ(\theta^{(i)})\Delta\theta^{(i+1)} = -J^T(\theta^{(i)})We(\theta^{(i)}) \tag{1-41}$$

Equation (1-35) is generalized to (noticing that $W^T = W$)

$$\text{Cov}(\hat{\theta}_{\text{WNLS}}(N)) \approx (J^T(\theta)WJ(\theta))^{-1}J^T(\theta)WC_{n_y}WJ(\theta)(J^T(\theta)WJ(\theta))^{-1}\Big|_{\theta = \hat{\theta}_{\text{WNLS}}(N)} \tag{1-42}$$

with $C_{n_y} = \text{Cov}(n_y)$. By choosing $W = C_{n_y}^{-1}$, the expression simplifies to

$$\text{Cov}(\hat{\theta}_{\text{WNLS}}(N)) \approx [J^T(\hat{\theta}_{\text{WNLS}}(N))C_{n_y}^{-1}J(\hat{\theta}_{\text{WNLS}}(N))]^{-1} \tag{1-43}$$

In Exercise 1.15 it is shown that among all possible positive definite choices for W, the best one is $W = C_{n_y}^{-1}$ because this minimizes the covariance matrix. The results for models that are linear-in-the-parameters are immediately found, analogous to the least squares estimator. Also, in this case, the weighted least squares is unbiased under the same conditions as the least squares estimator.

1.5.3 The Maximum Likelihood Estimator

Using the covariance matrix of the noise as the weighting matrix allows prior knowledge about the noise on the measurements. However, a full stochastic characterization requires the pdf of the noise distortions. If this knowledge is available, it may be possible to get better results than those attained with a weighted least squares. Maximum likelihood estimation offers a theoretical framework to incorporate the knowledge about the distribution in the estimator. The pdf f_{n_y} of the noise also determines the conditional pdf $f(y|\theta_0)$ of the measurements, given the hypothetical exact model, $y_0 = G(u_0, \theta_0)$, that describes the system and the inputs that excite the system. Assuming, again, an additive noise model $y = y_0 + n_y$, with $y, y_0, n_y \in \mathbb{R}^N$, the likelihood function becomes:

$$f(y|\theta_0, u_0) = f_{n_y}(y - G(u_0, \theta_0)) \tag{1-44}$$

The maximum likelihood procedure consists of two steps. First the numerical values y_m of the actual measurements are plugged into expression (1-44) for the variables y, and next the model parameters θ_0 are considered as the free variables. This results in the so-called likelihood function. The maximum likelihood estimate is then found as the maximizer of the likelihood function

$$\hat{\theta}_{\mathrm{ML}}(N) = \arg\max_{\theta} f(y_m | \theta, u_0) \tag{1-45}$$

From now on, we will no longer explicitly indicate the numerical values y_m but just use the symbol y for the measured values.

Example 1.8 (Weighing a Loaf of Bread—Continued): Consider Example 1.7 again, but assume that more information about the noise is available. This time John knows that the distribution f_y of n_y is normal with zero mean and standard deviation σ_y. With this information he can build an ML estimator:

$$f(y|\theta) = \frac{1}{\sqrt{2\pi\sigma_y^2}} e^{-\frac{(y-\theta)^2}{2\sigma_y^2}} \tag{1-46}$$

and the estimated weight becomes $\hat{\theta}_{\mathrm{ML}} = y$. It is therefore not possible to give a better estimate than the measured value itself. If John makes repeated independent measurements $y(1), \ldots, y(N)$, the likelihood function is

$$f(y|\theta) = \frac{1}{(2\pi\sigma_y^2)^{N/2}} e^{-\frac{1}{2\sigma_y^2}\sum_{k=1}^{N}(y(k)-\theta)^2} \tag{1-47}$$

The ML estimate is given by the minimizer of $\frac{1}{2\sigma_y^2}\sum_{k=1}^{N}(y(k)-\theta)^2$ ($(2\pi\sigma_y^2)^{-N/2}$ is parameter independent) and becomes

$$\hat{\theta}_{\mathrm{ML}}(N) = \frac{1}{N}\sum_{k=1}^{N} y(k) \tag{1-48}$$

This is nothing other than the sample mean of the measurements. It is again easy to check that this estimate is unbiased. Note that in this case the ML estimator and the (weighted) least squares estimator are the same. This is the case only for normally distributed errors. □

The unbiased behavior may not be generalized because the MLE can also be biased. For example, the sample mean and sample variance are shown to be the ML estimates for the mean and the variance of measurements that are identically independent and normally distributed: $\hat{\mu}_{\mathrm{ML}} = \frac{1}{N}\sum_{k=1}^{N} y(k)$, $\hat{\sigma}_{\mathrm{ML}}^2 = \frac{1}{N}\sum_{k=1}^{N}(y(k) - \hat{\mu}_{\mathrm{ML}})^2$. Although the first estimate is unbiased, the second one can be shown to be prone to a bias of σ^2/N that asymptotically disappears in N: $\mathcal{E}\{\hat{\sigma}_{\mathrm{ML}}\} = \sigma^2(N-1)/N$. This shows that there is a clear need to understand the properties of the ML estimator better. In the literature, a series of important properties is tabled assuming well-defined experimental conditions. Each time these conditions are met, the user knows in advance, before passing through the complete development process, what the properties of the estimator would be. On the other hand, if the conditions are not met, nothing is guaranteed anymore and a dedicated analysis is, again, required. In this introductionary chapter we just make a loose statement of the properties; a very precise description can be found in the literature (Goodwin and Payne, 1977; Caines, 1988).

Properties of the ML Estimator

- *Principle of invariance:* if $\hat{\theta}_{\mathrm{ML}}$ is an ML estimator of $\theta \in \mathbb{R}^{n_\theta}$, then $\hat{\theta}_g = g(\hat{\theta}_{\mathrm{ML}})$ is an ML estimator of $g(\theta)$ where g is a function, $\hat{\theta}_g \in \mathbb{R}^{n_g}$ and $n_g \le n_\theta$, with n_θ a finite number.
- *Consistency:* if $\hat{\theta}_{\mathrm{ML}}(N)$ is an ML estimator based on N iid random variables, with n_θ independent of N, then $\hat{\theta}_{\mathrm{ML}}(N)$ converges to θ_0 almost surely: $\underset{N \to \infty}{\text{a.s.lim}}\, \hat{\theta}_{\mathrm{ML}}(N) = \theta_0$.

  If n_θ depends on N, the property is no longer valid, and the consistency should be checked again. See, for example, the errors-in-variables estimator in the previous section where not only is the resistance value estimated, but also the currents $i(1), \ldots, i(N)$ and voltages $u(1), \ldots, u(N)$ (In this case $n_\theta = N + 1$, e.g., the N current values and the unknown resistance value, and the voltage is calculated from the estimated current and resistance value).
- *Asymptotic normality:* if $\hat{\theta}_{\mathrm{ML}}(N)$ is an ML estimator based on N iid random variables, with n_θ independent of N, then $\hat{\theta}_{\mathrm{ML}}(N)$ converges in law to a normal random variable.

  The importance of this property is that it not only allows one to calculate uncertainty bounds on the estimates but also guarantees that most of the probability mass gets more and more unimodally concentrated around its limiting value.
- *Asymptotic efficiency:* if $\hat{\theta}_{\mathrm{ML}}(N)$ is an ML estimator based on N iid random variables, with n_θ independent of N, then $\hat{\theta}_{\mathrm{ML}}(N)$ is asymptotically efficient $(\mathrm{Cov}(\hat{\theta}_{\mathrm{ML}}(N))$ reaches asymptotically the Cramér-Rao lower bound).

1.5.4 The Bayes Estimator

As described before, the Bayes estimator requires the most prior information before it is applicable, namely the pdf of the noise on the measurements and the pdf of the unknown parameters. The kernel of the Bayes estimator is the conditional pdf of the unknown parameters θ with respect to the measurements y: $f(\theta | u, y)$. This pdf contains complete information about the parameters θ, given a set of measurements y. This makes it possible for the experimenter to determine the best estimate of θ for the given situation. To select this best value, it is necessary to lay down an objective criterion, for example, the minimization of a risk function $C(\theta | \theta_0)$ that describes the cost of selecting the parameters θ if θ_0 are the true but unknown parameters. The estimated parameters $\hat{\theta}$ are found as the minimizers of the risk function weighted with the probability $f(\theta | u, y)$:

$$\hat{\theta}(N) = \arg\min_{\theta_0} \int_{\theta \in \mathscr{D}} C(\theta | \theta_0) f(\theta | u, y) d\theta \tag{1-49}$$

For some specific choices of $C(\theta | \theta_0)$, the solution of expression (1-49) is well known; for example, $C(\theta | \theta_0) = |\theta - \theta_0|^2$ leads to the mean value, and $C(\theta | \theta_0) = |\theta - \theta_0|$ results in the median, which is less sensitive to outliers because these contribute less to the second criterion than to the first (Eykhoff, 1974).

Another objective criterion is to choose the estimate as

$$\hat{\theta}_{\text{Bayes}}(N) = \arg\max_{\theta} f(\theta|u, y) \tag{1-50}$$

The first and second examples are "minimum risk" estimators, and the last is the Bayes estimator. In practice, it is very difficult to select the best out of these. In the next section, we study the Bayes estimator in more detail. To search for the maximizer of (1-50) the Bayes rule is applied:

$$f(\theta|u, y) = \frac{f(y|\theta, u)f(\theta)}{f(y)} \tag{1-51}$$

In order to maximize the right-hand side of this equation it is sufficient to maximize its numerator, because the denominator is independent of the parameters θ, so that the solution is given by looking for the maximum of $f(y|\theta, u)f(\theta)$. This simple analysis shows that a lot of a priori information is required to use the Bayes estimator: $f(y|\theta, u)$ (also appearing in the ML estimator) and $f(\theta)$. In many problems the parameter distribution $f(\theta)$ is unavailable, and this is one of the main reasons why the Bayes estimator is rarely used in practice (Norton, 1986).

Example 1.9 (Use of the Bayes Estimator in Daily Life): We commonly use some important principles of the Bayes estimator without being aware of it. This is illustrated in the following story: Joan was walking at night in Belgium and suddenly saw a large animal in the far distance. She decided that it was either a horse or an elephant Prob(observationlelephant) = Prob(observationlhorse). However, the probability of seeing an elephant in Belgium is much lower than that of seeing a horse: Prob(elephant in Belgium) « Prob(horse in Belgium) so that from the Bayes principle Joan concludes she was seeing a horse. If she was on safari in Kenya instead of Belgium, the conclusion would be opposite, because Prob(elephant in Kenya) » Prob(horse in Kenya).

Joan continued her walk. When she came closer she saw that the animal had big feet, a small tail, and also a long trunk so that she had to review her previous conclusion on the basis of all this additional information: there was an elephant walking on the street. When she passed the corner, she saw that a circus had arrived in town. □

From the previous example it is clear that in a Bayes estimator the prior knowledge of the pdf of the estimated parameters is very important. It also illustrates that it balances our prior knowledge with the measurement information. This is more quantitatively illustrated in the next example.

Example 1.10 (Weighing a Loaf of Bread—Continued): Consider again Example 1.8 but assume this time that the baker told John that the bread normally weighs about $w = 800$ g. However, the weight can vary around this mean value as a result of humidity, the temperature of the oven, and so on, in a normal way with a standard deviation σ_w. With all this information, John knows enough to build a Bayes estimator. Using normal distributions and noticing that $f(y|\theta) = f_y(n_y) = f_y(y-\theta)$, the Bayes estimator is found by maximizing

$$f(y|\theta)f(\theta) = \frac{1}{\sqrt{2\pi\sigma_y^2}}e^{-\frac{(y-\theta)^2}{2\sigma_y^2}}\frac{1}{\sqrt{2\pi\sigma_w^2}}e^{-\frac{(\theta-w)^2}{2\sigma_w^2}} \tag{1-52}$$

and the estimated weight becomes

$$\hat{\theta}_{\text{Bayes}} = \frac{y/\sigma_y^2 + w/\sigma_w^2}{1/\sigma_y^2 + 1/\sigma_w^2} \tag{1-53}$$

In this result, two parts can be distinguished: y, the information derived from the measurement, and w, the a priori information from the baker. If the quality of the prior information is high compared with that of the measurements ($\sigma_w \ll \sigma_y$), the estimate is determined mainly by the prior information. If the quality of the prior information is very low compared with the measurements ($\sigma_w \gg \sigma_y$), the estimate is determined mainly by the information from the measurements.

After making several independent measurements $y(1), \ldots, y(N)$ the Bayes estimator becomes

$$\hat{\theta}_{\text{Bayes}}(N) = \frac{\sum_{k=1}^{N} y(k)/\sigma_y^2 + w/\sigma_w^2}{N/\sigma_y^2 + 1/\sigma_w^2} \tag{1-54}$$

The previous conclusions remain valid. However, when the number of measurements increases, the first term dominates the second one such that the impact of the prior information is reduced (Sörenson, 1980). Finally, when N becomes infinite, the estimate is completely determined by the measurements. □

Conclusion. From these examples it is seen that a Bayes estimator combines prior knowledge of the parameters with information from measurements. When the number of measurements is increased, the measurement information becomes more important and the influence of the prior information decreases. If there is no information about the distribution of the parameters, the Bayes estimator reduces to the ML estimator. If the noise is normally distributed, the ML estimator reduces to the weighted least squares. If the noise is white, the weighted least squares boils down to the least squares estimator.

1.5.5 Instrumental Variables

In this section we will discuss a final parameter estimation method that is very suitable when both the input and the output are disturbed by noise. Although it does not belong directly to the previous family of estimators, we include it in this chapter for use later, to interpret one of the proposed identification schemes. In the resistance estimation examples, it was shown that the least squares method $\hat{R}_{\text{LS}}(N)$ is biased because of the quadratic noise contributions appearing in the denominator:

$$\hat{R}_{\text{LS}}(N) = \frac{\frac{1}{N}\sum_{k=1}^{N} u(k)i(k)}{\frac{1}{N}\sum_{k=1}^{N} i^2(k)}, \text{ with } \lim_{N\to\infty}\hat{R}_{\text{LS}}(N) = R_0\frac{1}{1+\sigma_i^2/i_0^2} \tag{1-55}$$

This systematic error can be removed by replacing $i(k)$ in the numerator and denominator by $i(k-1)$ so that the new estimate becomes:

$$\hat{R}_{\mathrm{IV}}(N) = \frac{\frac{1}{N}\sum_{k=1}^{N} u(k)i(k-1)}{\frac{1}{N}\sum_{k=1}^{N} i(k)i(k-1)} \tag{1-56}$$

Making the same analysis as in Section 1.2.2.1, it is seen that all quadratic noise contributions are eliminated by this choice, so that

$$\lim_{N\to\infty} \hat{R}_{\mathrm{IV}}(N) = R_0 \tag{1-57}$$

The idea used to generate (1-56) can be generalized as follows. Consider the linear-in-the-parameters model structure $y_0 = K(u_0)\theta_0$ in Section 1.5.1, and replace K^T in Eq. (1-36) by G^T, to get

$$\hat{\theta}_{\mathrm{IV}}(N) = (G^T K(u))^{-1} G^T y \tag{1-58}$$

The choice of G, a matrix of the same size as $K(u)$, will be defined later. $\hat{\theta}_{\mathrm{IV}}(N)$ is the instrumental variables estimate. Consistency is proved by considering the plim for $N \to \infty$ (Norton, 1986). For simplicity, we assume all the plim exists, namely

$$\begin{aligned}\text{plim } \hat{\theta}_{\mathrm{IV}} &= \text{plim } \{(G^T K(u))^{-1} G^T y\} \\ &= (\text{plim } \{G^T K(u_0 + n_u)\})^{-1}(\ \text{plim}\{G^T y_0 + G^T n_y\}) \\ &= (\text{plim } \{G^T K(u_0 + n_u)/N\})^{-1}(\text{plim } \{G^T K(u_0)/N\}\theta_0 + \text{plim } \{G^T n_y/N\})\end{aligned}$$

If

$$\text{plim } \{G^T K(u_0 + n_u)/N\} = \text{plim } \{G^T K(u_0)/N\} \quad \text{and} \quad \text{plim } \{G^T n_y/N\} = 0 \tag{1-59}$$

then

$$\underset{N\to\infty}{\text{plim}}\, \hat{\theta}_{\mathrm{IV}}(N) = \theta_0 \tag{1-60}$$

Equation (1-59) defines the necessary conditions for G to get a consistent estimate. Loosely stated, G should not be correlated with the noise on $K(u_0 + n_u)$ and the output noise n_y. The variables used for building the entries of G are called the instrumental variables.

If the covariance for $C_{n_y} = \sigma^2 I_N$, then an approximate expression for the covariance matrix of the estimates is (Norton, 1986):

$$\text{Cov}(\hat{\theta}_{\mathrm{IV}}(N)) \approx \sigma^2 R_{GK}^{-1} R_{GG} R_{GK}^{-T} \text{ with } R_{GK} = G^T K(u)/N \text{ and } R_{GG} = G^T G/N \tag{1-61}$$

This reveals another condition on the choice of the instrumental variables G: although they should be "uncorrelated" with the noise on the output observation n_y, they should be correlated maximally with K, otherwise R_{GK} tends to zero and $\text{Cov}(\hat{\theta}_{\mathrm{IV}}(N))$ would become very large. In the case of the resistance estimate, the instrumental variables are the shifted input. Because we used a constant current, no problem arises. In practice, this technique can be generalized to varying inputs under the condition that the power spectrum of the noise is much wider than the power spectrum of the input. In the following Exercises the instrumental variables method is applied to the resistance example.

1.6 EXERCISES

1.1. Set up a simulation to measure the value of the resistance using

$$i(k) = i_0 + n_i(k) \qquad u(k) = u_0 + n_u(k) \tag{1-62}$$

Use for n_i and n_u zero mean iid noise with standard deviation σ_i and σ_u. Consider uniformly and normally distributed noise and use $i_0 = 1$ A, $u_0 = 1$ V, $\sigma_i = 0.5(1)$ A, and $\sigma_u = 0.5(1)$ V. Plot $R(k) = u(k)/i(k)$ for $k = 1, \ldots, 100$.

1.2. Apply the estimators $\hat{R}_{LS}$, $\hat{R}_{EV}$, $\hat{R}_{SA}$ from Eqs. (1-1) to (1-3) to the results of the simulator in Exercise 1.1 and plot the results as a number of the processed measurements N.

1.3. Measure the histogram for the three estimators of Exercise 1.2 for $N = 10, 100, 1000$ and plot the approximated pdf.

1.4. Use the simulator of Exercise 1.1 to estimate the variance of the three estimators of Exercise 1.2 as a function of N and plot the results on a log-log scale. Check the $1/\sqrt{N}$ rule of thumb. Vary N between 1 and 10,000.

1.5. Derive the variance expressions $\text{var}(\hat{R}_{LS}(N))$, $\text{var}(\hat{R}_{EV}(N))$, $\text{var}(\hat{R}_{SA}(N))$ under Assumption 1.1 using linear approximations as illustrated in Eqs. (1-16) and (1-17).

1.6. Use the simulator of Exercise 1.1 to estimate the variance of the three estimators of Exercise 1.2 for $N = 100$ as a function of the SNR of the current and the voltage measurements. Compare the results with the theoretical level (see Eqs. 1-17 and 1-18) and discuss the results.

1.7. Derive the estimators $\hat{R}_{LS}(N)$, $\hat{R}_{EV}(N)$, and $\hat{R}_{SA}(N)$ by minimizing the cost functions (1-20), (1-21), and (1-22).

1.8. Reformulate the cost functions (1-20), (1-21), and (1-22) for the case that the current is varying from measurement to measurement (the current is no longer a DC source), and derive the new expressions of the estimators.

1.9. Consider a signal

$$y_0(k) = \sin(2\pi f k T_s + \varphi) \tag{1-63}$$

and its measurement

$$y(k) = y_0(k) + n_y(k)\text{, for } k = 1, \ldots, 1024 \tag{1-64}$$

where $n_y(k)$ is iid normally distributed noise with zero mean and variance σ_y^2. Calculate the Cramér-Rao for the estimates (f, φ). What is the best choice for T_s if we want to estimate the frequency with minimum variance?

1.10. Consider a polynomial model:

$$y_0(k) = \sum_{p=1}^{P} a_p u^p(k) \tag{1-65}$$

that is identified from a set of measurements $y(k) = y_0(k) + n_y(k)$, with $u(k) = [-N{:}N]/N$ and $n_y(k)$ zero mean iid distributed noise with variance σ_y^2. Set up the least squares estimator for this problem, and observe the condition number for growing values of P (put $N = 1000$). What is the maximum order that can be reliably identified?

1.11. Consider the least squares solution $\hat{\theta}_{LS}(N) = (J^TJ)^{-1}J^Ty$ of the overdetermined set $J\theta = y$ (as they appear in Eq. 1-36). Show that this solution can be calculated using the SVD method of Section 13.5 on matrix algebra without forming the product J^TJ as $\hat{\theta}_{LS}(N) = J^+y$, with $J^+ = V\Sigma^+U^T$.

1.12. Apply the method of Exercise 1.11 to the polynomial problem of Exercise 1.10, and find the maximum order that can be identified reliably.

1.13. The polynomial identification problem is an ill-posed problem because of the poor numerical conditioning of the normal equations. Using the SVD method, it is already possible to solve higher order problems, but even then the numerical conditioning decreases fast. A much better solution is to change the model representation and to use orthogonal polynomials $T_p(u)$ such that

$$y_0(k) = \sum\nolimits_{p=1}^{P} a_p u^p(k) = \sum\nolimits_{p=1}^{P} t_p T_p(u(k)) \tag{1-66}$$

where $T_p(u) = \sum_{k=1}^{p} a_{pk}u^k$ is a polynomial of degree p. The coefficients a_{pk} are set s.t. $\sum_{k=1}^{N} T_r(u(k))T_s(u(k)) = \delta_{rs}$. Note that the actual form of $T_p(u)$ (the choice of a_{pk}) depends on the set of input values $u(k)$ that appears in the problem. Reformulate the polynomial identification problem using the orthogonal basis and discuss the condition number of the new estimator.

Remarks:

For the given set of input values, the orthogonal polynomials $T_p(u)$ are given by the following recurrence relation (Ralston and Rabinowitz, 1984):

$$\begin{aligned} &\frac{1}{\alpha_{j+1}}T_{j+1}(u) = \frac{u}{\alpha_j}T_j(u) - \frac{\beta_j}{\alpha_{j-1}}T_{j-1}(u) \\ &\beta_j = \frac{j^2[(2N+1)^2 - j^2]}{4(4j^2-1)}, \qquad \alpha_j = \frac{(2j)!}{(j!)^2(2N)^j} \end{aligned} \tag{1-67}$$

with $T_0(u) = 1$ and $T_{-1}(u) = 0$ for $j = 0, 1, \ldots$.

When using orthogonal polynomials the reader should take care not to use the explicit polynomial expressions, but only the values of the orthogonal polynomials. Otherwise the numerical stability is not guaranteed. As a result, it is also not possible to calculate the coefficients a_p of the original solution; only the value of the solution can be calculated (see Ralston and Rabinowitz, 1984).

1.14. Prove expression (1-42) for the covariance matrix of a weighted least squares for models that are linear-in-the-parameters.

1.15. Show that the covariance matrix of the weighted least squares estimator becomes minimal for $W = C_{n_y}^{-1}$ (hint: use the Schwarz inequality $B^TB \geq (B^TA)(A^TA)^{-1}(A^TB)$, see Eykhoff, 1974, p. 525, and put $C_{n_y}^{-1} = C^TC$, $B = CJ$, and $A = C^{-T}WJ$).

1.16. Consider the linear-in-the-parameters model $y_0 = K(u_0)\theta_0$ and calculate the variance of the modeled output $\hat{y} = K(u_0)\hat{\theta}$ starting from the covariance matrix C_θ given in (1-43).

1.17. Show that the variance on the output of the polynomial model in Exercise 1.10 is independent of the model representation $y_0(k) = \sum_{p=1}^{P} a_p u^p(k)$ or $y_0(k) = \sum_{p=1}^{P} t_p T_p(u(k))$. Check this by a simulation using the estimators of Exercises 1.10 and 1.13 for a polynomial of degree 5 (so that the numerical conditioning of the problem remains acceptable for the direct estimation).

1.18. Consider the system $y_0 = au$. Construct the least squares and the weighted least squares estimator for a starting from the measurements $y(k) = au(k) + n_y(k)$ with $\mathcal{E}\{n_y(k)\} = 0$ and $\sigma^2_{n_y}(k) = u(k)$. Compare the bias and the variance of both estimators for $u(k) = 1, 2, \ldots, 10$. Verify your results by means of a simulation.

1.19. Construct $\hat{R}_{\mathrm{IV}}(N)$ for the resistance example of Section 1.2.1 using Eq. (1-58). Use the time-shifted current as an instrumental variable. Study the behavior of the estimator (mean value and variance) as a function of the shift by means of a simulation.

1.20. Study the behavior of $\hat{R}_{\mathrm{IV}}(N)$ (mean value and variance) of the previous exercise for the situation where $i_0(k)$ is generated as low-pass filtered noise (bandwidth of the filter at $f_s/50$) as a function of the applied delay by means of a simulation.

2

Measurements of Frequency Response Functions

Abstract: Frequency response function (FRF) measurements are an interesting intermediate step in the identification process. The complexity of the modeling problem is visualized before starting the parametric modeling; the quality of the measurements is assessed in an early phase. In this chapter a number of basic and advanced FRF measurement methods are discussed. An analysis of the bias and efficiency of the FRF measurements is made, and their dependence on the experimental conditions and on the excitation signal is analyzed. Simple and more advanced averaging techniques are proposed to improve the quality of the FRF measurement. Guidelines given at the critical steps of the FRF measurement process enable the less experienced user to start modeling from good raw data.

2.1 INTRODUCTION

Consider the linear dynamic system $G(j\omega)$ between the input $u(t)$ and the output $y(t)$ as shown in Figure 2-1. The aim of this book is to build a parametric model for this system, identifying, for example, a transfer function $G(j\omega, \theta)$. Such a model is called a parametric because it employs a finite-dimensional parameter vector. Parametric modeling requires a series of user decisions (e.g., selection of the order of numerator and denominator of $G(s)$), thus it is strongly advised to get a good initial idea about the system under test. Step or impulse response measurements provide this information. Also, frequency response function (FRF) measurements are very valuable. An FRF consists of transfer function measurements $G(j\omega_k)$ at a discrete set frequencies ω_k, $k = 1, ..., F$. All these models are called nonparametric because the information is not condensed into a small set of parameters. In this chapter we focus, exclusively, on FRF measurements. A series of basic questions is addressed:

- How are the bias and efficiency of the FRF measurements influenced by the experimental conditions?
- How should the excitation signal be chosen?
- Can we improve the quality of the FRF using averaging methods?
- Can we quantify the quality of the FRF measurements?

System Identification: A Frequency Domain Approach. Rik Pintelon and Johan Schoukens

- What is the impact of nonlinear distortions on the measured FRF and how can we detect their existence?

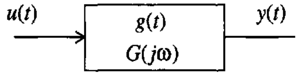


Figure 2-1. Block diagram of the system.

All these aspects are discussed here or in the next chapter for the last question. Starting from a straightforward solution, the more advanced techniques are introduced step by step, showing each time what additional problems are addressed by these more advanced techniques. As FRF measurement techniques rely heavily on the transformation of sampled signals from the time to the frequency domain, we will spend some time on the most important aspects of the discrete Fourier transform.

2.2 AN INTRODUCTION TO THE DISCRETE FOURIER TRANSFORM

In most situations, real-life systems are naturally continuous in time. However, most signal processing is now done on digital computers that operate on discrete-time signals. In practice the continuous-time signals are discretized (sampled) and quantized (digitized) so that the signal can finally be stored in the memory of a digital computer. Next, the spectrum of these signals is needed in order to calculate the FRF of the system. This is done using the discrete Fourier transform (DFT), usually calculated with the fast Fourier transform (FFT) algorithm. Each of these steps creates errors, and it is important for a user to understand their behavior to minimize the impact of the errors on his results. In this section only a brief introduction is given. For an extended overview, the reader is referred to Brigham's (1974) book. First we discuss, briefly, the sampling process, next we show how to "measure" the Fourier spectrum of a signal, and finally we focus on the spectral properties of periodic excitations and how to exploit them to minimize the measurement errors.

2.2.1 The Sampling Process

The continuous-time signal is sampled at an equidistant time grid and is represented by the equivalent discrete-time sequence $u_{\mathrm{d}}(n) = u(nT_{\mathrm{s}})$. In the time domain, the sampling process can be formulated as a multiplication with a periodically repeated Dirac impulse (Brigham, 1974):

$$\tilde{u}_{\mathrm{d}}(t) = u(t)\delta_{T_{\mathrm{s}}}(t) \text{ with } \delta_{T_{\mathrm{s}}}(t) = \sum\nolimits_{n=-\infty}^{\infty}\delta(t-nT_{\mathrm{s}}) \tag{2-1}$$

Note that in this framework the discrete-time signal $u_{\mathrm{d}}(n)$ is formally represented by a continuous-time signal $\tilde{u}_{\mathrm{d}}(t)$ that carries all its power at the discrete-time instances nT_{s}. Define the spectrum of the discrete-time signal as

$$U_{\mathrm{d}}(e^{j2\pi fT_{\mathrm{s}}}) = \sum\nolimits_{n=-\infty}^{\infty} u_{\mathrm{d}}(n)e^{-j2\pi fnT_{\mathrm{s}}} \tag{2-2}$$

Then the following relation exists:

$$U_{\mathrm{d}}(e^{j2\pi fT_{\mathrm{s}}}) = \tilde{U}_{\mathrm{d}}(j2\pi f) = F\{\tilde{u}_{d}(t)\} = \int_{-\infty}^{\infty}\tilde{u}_{\mathrm{d}}(t)e^{-j2\pi ft}dt \tag{2-3}$$

The spectrum $U_d(e^{j\omega T_s})$ is linked to $U(j\omega)$ by noticing that a multiplication in the time domain, $u(t)\delta_{T_s}(t)$, corresponds to the convolution of the spectra in the frequency domain, $U(j2\pi f)*(f_s\delta_{f_s}(f))$, with $f_s\delta_{f_s}(f)$ the spectrum of $\delta_{T_s}(t)$, and $\delta_{f_s}(f)$ a periodically repeated Dirac impulse with period $f_s = 1/T_s$

$$\delta_{f_s}(f) = \sum_{k=-\infty}^{+\infty}\delta(f-kf_s) \tag{2-4}$$

Using (2-4), we get

$$U_d(e^{j2\pi fT_s}) = U(j2\pi f)*(f_s\delta_{f_s}(f)) = \frac{1}{T_s}\sum_{k=-\infty}^{+\infty}U(j2\pi(f-kf_s)) \tag{2-5}$$

The convolution of the spectra is illustrated in Figure 2-2. It shows that the sampling process results in a repeated spectrum in the frequency domain with period f_s. If the bandwidth f_B of the sampled signal is larger than half the sampling frequency, the shifted spectra overlap and information is lost. Therefore, it is important to restrict the bandwidth below half the sampling frequency $f_B < f_s/2$ in order to avoid errors. This error is called the aliasing error and the condition on the sample frequency is known as Shannon's sampling theorem. In practice, it is often necessary to put antialias filters to eliminate the high-frequency spectral content of the signal.

2.2.2 The Discrete Fourier Transform (DFT-FFT)

Three basic steps have to be taken to measure the spectrum of a continuous-time signal:

- Discretization in time: sample the continuous-time signal at an equidistant time grid.
- Restrict the length of the data record: our computers can deal with only a finite number of data. Thus, the length of the record is restricted to N samples, excluding the rest. This is called windowing.
- Discretization in frequency: the finite length discrete-time signal still has a continuous frequency spectrum. The value of this spectrum will be calculated only at an equidistant set of frequencies.

The impact of all these steps is illustrated in more detail in the following, in a simple example. The continuous-time signal $u(t) = \cos 2\pi f_0 t$, with $f_0 = 5.5$ Hz is sampled at $f_s = 64$ Hz during 1 second. From these measurements we will calculate the discrete Fourier transform step by step.

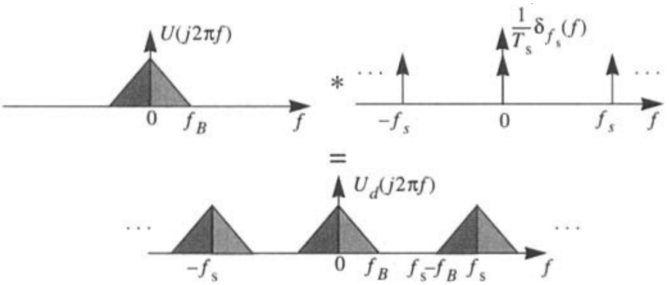


Figure 2-2. Impact of the time domain discretization (sampling) on the spectrum.

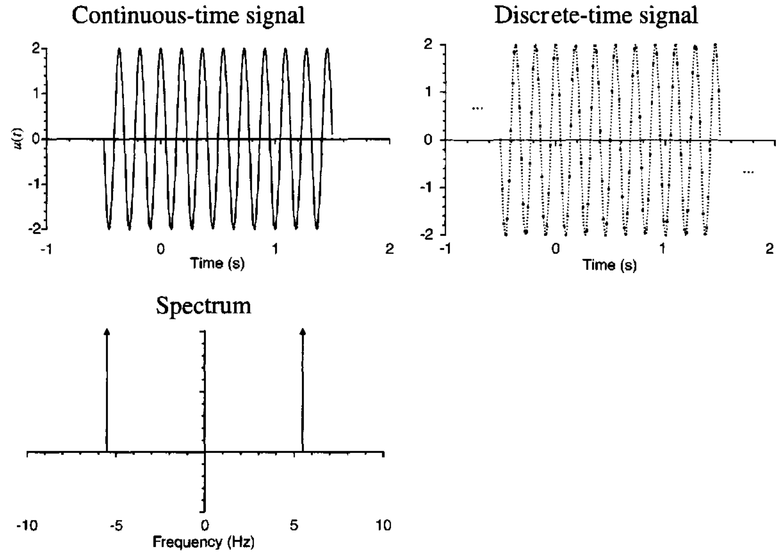


Figure 2-3. The time signal before and after sampling together with the spectrum in the frequency band [−10 Hz, 10 Hz].

2.2.2.1 Discretization in Time. The sampling process has already been discussed in the previous section. Figure 2-3 shows the signal together with its spectrum before and after sampling. In order to keep enough detail in the figures shown, a zoom is made in the frequency band [−10 Hz, 10 Hz]. The periodic repetitions of the spectrum of the discrete-time sequence are not shown. Note that if no aliasing appears, the spectra of the continuous-time and the discrete-time signal are equal to each other within a scale factor.

Mathematical operation:

$$\begin{aligned} &\text{time domain:} && \tilde{u}_{\text{d}}(t) = \textstyle\sum_{n=-\infty}^{+\infty} u(t)\delta(t-nT_{\text{s}}) \\ &\text{frequency domain:} && U_{\text{d}}(e^{j2\pi f T_{\text{s}}}) = T_{\text{s}}^{-1}\textstyle\sum_{k=-\infty}^{+\infty} U(j2\pi(f-kf_{\text{s}})) \end{aligned} \tag{2-6}$$

2.2.2.2 Windowing. The sampled signal still has an infinite length (]-∞, ∞[ ). Because the computer can process only a finite number of samples, we have to restrict the measurement length. We consider only samples that appear in the measurement window:

$$w(t) = 1 \text{ if } 0 \le t < T \quad \text{and} \quad w(t) = 0 \text{ elsewhere} \tag{2-7}$$

This rectangular window, together with its spectrum (the phase is omitted), is shown in Figure 2-4. This window is called a rectangular window and its major characteristic is its width T. Its spectrum $W(j2\pi f)$ is a sinc-like signal, see Eq. (2-8), with zero crossings at the multiples of $1/T$. In this example $T = 1$ s. This window is multiplied with the sampled signal to obtain a new signal that is different from zero in only a finite number of samples.

The spectra have to be convoluted in the frequency domain. Remembering that a convolution with a Dirac impulse is nothing other than a shift of the origin to the position of the impulse, the result of Figure 2-5 is found. The broken lines in the spectra indicate the position of the original frequency components. As can be seen, the restriction of the signal to a finite interval in the time domain smears the power in the frequency domain over the neighboring frequencies. This phenomenon is called leakage.

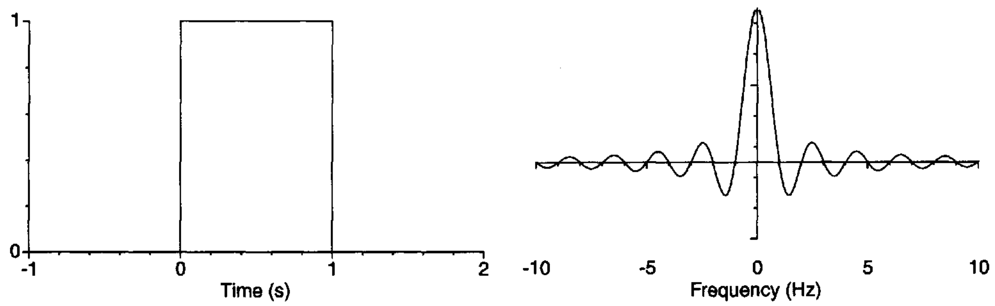


Figure 2-4. Rectangular window and its spectrum (the phase is omitted).

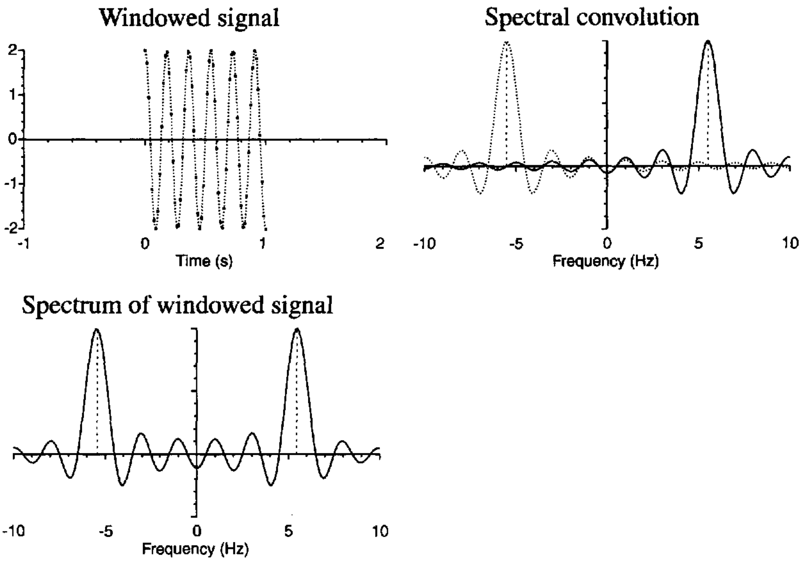


Figure 2-5. Spectrum of the sampled signal after applying a rectangular window.

Mathematical description:

$$\begin{array}{ll} \text{time domain:} & w(t)\tilde{u}_{\mathrm{d}}(t) \\ \text{frequency domain:} & W(j2\pi f) * U_{\mathrm{d}}(e^{j2\pi f T_s}) \end{array} \qquad (2\text{-}8)$$

with $W(j\omega) = Te^{-j\omega T/2}\mathrm{sinc}(\omega T/2)$ and $\mathrm{sinc}(x) = \sin(x)/x$.

2.2.2.3 Discretization in Frequency. As can be seen in Figure 2-5, the spectrum of the sampled and windowed signal is still a continuous frequency signal. Because the spectrum can be calculated in only a finite number of frequencies, the frequencies considered should also be restricted to a discrete grid. An equidistant grid with spacing $1/T$ is selected. Hence, the spectrum is calculated only at the frequencies $f_k = k/T$ Hz. This can be considered as frequency sampling or discretization in frequency. The resulting sampled spectrum shown in Figure 2-6 is quite disappointing. Although the shape of the original spectrum (Figure 2-3) can still be recognized, it seems that all detailed information about it has definitely been lost. The basic reason for this problem is that the original frequency (5.5 Hz) does

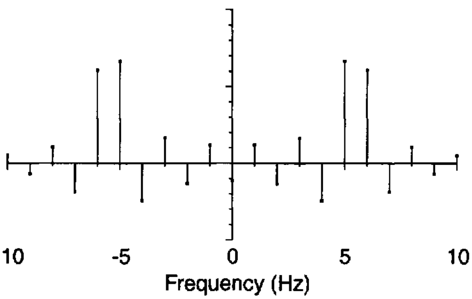


Figure 2-6. DFT result.

not correspond to one of the sampled frequencies in the DFT (multiples of $1/T = 1$ Hz). This can also be seen in the time domain representation of the DFT result. Sampling in the frequency domain at multiples of $1/T$ is described as a multiplication with a Dirac train (see Section 2.2.1) so that in the time domain a convolution should be made with a Dirac train $T\delta_T(t)$. This results in a periodic repetition with period T of the sampled and windowed signal as shown in Figure 2-7. However, T is not a multiple of the signal period, resulting in a discontinuity that appears at the borders of the window as seen in Figure 2-7 ($T = 1$ s in this case).

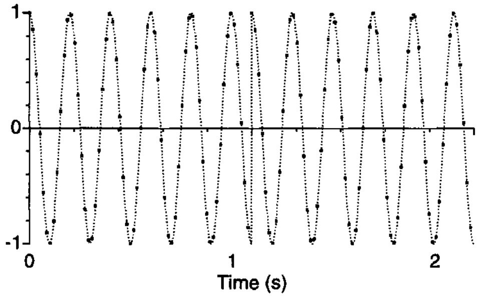


Figure 2-7. Interpretation of the DFT result in the time domain.

Mathematical description:

$$\begin{aligned} &\text{time domain:} && (w(t)\tilde{u}_d(t))*(T\delta_T(t)) \\ &\text{frequency domain:} && (W(j2\pi f)*U_d(e^{j2\pi fT_s}))\delta_{1/T}(f) \end{aligned} \tag{2-9}$$

From (2-9) it follows that the relationship between the time domain samples $u_d(n) = u(nT_s)$ (amplitudes of the Dirac impulses of the time domain signal in (2-9)) and the frequency domain samples $U_{\text{DFT}}(k)$ (amplitudes of the Dirac impulses of the spectrum in (2-9)), is given by

$$U_{\text{DFT}}(k) = \sum\nolimits_{n=0}^{N-1} u(nT_s)e^{-j2\pi nk/N}, \quad k = 0, 1, \ldots, N-1 \tag{2-10}$$

Equation (2-10) is called the discrete Fourier transform (DFT) of the samples $u(nT_s)$, $n = 0, 1, \ldots, N-1$.

If an integer number of periods is measured, the DFT will give an exact copy of the discrete spectrum of the periodic signal. This is illustrated in Figure 2-8 showing the spectra after windowing and after discretization for $u(t) = \cos 2\pi f_0 t$, $f_0 = 5$ Hz, $T = 1$ s. This time no leakage is observed. The basic reason for this remarkable difference is that the continuous-time spectrum equals zero at the frequencies where the spectrum is sampled because the window length is an exact multiple of the period length. Also, the time domain interpretation in Figure 2-9 illustrates the result: this time the periodic repetition coincides with the period of the signal (no discontinuities appear at the multiples of T).

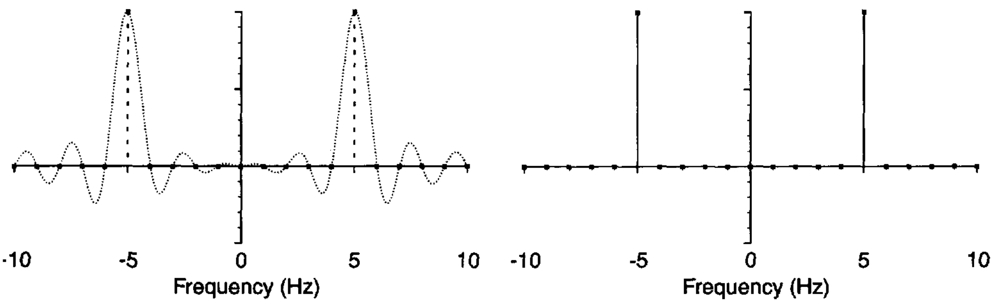


Figure 2-8. DFT spectrum for a periodic signal when an integer number of periods is measured.

At a glance, this seems to be a theoretical result without practical value. The probability of getting an exact match between the signal and the window length is in general, indeed, zero. However, in many FRF measurements, the user masters the generator and the acquisition. In these experimental setups both systems are driven by mother clocks that are synchronized with each other. It is therefore possible for the user to create this ideal match, which eliminates the leakage effect completely. We strongly advise realization of such a setup whenever possible. If for some reason it is impossible to get synchronized measurements, there exist other less attractive alternatives based on windows other than the rectangular window. An extended discussion of the window properties can be found in Harris (1978). In Section 2.2.3 we will briefly touch on this topic.

2.2.2.4 The DFT Expressions. For the samples $u(nT_s)$, $n = 0, 2, \ldots, N-1$, the DFT relations between the time and frequency domain sequences are

$$U_{\mathrm{DFT}}(k) = \sum_{n=0}^{N-1} u(nT_s)e^{-j2\pi nk/N} \quad \text{and} \quad u(nT_s) = \frac{1}{N}\sum_{k=0}^{N-1} U_{\mathrm{DFT}}(k)e^{j2\pi kn/N}$$

(see (2-10)). In this book the scaling factor $1/N$ is symmetrically distributed over both transforms using $1/\sqrt{N}$, and the notation $U_{\mathrm{DFT}}(k)$ will be replaced by $U(k)$ in order not to overload the equations. This gives

$$U(k) = \frac{1}{\sqrt{N}}\sum_{n=0}^{N-1} u(nT_s)e^{-j2\pi nk/N} \quad \text{and} \quad u(nT_s) = \frac{1}{\sqrt{N}}\sum_{k=0}^{N-1} U(k)e^{j2\pi kn/N} \tag{2-11}$$

The straightforward evaluation of Eq. (2-11) requires $O(N^2)$ operations. However, if N is a power of two, a very efficient implementation known as the FFT (fast Fourier transform) is available: it calculates the transforms in $O(N\log_2 N)$ operations (Brigham, 1974). If N is not

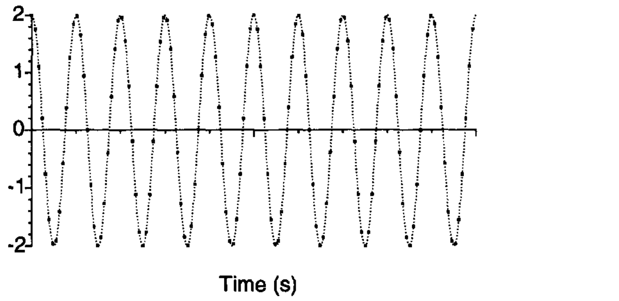


Figure 2-9. Interpretation of the DFT result in the time domain when an integer number of periods is measured.

a power of two, there still exist fast implementations such as the chirp-z transform (Rabiner and Gold, 1975). The FFT algorithm is available in many numerical packages.

2.2.3 DFT Properties of Periodic Signals

2.2.3.1 Integer Number of Periods Measured. Consider the periodic signal $u(t) = \sum_{k=1}^{15}\cos(2\pi k f_0 t + \phi_k)$ in Figure 2-10. Using the same sample frequency, this signal is measured over 1 period and over 10 periods. For both data records the DFT is calculated and the first 150 lines of the DFT spectrum are plotted in Figure 2-11. In both cases an exact recovery of the signal spectrum is made because each time an integer number of periods is measured. However, by measuring 10 times longer ($10N$ data points), the spectral resolution is increased from $1/T = f_s/N$ to $1/(10T) = f_s/(10N)$. Whereas in the former time the spectral lines appear at harmonics $k = 1, 2, ..., 15$, they are placed in the latter time at $k = 10, 20, ..., 150$. The gaps between these spectral lines can be used later on to extract noise information because the noise is nonperiodic and excites all spectral lines.

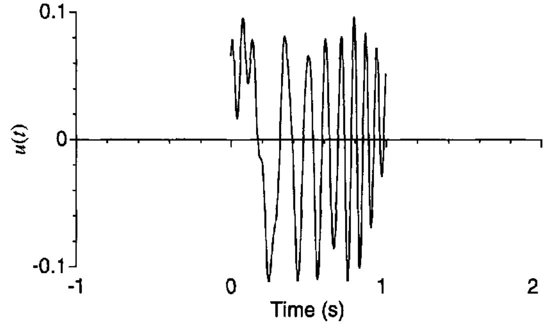


Figure 2-10. Example of a periodic excitation consisting of the sum of 15 sines with equal amplitude and frequencies kf_0, $k = 1, 2, ..., 15$.

2.2.3.2 No Integer Number of Periods Measured. From Section 2.2.2 it is known that leakage errors appear if no integer number of periods is measured. A sound solution for this problem is to change the setup and measure an integer number. If this is impossible we can try to minimize the impact of the leakage on the measurement. A classical technique is to apply a window other than the rectangular one. A concise review of windows and their properties is given by Harris (1978). Here, we present only one of the possibilities: the Hanning or cosine window

$$w(t) = 1 - \cos(2\pi t/T) \text{ if } 0 \le t < T \text{ and } w(t) = 0 \text{ elsewhere} \tag{2-12}$$

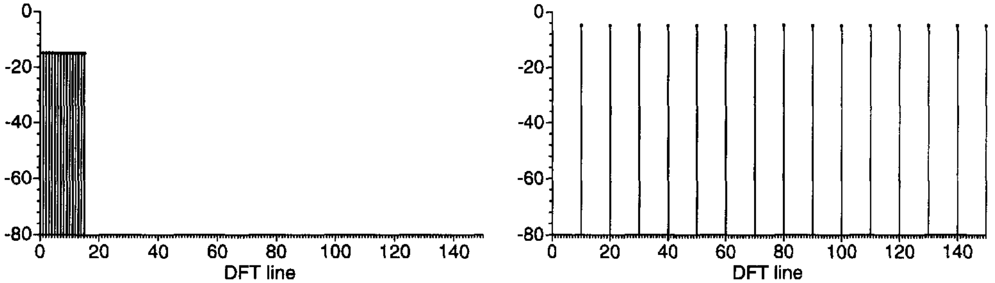


Figure 2-11. DFT spectrum (amplitude in dB) of a periodic signal with 15 components. On the left 1 period in the window, on the right 10 periods in the window.

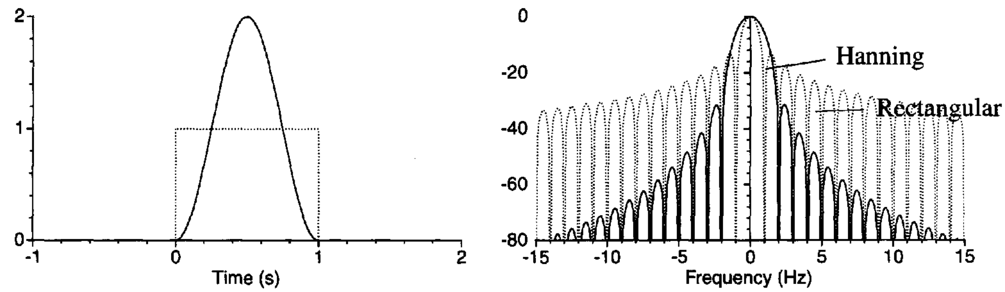


Figure 2-12. Comparison of the rectangular window with the Hanning window in the time domain (left) and the frequency domain (right, amplitude spectrum in dB).

The aim of all the alternative windows is to taper the signal at the beginning and at the end of the window in order to decrease the discontinuities of the periodically reconstructed signal because they are the basic source of the leakage errors. In Figure 2-12 the rectangular (2-7) window and the Hanning (2-12) window are shown. By applying such an alternative window, we do not eliminate the leakage effect but only reshape its impact. Windowing in the time domain is equivalent to a convolution with its spectrum in the frequency domain. The spectrum of the Hanning window decreases much faster than that of a rectangular window, keeping the leakage effect more localized. On the other hand, the main lobe of the Hanning window (first lobe around zero) is two times wider than that of the rectangular window; hence, for components that are close to each other (less than four DFT bins) the interference will increase. This is a typical effect of these windows: they minimize the far leakage effects (far from the position of the original frequency) at a cost of a loss in resolution. The choice of the window also affects the noise sensitivity, the maximum error on the amplitude of the spectral components, etc. We refer the interested reader to Harris (1978) for more information.

To illustrate the effect of the window on the spectrum, we considered 10.5 periods of the periodic signal and calculated the DFT, first with a rectangular window and second with the Hanning window (Figure 2-13). The separation between the components becomes much more visible for the Hanning than for the rectangular window. The interference is reduced from –30 dB (3%) to less than –60 dB (0.1%).

Conclusion. The best solution is to measure an integer number of periods. If this is impossible, the leakage interference between the different spectral components can be reduced by measuring enough periods and using, for example, a Hanning window. For M measured periods, the leakage errors are an $O(M^{-1})$ effect for the rectangular window and an

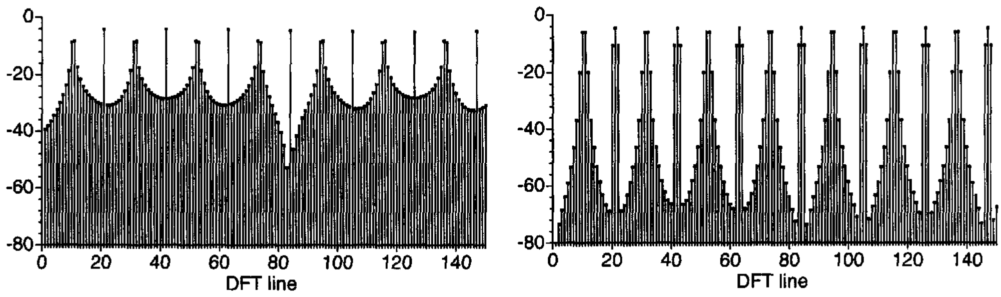


Figure 2-13. Impact of the rectangular (left) and Hanning window (right) on the spectrum for 10.5 measured periods.

$O(M^{-2})$ for the Hanning window. Notice that if at least three or more integer periods are measured, the Hanning window also allows perfect recovery of the original spectral lines. This is a very specific property of the Hanning window that is due to the fact that its zeros coincide with those of the rectangular window except for the main lobe (Figure 2-12). We will make use of this fact in the case study of the CD player (Section 12.2) to eliminate a nonsynchronous periodic disturbance that is about 30 dB above the signal level.

2.2.4 DFT of Burst Signals

The study of the DFT properties showed that no leakage errors occur if periodic signals are analyzed and an integer number of periods is measured. There is an important exception to this general rule: using a DFT, it is possible to sample the continuous spectrum of a burst signal.

Definition 2.1 (Burst Signal): $u(t)$ is a burst signal if $u(t) = 0 \ \forall t \notin [0, T_B]$.

Remark. A time-limited signal cannot be band limited ($|U(j2\pi f)| = 0$ if $|f| > f_B$); thus the time discretization of such a signal always creates aliasing errors. In practice, most burst signals are low-pass filtered signals, which minimize these aliasing effects if a reasonable design is made. In Figure 2-14 an example of such a signal is given. This is an exponentially damped signal that is not exactly zero at the end of the window. So also the "burst" condition is not exactly met, but again the errors are negligible for a good design.

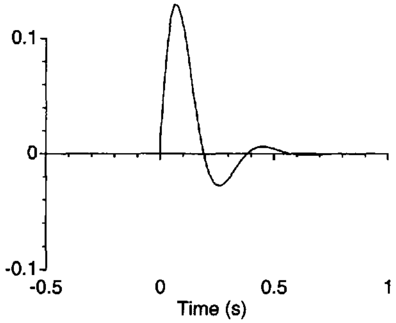


Figure 2-14. Burst signal.

In Section 2.2.2 it was shown that the DFT eventually makes a periodic reconstruction of the original sequence. Because this sequence is zero outside the window ($T > T_B$) this reconstruction does not create discontinuities at the borders and hence the calculated spectrum is a perfect copy of the original one at the DFT lines. This is illustrated in Figure 2-15, where

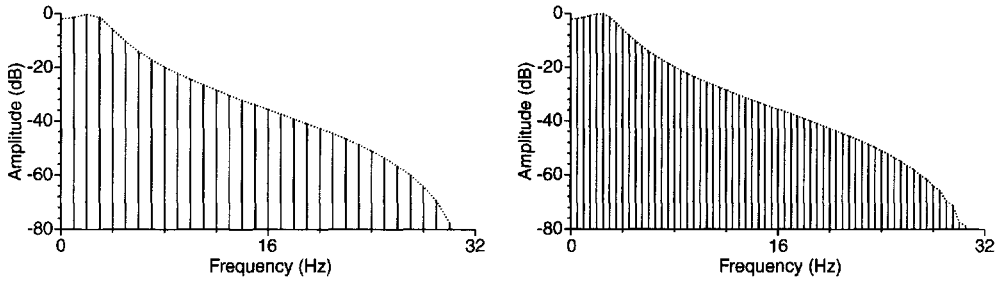


Figure 2-15. DFT spectrum of a burst signal. Left: window length 1 s (64 points); right: window length 2 s (128 points). The dotted line is the original continuous spectrum.

the DFT spectrum of the burst signal in Figure 2-14 is shown. In the first case the window length was 1 s, resulting in a frequency resolution of 1 Hz, and in the second case the window length was 2 s (this can be done by zero appending: N zeros are appended to extend the record length to $2N$), resulting in a frequency resolution of 0.5 Hz.

2.2.5 Conclusion

It is possible to calculate the spectra of sampled signals using the DFT (FFT), but two errors can occur. The first one is the aliasing error: the power at higher frequencies is mirrored at the lower frequencies. To avoid this, the sampling frequency should be set high enough ($f_s > 2f_B$). The second error is leakage: the spectrum of the signal is smeared out due to the finite length of the measurements. In two special, but in practice very important, situations it can be completely avoided. For example, the spectrum of periodic signals measured over an integer number of periods is perfectly calculated by the DFT. It is an exact copy of the spectrum of the continuous-time signals, at least up to half the sample frequency for band-limited signals. We strongly advise the readers to get as close as possible to this ideal situation whenever they have enough freedom during the experiment design. If it is not possible to realize the previous conditions, errors will appear, but it is still possible to reshape these errors to minimize their effect on the results.

2.3 SPECTRAL REPRESENTATIONS OF PERIODIC SIGNALS

In this book we will use three different spectral representations of a periodic signal: the Fourier series, the Fourier transform, and the discrete Fourier transform. Because these all describe the same signal, it is clear that there are close connections between them. Consider a periodic signal, described by its Fourier series representation:

$$u(t) = \sum_{k=1}^{F} |A_k| \cos(k\omega_0 t + \angle A_k) = \sum_{k=-F, k\neq 0}^{F} (A_k/2) e^{jk\omega_0 t} \qquad (2\text{-}13)$$

with $A_k = |A_k| e^{j\angle A_k}$.

- The Fourier coefficient at line k is then

$$U_k = A_k/2 \qquad (2\text{-}14)$$

- The Fourier transform is $F\{u(t)\} = U(j\omega) = \int_{-\infty}^{\infty} u(t) e^{-j\omega t} dt$, with

$$U(j\omega) = \sum_{k=-F, k\neq 0}^{F} (A_k/2)\delta(f - kf_0) \qquad (2\text{-}15)$$

The Dirac impulses account for the convergence problems of the Fourier integral on periodic signals.

- The discrete Fourier transform of one period ($f_s = 1/T_s = Nf_0$) of $u_d(n) = u(nT_s)$ is given by

$$U(k) = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} u_d(n) e^{-j2\pi kn/N} = \sqrt{N} A_k/2 \qquad (2\text{-}16)$$

The difference in notation between the Fourier transform $U(j\omega)$, the discrete Fourier transform $U(k)$, and the Fourier coefficient U_k is indicated by the argument ($j\omega$, k, or subscript k).

2.4 ANALYSIS OF FRF MEASUREMENTS USING PERIODIC EXCITATIONS

In this section we study the principal techniques to measure the FRF of a linear system. During the first part of this analysis we assume that the plant is periodically excited and that an integer number of periods of the steady-state response is measured. The aim of the study is to understand the impact of the disturbing noise on the measured transfer function. Next, we will also consider the use of arbitrary excitations.

2.4.1 Measurement Setup

The typical measurement setup for an FRF measurement is given in Figure 2-16. The generator signal (e.g., a ZOH-reconstructed signal) is applied to the plant (e.g., a mechanical system) using an actuator (e.g., an electromechanical shaker). The input $u_1(t)$ and output $y_1(t)$ (e.g., the applied force and resulting acceleration) are passed through the antialias filter before sampling, resulting in $u_{AA}(t)$ and $y_{AA}(t)$. For simplicity we assume that the antialiasing filters are perfect, leading to the following assumption:

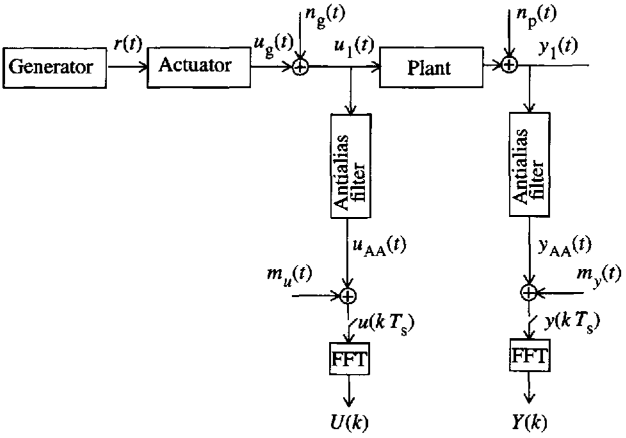


Figure 2-16. Principal measurement setup and notations for periodic signals.

Assumption 2.2 (Band-Limited Measurements): $u_{AA}(t)$, $y_{AA}(t)$ are band-limited copies of $u_1(t)$, $y_1(t)$ obeying the Shannon theorem: e.g., $U_{AA}(j\omega) = U_1(j\omega)$ for $|\omega| < \omega_s/2$, and $U_{AA}(j\omega) = 0$ for $|\omega| \geq \omega_s/2$.

These time domain signals are finally transformed to the frequency domain using the discrete Fourier transform (DFT), implemented as an FFT (fast Fourier transform). In this section we assume that an integer number of periods is measured so that no leakage errors appear. The FRF at frequency f_k is eventually given by

$$G(j\omega_k) = Y(k)/U(k) \tag{2-17}$$

with $f_k = k/T$, and $T = NT_s$ the length of the measured record. This process is disturbed at different points with noise as shown in Figure 2-16. Generator noise $n_g(t)$ distorts the actual, applied excitation; $m_u(t)$ models the measurement noise (e.g., amplifier noise, quantization noise) on the measured input; $m_y(t)$ stands for the output measurement noise; and the process noise (generated by the plant itself) is given by $n_p(t)$. Notice that although the generator noise $n_g(t)$ acts as a proper excitation signal, it is considered in the periodic setup as a noise source because it is a nonperiodic signal. Later in the chapter the consequences of this decision will be analyzed in detail. After the DFT we find, at frequency f_k, that

$$\begin{aligned} Y(k) &= Y_0(k) + N_Y(k) \\ U(k) &= U_0(k) + N_U(k) \end{aligned} \tag{2-18}$$

where $N_U(k)$ and $N_Y(k)$ are the contributions of the noise to the measured Fourier coefficients. The impact of the DFT on the noise is intensively studied. Under very mild conditions on the time domain noise, it is shown that (see Section 14.16) these noise contributions are circular complex normally distributed. For our purpose the most important properties of such a distribution are repeated in the following assumption:

Assumption 2.3 (Disturbing Noise): The input $N_U(k)$ and output $N_Y(k)$ errors satisfy

$$\begin{aligned} &\mathcal{E}\{N_U^l(k)\} = 0, \ \ \mathcal{E}\{N_Y^l(k)\} = 0, \ \ l = 1, 2, \dots \\ &\mathcal{E}\{|N_U(k)|^2\} = \sigma_U^2(k), \ \ \mathcal{E}\{|N_Y(k)|^2\} = \sigma_Y^2(k) \\ &\mathcal{E}\{N_Y(k)\bar{N}_U(k)\} = \sigma_{YU}^2(k) = \bar{\sigma}_{UY}^2(k), \ \ \mathcal{E}\{N_Y(k)N_U(k)\} = 0 \end{aligned} \tag{2-19}$$

for $k = 1, 2, \dots, F$.

At a glance it can be surprising that a squared variable has a zero mean ($\mathcal{E}\{x^2\} = 0$), but the reader should keep in mind that we deal here with complex variables (see also Exercise 14.8). Using these properties, it is easy to carry out a simplified calculation of $\mathcal{E}\{G(j\omega_k)\}$ and $\sigma_G^2(k) = \operatorname{var}(G(j\omega_k))$.

2.4.2 Error Analysis

In this section we calculate the bias (systematic error) and the variability (variance) of the measured FRF. In order to address the essential aspects carefully, the analysis is simplified significantly using a Taylor series, assumed to converge. At the end of the section more precise results are included.

Consider the measured FRF $G(j\omega_k)$:

$$G(j\omega_k) = \frac{Y_0(k) + N_Y(k)}{U_0(k) + N_U(k)} = G_0(j\omega_k)\frac{1 + N_Y(k)/Y_0(k)}{1 + N_U(k)/U_0(k)} \tag{2-20}$$

The Taylor series expansion of $G(j\omega_k)$ is

$$G(j\omega_k) = G_0(j\omega_k)\left(1 + \frac{N_Y(k)}{Y_0(k)}\right)\left(1 - \frac{N_U(k)}{U_0(k)} + \left(\frac{N_U(k)}{U_0(k)}\right)^2\right) + \text{higher order terms} \tag{2-21}$$

In order to calculate the mean value and the variance of $G(j\omega_k)$ it is necessary to make an assumption on the relation between the noise $N_U(k), N_Y(k)$ and the undisturbed signals $U_0(k), Y_0(k)$:

Assumption 2.4 (Disturbing Noise—Continued): The disturbing noise $N_U(k)$, $N_Y(k)$ is independent of the undisturbed signals $U_0(k)$, $Y_0(k)$.

In many cases this is not a difficult assumption. However, in some applications such as measurements in feedback, this assumption is not met if arbitrary excitations are used, leading to systematic errors.

2.4.2.1 Bias Error on the FRF. Under Assumptions 2.3 and 2.4 it follows directly from (2-21) that $\mathcal{E}\{G(j\omega_k)\} = G_0(j\omega_k)$. This result can be extended easily to the higher order terms of the Taylor expansion. It shows that if the Taylor series converges, the expected value equals the exact value. However, it is well known that the Taylor series of $1/(1+x)$ converges only if $|x| < 1$, or in this case $|N_U(k)/U_0(k)| < 1$. For normally distributed noise this condition is always violated by a fraction of the realizations. For high SNR ($\sigma_U(k) < |U_0(k)|$) the previous result will be a very good approximation, but for low SNR a significant bias pops up. If $U_0(k)$ is fixed and the noise is normally distributed, an exact calculation of the expected value can be obtained without using the Taylor series approximation (Guillaume et al., 1992b). For uncorrelated input-output noise ($\sigma_{YU}(k) = 0$) the relative bias $b(k)$ is

$$b(k) = \frac{\mathcal{E}\{G(j\omega_k)\}}{G_0(j\omega_k)} - 1 = -\exp(-|U_0(k)|^2/\sigma_U^2(k)) \tag{2-22}$$

This shows that, even for a moderate SNR, small bias errors exist, for example, for an SNR of 6 dB ($|U_0(k)|/\sigma_U(k) = 2$), $|b(k)| = 0.018$, but the reader should be aware that significant outliers on $G(j\omega_k)$ can appear. For an SNR of 10 dB, $|b(k)| = 5\times10^{-5}$.

If the input noise and output noise are linearly correlated, as in the case of feedback, a more complicated expression is found (see Appendix 7.G):

$$b(k) = -\exp(-|U_0(k)|^2/\sigma_U^2(k))\left(1 - \rho(k)\frac{U_0(k)/\sigma_U(k)}{Y_0(k)/\sigma_Y(k)}\right) \tag{2-23}$$

with $\rho(k) = \sigma_{YU}^2(k)/(\sigma_U(k)\sigma_Y(k))$. Also in this case the maximal relative bias (2-23) is quite small. It is smaller than 1×10^{-4} if the worst case input and output signal-to-noise ratios $|U_0(k)|\sigma_U(k)$, $|Y_0(k)|/\sigma_Y(k)$ are larger than 10 dB.

This good behavior is due to the use of periodic excitations. If $U_0(k)$ is also a stochastic variable, as is the case for random excitations, the analysis is much more involved. It turns out that in this case the FRF methods are much more sensitive to the noise, leading rapidly to large systematic errors. This discussion is postponed to Section 2.6 but, just as an illustration, it can be mentioned that the bias errors in this case grow to more than 20% for an SNR of 6 dB.

2.4.2.2 Variance Analysis of the FRF. Under Assumption 2.4 and restricting the Taylor expansion in (2-21) to the first-order terms,

$$\begin{aligned} G(j\omega_k) &\approx G_0(j\omega_k)\left(1+\frac{N_Y(k)}{Y_0(k)}-\frac{N_U(k)}{U_0(k)}\right) = G_0(j\omega_k)+N_G(k) \\ N_G(k) &= G_0(j\omega_k)\left(\frac{N_Y(k)}{Y_0(k)}-\frac{N_U(k)}{U_0(k)}\right) \end{aligned} \tag{2-24}$$

and because $\mathcal{E}\{N_G(k)\} = 0$ the variance is given by

$$\sigma_{\hat{G}}^2(k) = \mathcal{E}\{|N_G(k)|^2\} = |G_0(j\omega_k)|^2\left(\frac{\sigma_{\hat{Y}}^2(k)}{|Y_0(k)|^2}+\frac{\sigma_{\hat{U}}^2(k)}{|U_0(k)|^2}-2\mathrm{Re}(\frac{\sigma_{\hat{Y}U}^2(k)}{Y_0(k)\bar{U}_0(k)})\right) \tag{2-25}$$

The variance is inversely proportional to the square of the SNR of the measurements. This result facilitates the excitation design and answers the question of how the power spectrum of the excitation signal should be chosen to cause a small uncertainty.

Remark. In the previous calculations, an approximate expression for the variance is obtained. A detailed analysis (Broersen, 1995) shows, however, that the variance of $G(j\omega_k)$ does not exist because of the presence of outliers that appear when the denominator comes very close to zero. This risk disappears for improving SNR. The variance (2-25) can then be interpreted as that of the limiting distribution. Guillaume et al. (1992b) showed that the problem can be removed by eliminating the measurements with a "too small" denominator so that no outliers appear anymore.

For the special case that the generator noise dominates ($m_u(t) = 0$, $m_y(t) = 0$ and $n_p(t) = 0$), the following relations exist: $n_y(t) = g_0(t)*n_u(t)$, with $g_0(t)$ the plant impulse response, so that

$$\sigma_{\hat{Y}}^2(k) \approx |G_0(j\omega_k)|^2\sigma_{\hat{U}}^2(k) \quad \text{and} \quad \sigma_{\hat{Y}U}^2(k) = \mathcal{E}\{N_Y(k)\bar{N}_U(k)\} \approx G_0(j\omega_k)\sigma_{\hat{U}}^2(k)$$

The approximations are due to the leakage effect that appears when random signals are Fourier transformed with the DFT. Substituting these results into (2-25) gives $\sigma_{\hat{G}}^2(k) \approx 0$, which implies that the generator noise does not contribute to the uncertainty on the FRF measurements. It also does not contribute to better knowledge of the system because the $n_g(t)$ contributions disappear in the periodic averaging process. This means that some information is lost because $n_g(t)$ can also be considered as an excitation signal.

A number of possibilities are available to reduce the variance $\sigma_{\hat{G}}^2(k)$. The most simple solution is to inject more power into the system, increasing $|U_0(k)|$ and $|Y_0(k)|$. In Chapter 4 methods are proposed to maximize this power, while the peak value of the excitations remains below a user-specified level, so that nonlinear operation of the plant is avoided. A second possibility is to measure the FRF frequency by frequency, making stepped sine measurements that concentrate all power at one frequency at a time, so that the SNR is maximized. The disadvantage of this method is that it can become extremely slow because at each frequency point sufficient waiting time should be added until all transients due to the frequency change have disappeared. The alternative is to use well-designed broadband excitations in combination with good averaging methods. This solution depends, again, strongly on the periodic or random nature of the excitation signal, leading to completely different methods.

2.5 REDUCING FRF MEASUREMENT ERRORS FOR PERIODIC EXCITATIONS

In this section it is shown how to reduce the bias and the variance of FRF measurements using well-designed averaging techniques. Because the solutions strongly depend on the periodic or random behavior of the excitation, the discussion is split into two parts. In the first part we deal with periodic signals because they lead to the best solutions, while the algorithms are very simple. In the next section random excitations are considered because they are still very popular, even if they lead to inferior results compared with periodic excitations.

All the FRF averaging techniques start from M input-output data blocks $u^{[l]}(t)$, $y^{[l]}(t)$, $l = 1, 2, ..., M$. To study the stochastic behavior of theses averaging methods we need an assumption concerning the way the data blocks $u^{[l]}(t)$, $y^{[l]}(t)$, $l = 1, 2, ..., M$ are collected.

Assumption 2.5 (Measurement Data Blocks): The M input-output data blocks $u^{[l]}(t)$, $y^{[l]}(t)$, $l = 1, 2, ..., M$ stem either (i) from M independent (possibly repeated) experiments where the disturbing noise $n_u^{[l]}(t)$, $n_y^{[l]}(t)$ has finite Pth order moments and is independent over l or (ii) from a single experiment where the disturbing noise $n_u(t)$, $n_y(t)$ can be written as filtered white noise with finite Pth order moments.

Intuitively, Assumption 2.5(ii) boils down to saying that the correlation length of the noise should be much smaller than the total measurement time.

2.5.1 Basic Principles

In this section we assume again explicitly that the excitation signal $u_0(t)$ is periodic with period T, such that the sampled signal $u_0(nT_s) = u_0((n+N_p)T_s)$. Notice that this also imposes a constraint on the sampling period because the signal period should be a multiple of the sampling period $T = N_pT_s$. For notational simplicity, we drop the sampling period T_s in the argument of the signals; for example, $x(nT_s)$ is denoted as $x(n)$. When periodic excitations are applied, it is possible to collect M successive periods (with length N_p) and to average the measurements in the time domain over these repeated periods, exemplified by the output measurement (Figure 2-17):

$$\hat{y}(n) = \frac{1}{M}\sum_{l=0}^{M-1} y(n+lN_p) = \frac{1}{M}\sum_{l=1}^{M} y^{[l]}(n) \text{ with } y^{[l]}(n) = y(n+(l-1)N_p) \tag{2-26}$$

and the DFT is $\hat{Y}(k) = \text{DFT}(\hat{y}(n))$. The FRF estimate is

$$\hat{G}_{\text{ML}}(j\omega_k) = \frac{\hat{Y}(k)}{\hat{U}(k)} \tag{2-27}$$

$\hat{G}_{\text{ML}}$ is the maximum likelihood solution for Gaussian disturbances if the repeated measurements $u^{[l]}$, $y^{[l]}$ can be considered to be independent over l. It is clear that due to the averaging process, the noise is reduced as $1/\sqrt{M}$ under Assumptions 2.4 and 2.5 $(P = 2)$, so that, asymptotically,

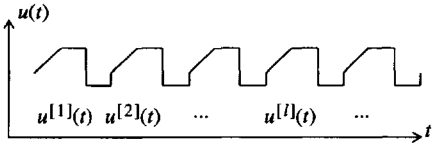


Figure 2-17. Processing periodic excitations.

$$\begin{aligned}\underset{M\to\infty}{\text{a.s.lim}}\hat{G}_{\text{ML}}(j\omega_k) &= \frac{\underset{M\to\infty}{\text{a.s.lim}}\hat{Y}(k)}{\underset{M\to\infty}{\text{a.s.lim}}\hat{U}(k)} = \frac{Y_0(k)}{U_0(k)} = G_0(j\omega_k)\\ \hat{G}_{\text{ML}}(j\omega_k) &= G_0(j\omega_k) + O_{\text{P}}(M^{-1/2})\end{aligned} \tag{2-28}$$

in the absence of other systematic error sources typified by instrumentation errors (proof: see Appendix 2.A). Moreover, under Assumption 2.4 and Assumption 2.5 (i, $P = 2 + \varepsilon$) or (ii, $P = \infty$), the FRF estimate $\hat{G}_{\text{ML}}(j\omega_k)$ (2-27) is asymptotically normally distributed (see Appendix 2.A). Many dynamic signal analyzers offer this averaging option; for example, $M = 128$ averages are made over $N_{\text{p}} = 2048$ points. Because this improves the results at a very low computational cost, it is strongly advised to make full use of this option. In practice, M is determined by the maximum measurement time T and the minimum required frequency resolution f_0: $M = Tf_0$.

Although the computational effort is minimized by first averaging the measurements in the time domain before calculating the DFTs, it also makes sense to calculate the spectrum of each individual subrecord and perform the averaging in the frequency domain. In the latter case it is also possible to estimate the noise (co-)variance. Because the DFT is a linear operator, the order of the operations does not influence the result. Consider the DFTs of the subrecords

$$U^{[l]}(k) = \text{DFT}(u^{[l]}(n)), \qquad Y^{[l]}(k) = \text{DFT}(y^{[l]}(n)) \tag{2-29}$$

and calculate the sample mean

$$\hat{U}(k) = \frac{1}{M}\sum\nolimits_{l=1}^{M} U^{[l]}(k),\ \hat{Y}(k) = \frac{1}{M}\sum\nolimits_{l=1}^{M} Y^{[l]}(k),\ \text{with } \hat{G}_{\text{ML}}(j\omega_k) = \frac{\hat{Y}(k)}{\hat{U}(k)} \tag{2-30}$$

and the sample (co-)variances

$$\begin{aligned}\hat{\sigma}^2_{\hat{U}}(k) &= \frac{1}{M-1}\sum\nolimits_{l=1}^{M}\left|U^{[l]}(k) - \hat{U}(k)\right|^2,\ \hat{\sigma}^2_{\hat{Y}}(k) = \frac{1}{M-1}\sum\nolimits_{l=1}^{M}\left|Y^{[l]}(k) - \hat{Y}(k)\right|^2\\ \hat{\sigma}^2_{\hat{Y}U}(k) &= \frac{1}{M-1}\sum\nolimits_{l=1}^{M}(Y^{[l]}(k) - \hat{Y}(k))\overline{(U^{[l]}(k) - \hat{U}(k))}\end{aligned} \tag{2-31}$$

These are unbiased estimates of the true (co-)variances. Under Assumptions 2.4 and 2.5 (i, $P = 2$), the asymptotic variance of $\hat{G}_{\text{ML}}(j\omega_k)$ (2-27) is given by (2-25) (see Appendix 2.A). Using (2-31), it can be approximated as

$$\sigma^2_{\hat{G}}(k) \approx \frac{\left|\hat{G}_{\text{ML}}(j\omega_k)\right|^2}{M}\left(\hat{\sigma}^2_{\hat{Y}}(k)/\left|\hat{Y}(k)\right|^2 + \hat{\sigma}^2_{\hat{U}}(k)/\left|\hat{U}(k)\right|^2 - 2\text{Re}(\hat{\sigma}^2_{\hat{Y}U}(k)/(\hat{Y}(k)\overline{\hat{U}(k)}))\right) \tag{2-32}$$

The additional division by M is due to the averaging effect that reduces the noise variance by a factor M if the noise can be considered to be uncorrelated from one subrecord to the other.

2.5.2 Processing Repeated Measurements

Many instruments do not have enough memory to store long data records. Instead they make repeated synchronized (start each time at the same point in the period) measurements of the periodic signal by using a good trigger. In practice, a slight variation appears from measurement to measurement, resulting in time jitter. Consider, for simplicity, noiseless measurements. Then

$$u^{[l]}(nT_s) = u_0(nT_s - \tau^{[l]}) \tag{2-33}$$

with $\tau^{[l]}$ the variation with respect to the perfect starting point of the measurement. The expected value becomes

$$\mu_u(nT_s) = \mathcal{E}\{u^{[l]}(nT_s)\} = \int_{-\infty}^{\infty} u_0(nT_s - \tau) f_\tau(\tau) d\tau \tag{2-34}$$

with $f_\tau(\tau)$ the probability density function of the jitter, and its spectrum is

$$M_u(e^{j\omega T_s}) = U_0(e^{j\omega T_s}) F_\tau(j\omega) \tag{2-35}$$

with $F_\tau(j\omega) = F\{f_\tau(\tau)\}$ the characteristic function of $f_\tau(\tau)$. This shows that the jitter acts as a linear filter on the data (Souders et al., 1990). It creates no systematic errors if the jitter is the same for the input and the output error. However, the uncertainty on the FRF measurement increases, especially at the higher frequencies because $F_\tau(j\omega)$ has a low-pass behavior. For example, for normally distributed jitter $N(0, \alpha^2 T_s^2)$,

$$F_\tau(j\omega) = e^{-(\omega^2 \alpha^2 T_s^2)/2} = e^{-\alpha(\omega/\omega_s)^2 2\pi^2} \tag{2-36}$$

For jitter with a standard deviation of one sample, a loss of 11 dB appears at $f_s/4$ and 43 dB at $f_s/2$. This clearly shows that it is extremely important to pay sufficient attention to the quality of the triggering if full band measurements are made.

2.5.3 Improved Averaging Methods for Nonsynchronized Measurements

Sometimes it is impossible to get a proper trigger signal that guarantees good synchronization of the measurements. A prime possibility to solve this problem is to perform a postsynchronization, estimating each time the delay with respect to the reference record (for example, the first one) and adding a corresponding phase shift $e^{j\omega\tau^{[l]}}$ to the measurements. An alternative is to calculate the FRF of each individual measurement (the division $Y(k)/U(k)$ eliminates the varying delay). As explained in Section 2.4.2.1, this can create bias errors if the simple arithmetic mean is used to average the individual FRF measurements. In Guillaume et al. (1992b) nonlinear averaging methods have been developed that are more robust on this aspect, without increasing the variance significantly. The most robust method turned out to be

$$\begin{aligned} \left|\hat{G}_{\mathrm{H}_{\log}}(j\omega_k)\right| &= \exp\left(\frac{1}{M}\sum\nolimits_{l=1}^{M}\mathrm{Re}\left(\log\frac{Y^{[l]}(k)}{U^{[l]}(k)}\right)\right) \\ \angle\hat{G}_{\mathrm{H}_{\log}}(j\omega_k) &= \angle\hat{S}_{YU}(j\omega_k) \quad \text{with} \quad \hat{S}_{YU}(j\omega_k) = \frac{1}{M}\sum\nolimits_{l=1}^{M}Y^{[l]}(k)\overline{U}^{[l]}(k) \end{aligned} \tag{2-37}$$

The split between amplitude and phase is made to avoid the phase wrapping problems of the complex logarithm. For circular complex normally distributed errors it is shown under Assumptions 2.4 and 2.5(i) that the relative amplitude error $\left|\hat{G}_{\mathrm{H}_{\log}}(j\omega_k)\right| / \left|G_0(j\omega_k)\right| - 1$ converges for $M \to \infty$ to

$$\exp\left(\frac{1}{2}\mathrm{Ei}\left(-\frac{|U_0(k)|^2}{\sigma_U^2(k)}\right) - \frac{1}{2}\mathrm{Ei}\left(-\frac{|Y_0(k)|^2}{\sigma_Y^2(k)}\right)\right) - 1 \tag{2-38}$$

with Ei(.) the exponential integral functions (Gradshteyn and Ryzhik, 1980). This result is also valid for correlated input-output noise (still assuming that $U_0(k)$ is independent of the disturbing noise). This results in very small bias errors, even for poor SNR, as given in Figure 2-18. A comparison with other classical methods that were originally developed for random excitations is given in Section 2.6.

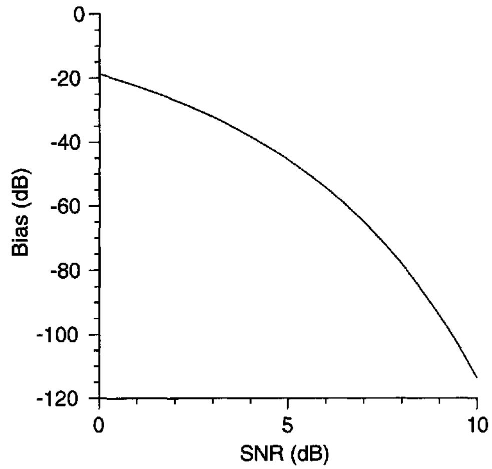


Figure 2-18. Maximum relative bias of $\left|\hat{G}_{\mathrm{H}_{\log}}(j\omega_k)\right|$ for a given worst case SNR (on input or output).

Remarks

(i) The relative amplitude error (2-38) is also valid in the presence of correlated noise because the log operator in (2-37) separates both noise sources.

(ii) The phase estimate is unbiased if the noise is uncorrelated $\sigma_{YU}(k) = 0$.

2.5.4 Coherence

A measure often used to quantify the quality of the obtained FRF is the coherence $\gamma^2(\omega)$ defined as

$$\gamma^2(\omega) = \frac{|S_{YU}(j\omega)|^2}{S_{UU}(j\omega)S_{YY}(j\omega)} \tag{2-39}$$

It measures how much of the output power is coherent (linearly related) with the input power (Bendat and Piersol, 1980; Cadzow and Solomon, 1987). It is shown to be captured between 0 and 1:

$$0 \le \gamma^2(\omega) \le 1 \tag{2-40}$$

If $\gamma(\omega)$ is smaller than 1 it indicates the presence of

- Extraneous noise in the measurements
- Leakage errors of the DFT
- A nonlinear distortion (only for random excitations)
- Other inputs besides $u(t)$ contributing to the output

For periodic signals Eq. (2-39) becomes

$$\hat{\gamma}^2(\omega_k) = \frac{\left|\frac{1}{M}\sum_{l=1}^{M} Y^{[l]}(k)\bar{U}^{[l]}(k)\right|^2}{\left(\frac{1}{M}\sum_{l=1}^{M}|U^{[l]}(k)|^2\right)\left(\frac{1}{M}\sum_{l=1}^{M}|Y^{[l]}(k)|^2\right)} = \frac{\left|1+\frac{\hat{\sigma}^2_{YU}(k)}{Y_0(k)\bar{U}_0(k)}\right|^2}{\left(1+\frac{\hat{\sigma}^2_U(k)}{|U_0(k)|^2}\right)\left(1+\frac{\hat{\sigma}^2_Y(k)}{|Y_0(k)|^2}\right)} \tag{2-41}$$

where the exact (co-)variances are replaced by sample (co-)variances. Notice that $\gamma^2(\omega_k) = 1$ when there is only generator noise and the leakage errors are neglected. Sometimes coherence is used to detect nonlinear distortions although its value is unity for periodic excitations in the absence of noise ($\sigma^2_U(k) = 0$, $\sigma^2_Y(k) = 0$ and $\sigma^2_{YU}(k) = 0$), independent of the presence of nonlinearities (McCormack et al., 1994b). Hence, better alternatives, given in Chapter 3, are sought for the detection of nonlinear distortions.

The variance on the measured FRF can be estimated directly from the coherence by

$$\hat{\sigma}^2_G(k) \approx |G(j\omega_k)|^2\frac{1-\gamma^2(\omega_k)}{\gamma^2(\omega_k)} \tag{2-42}$$

This follows directly from substitution of Eq. (2-41) into (2-42), assuming that

$$\frac{\hat{\sigma}^2_U(k)\hat{\sigma}^2_Y(k)}{|U_0(k)|^2|Y_0(k)|^2} \ll \frac{\hat{\sigma}^2_U(k)}{|U_0(k)|^2}+\frac{\hat{\sigma}^2_Y(k)}{|Y_0(k)|^2} \quad \text{and} \quad \left|\frac{\hat{\sigma}^2_{YU}(k)}{Y_0(k)\bar{U}_0(k)}\right| \ll 1$$

This estimate will be very useful in the case of random excitations, where it is impossible to estimate $\hat{\sigma}^2_U(k)$, $\hat{\sigma}^2_Y(k)$, and $\hat{\sigma}^2_{YU}(k)$ directly from the data.

2.6 FRF MEASUREMENTS USING RANDOM EXCITATIONS

In this section we focus on methods that are also applicable to random excitations. The major difference compared with periodic excitations is the variation of the excitation from one realization (subrecord) to the other. This requires other methods to get acceptable results. A comprehensive overview of dedicated FRF measurement techniques for random signals is given in the book of Bendat and Piersol (1980). In this section we give a brief introduction and an alternative to improve the classical methods.

2.6.1 Basic Principles

Consider a linear system driven with random excitations, so that $u_0(t)$ is no longer periodic. Under these conditions the analysis of the previous section is no longer valid. For example, it is no longer possible to consider a fixed value $U_0(k)$ in the Taylor expansion as was done in Section 2.5. A more detailed analysis is needed because the excitation signal varies from one realization to the other. These aspects will be tackled first and dedicated solutions to deal with random excitations are proposed in Section 2.6.2. Also, leakage errors appear (see Section 2.2.2). In general, the spectrum of random signals does not even exist (Bendat and Piersol, 1980; Papoulis, 1981) so that again a detailed analysis is required to understand exactly what is going on.

2.6.2 Reducing the Noise Influence

When measuring the FRF using random excitations, the same approach could be made as for periodic data. The full record is again split into M subrecords with input and output DFT spectra $U^{[l]}(k)$, $Y^{[l]}(k)$ for block l. Eventually, the FRF for block l is then $Y^{[l]}(k)/U^{[l]}(k)$. Broersen (1995) showed that this direct calculation has an infinite variance. From Eq. (2-20) it is also seen that bias errors are created because $\mathcal{E}\{1/(1+N_U(k)/U_0(k))\} \neq 1$. This bias is mainly induced by the nonlinear behavior of the division. The bias will be small only if $|N_U(k)/U_0(k)| \ll 1$. It is, therefore, necessary to reduce the noise by averaging before making the division. However, because $\mathcal{E}\{U^{[l]}(k)\} = 0$, it is clear that this cannot be done straightforwardly. The reason for this problem is that the vector $U^{[l]}(k)$ has a random phase, uniformly distributed between $[0, 2\pi[$ so that its averaged value is zero (see Figure 2-19). A possibility to avoid this problem is to eliminate the phase of $U^{[l]}(k)$ by multiplying it with its complex conjugate, to get vectors with a fixed phase as shown in Figure 2-20. It is also possible to average before making the division:

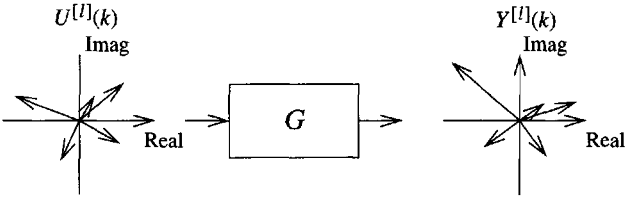


Figure 2-19. Successive realizations of $U^{[l]}(k)$ and $Y^{[l]}(k)$.

$$\hat{G}(j\omega_k) = \frac{\sum_{l=1}^{M} Y^{[l]}(k)\bar{U}^{[l]}(k)}{\sum_{l=1}^{M} |U^{[l]}(k)|^2} \tag{2-43}$$

Readers who are familiar with this field will observe that this expression is nothing other than the discrete implementation of the Wiener-Hopf equation (see Bendat and Piersol, 1980, Eq. (4.7), and Eykhoff, 1974, Eq. (8.10)), relating the cross-power with the autopower spectrum: $S_{YU}(j\omega) = G(j\omega)S_{UU}(j\omega)$. The asymptotic properties can be obtained easily by splitting the measurements into the undisturbed parts $U_0(k)$, $Y_0(k)$ (neglecting the leakage effects) and the distortions $N_U(k)$, $N_Y(k)$. Under Assumptions 2.4 and 2.5 ($P = 4$) the systematic errors and the variability can be calculated.

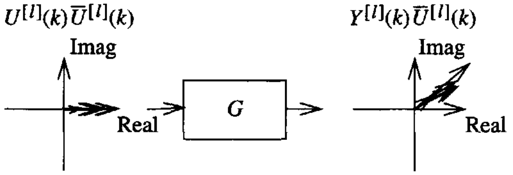


Figure 2-20. Successive realizations of $U^{[l]}(k)\bar{U}^{[l]}(k)$ and $Y^{[l]}(k)\bar{U}^{[l]}(k)$.

2.6.2.1 Systematic Errors. Under Assumptions 2.4 and 2.5 $(P = 4)$, the estimate (2-43) converges to

$$\underset{M\to\infty}{\text{a.s.lim}}\,\hat{G}(j\omega_k) = \frac{\underset{M\to\infty}{\text{a.s.lim}}\frac{1}{M}\sum_{l=1}^{M} Y^{[l]}(k)\bar{U}^{[l]}(k)}{\underset{M\to\infty}{\text{a.s.lim}}\frac{1}{M}\sum_{l=1}^{M} |U^{[l]}(k)|^2} = \frac{\mathcal{E}\{\,Y_0(k)\bar{U}_0(k)\,\} + \sigma^2_{YU}(k)}{\mathcal{E}\{\,|U_0(k)|^2\,\} + \sigma^2_U(k)} \tag{2-44}$$

at the rate $O_p(M^{-1/2})$ (see Appendix 2.A). Moreover, under Assumption 2.4 and Assumption 2.5(i, $P = 4+\varepsilon$) or (ii, $P = \infty$), the FRF estimate $\hat{G}(j\omega_k)$ (2-43) is asymptotically normally distributed (see Appendix 2.A). Neglecting the leakage effects, (2-44) becomes

$$\underset{M\to\infty}{\text{a.s.lim}}\,\hat{G}(j\omega_k) \approx G_0(j\omega_k)\frac{1 + \sigma^2_{YU}(k)/\mathcal{E}\{\,Y_0(k)\bar{U}_0(k)\,\}}{1 + \sigma^2_U(k)/\mathcal{E}\{\,|U_0(k)|^2\,\}} \tag{2-45}$$

Notice that for random signals $\mathcal{E}\{\,|U_0(k)|^2\,\}$ cannot be replaced by $|U_0(k)|^2$ because $U_0(k)$ varies from one realization to the other. Equation (2-45) shows that there is a systematic error that did not appear in the previous approach. This is the price to be paid for using random instead of periodic excitations. If the input signal can be measured free of noise, $\sigma_U(k) = 0$, the bias disappears. The method (2-43) is sometimes called the H_1 method. If the SNR at the output is much higher than that at the input, then it is better to use the following alternative:

$$\hat{G}(j\omega_k) = \frac{\sum_{l=1}^{M} |Y^{[l]}(k)|^2}{\sum_{l=1}^{M} U^{[l]}(k)\bar{Y}^{[l]}(k)} \tag{2-46}$$

which is called the H_2 method. The H_2 method (2-46) under the same noise assumptions has the same asymptotic $(M \to \infty)$ properties as the H_1 method (2-43). Neglecting the leakage effects, the asymptotic value of (2-46) is

$$\underset{M\to\infty}{\text{a.s.lim}}\,\hat{G}(j\omega_k) \approx G_0(j\omega_k)\frac{1 + \sigma^2_Y(k)/\mathcal{E}\{\,|Y_0(k)|^2\,\}}{1 + \sigma^2_{UY}(k)/\mathcal{E}\{\,U_0(k)\bar{Y}_0(k)\,\}} \tag{2-47}$$

For uncorrelated noise, $\sigma^2_{UY}(k) = 0$, (2-45) and (2-47) reduce to, respectively,

$$|G_0(j\omega_k)|/|1 + \sigma^2_U(k)/\mathcal{E}\{\,|U_0(k)|^2\,\}| \quad \text{and} \quad |G_0(j\omega_k)||1 + \sigma^2_Y(k)/\mathcal{E}\{\,|Y_0(k)|^2\,\}|$$

Hence,

$$\left|\underset{M\to\infty}{\text{a.s.lim}}\hat{G}_{\mathrm{H}_1}(j\omega_k)\right| \le |G_0(j\omega_k)| \le \left|\underset{M\to\infty}{\text{a.s.lim}}\hat{G}_{\mathrm{H}_2}(j\omega_k)\right| \tag{2-48}$$

where $\hat{G}_{\mathrm{H}_1}(j\omega_k)$ and $\hat{G}_{\mathrm{H}_2}(j\omega_k)$ are given by, respectively, (2-43) and (2-46). This result cannot be generalized to the case of correlated noise.

2.6.2.2 Variance. An approximate expression for the variance of $\hat{G}(j\omega_k)$ (valid for the H_1 and H_2) is found by considering only the linear noise contributions to (2-43):

$$\hat{G}(j\omega_k) = G_0(j\omega_k)\frac{1+N_1(k)}{1+N_2(k)} \approx G_0(j\omega_k)(N_1(k)-N_2(k)) \tag{2-49}$$

with

$$N_1(k) = \frac{\sum_{l=1}^{M} N_Y^{[l]}(k)\bar{U}_0^{[l]}(k) + Y_0^{[l]}(k)\bar{N}_U^{[l]}(k)}{\sum_{l=1}^{M} Y_0^{[l]}(k)\bar{U}_0^{[l]}(k)}, \quad N_2(k) = \frac{\sum_{l=1}^{M} N_U^{[l]}(k)\bar{U}_0^{[l]}(k) + U_0^{[l]}(k)\bar{N}_U^{[l]}(k)}{\sum_{l=1}^{M} |\bar{U}_0^{[l]}(k)|^2}$$

Next, the variance of (2-49) is obtained assuming that the M data blocks (subrecords) are independent, Assumption 2.5(i), and by taking the expected value $\mathscr{E}\{\,|G_0(j\omega_k)(N_1(k)-N_2(k))|^2\}$ with respect to the noise and not to the random excitation signal. This means that we calculate the variance that would be obtained if the experiment was repeated with the same noise realizations for the excitation signal. Neglecting the leakage errors,

$$G_0(j\omega_k) \approx \frac{Y_0^{[l]}(k)}{U_0^{[l]}(k)} \approx \frac{\sum_{l=1}^{M} Y_0^{[l]}(k)\bar{U}_0^{[l]}(k)}{\sum_{l=1}^{M} |U_0^{[l]}(k)|^2} \approx \frac{\sum_{l=1}^{M} |Y_0^{[l]}(k)|^2}{\sum_{l=1}^{M} \bar{Y}_0^{[l]}(k)U_0^{[l]}(k)}$$

we find

$$\sigma_{\hat{G}}^2(k) = |G_0(j\omega_k)|^2\left[\frac{\sigma_Y^2(k)}{\sum_{l=1}^{M}|Y_0^{[l]}(k)|^2} + \frac{\sigma_U^2(k)}{\sum_{l=1}^{M}|U_0^{[l]}(k)|^2} - 2\mathrm{Re}\left(\frac{\sigma_{YU}^2(k)}{\sum_{l=1}^{M} Y_0^{[l]}(k)\bar{U}_0^{[l]}(k)}\right)\right] \tag{2-50}$$

If the number of blocks $M \to \infty$ and we assume that the random excitation is stationary, the variance becomes

$$\underset{M\to\infty}{\text{a.s.lim}} M\sigma_{\hat{G}}^2(k) = |G_0(j\omega_k)|^2\left(\frac{\sigma_Y^2(k)}{S_{Y_0Y_0}(j\omega_k)} + \frac{\sigma_U^2(k)}{S_{U_0U_0}(j\omega_k)} - 2\mathrm{Re}\left(\frac{\sigma_{YU}^2(k)}{S_{Y_0U_0}(j\omega_k)}\right)\right) \tag{2-51}$$

so that for M sufficiently large, the following approximate expression can be used:

$$\sigma_{\hat{G}}^2(k) \approx \frac{|G_0(j\omega_k)|^2}{M}\left(\frac{\sigma_Y^2(k)}{S_{Y_0Y_0}(j\omega_k)} + \frac{\sigma_U^2(k)}{S_{U_0U_0}(j\omega_k)} - 2\mathrm{Re}\left(\frac{\sigma_{YU}^2(k)}{S_{Y_0U_0}(j\omega_k)}\right)\right) \tag{2-52}$$

This expression is similar to (2-32) and shows that the uncertainty $\sigma_G(k)$ decreases as $O(M^{-1/2})$. However, for small M,

$$S_{U_0U_0}^{(M)}(j\omega_k) = \frac{1}{M}\sum_{l=1}^{M}|U_0^{[l]}(k)|^2 \tag{2-53}$$

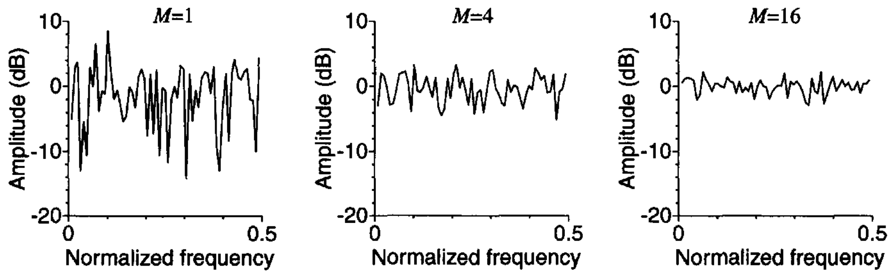


Figure 2-21. Realized power spectrum $S_{UU}^{(M)}(j\omega_k)$ for a white noise sequence ($M = 1, 4, 16$). Note that for a periodic signal a flat line at 0 dB would be found.

which can be significantly different from $S_{U_0U_0}(j\omega_k)$; thus Eq. (2-50) should be used. At some frequencies large drops in the realized power spectrum can appear, jeopardizing the FRF measurement completely. Therefore, it is strongly advised to average over a number of blocks to avoid these dips. In Figure 2-21 the realized power spectrum $S_{U_0U_0}^{(M)}(j\omega_k)$, after processing M blocks of a white noise excitation, is shown in dB (S_{xx} in dB is given by $10\log_{10}S_{xx}$). It is clearly seen that, compared with the limit value $S_{U_0U_0}^{(M)}(j\omega_k)$ (a constant value of 0 dB) for $M \to \infty$, a significant loss can occur. The normalized power spectrum $2MS_{U_0U_0}^{(M)}(j\omega)/S_{U_0U_0}(j\omega)$ is χ^2 distributed, having $2M$ degrees of freedom because it consists of the sum of $2M$ squared, independent, zero mean, normally distributed variables with equal variance (the real and imaginary part). In Table 2-1 the 95% uncertainty regions of the amplitude spectrum are described by their upper and lower bounds. The ratio of the lower bound to the rms value is also tabulated to illustrate the loss in SNR of the weakest components because of the stochastic nature of the excitations.

TABLE 2-1 Study of the Stochastic Behavior of the Averaged Spectrum of a Random Signal

N	Ratio 95% Upper/95% Lower Bound (dB)	Ratio 1/95% Lower Bound (dB)
1	22	13
2	14	7.5
4	9	4.7
8	6.2	3.0
16	4.3	2.1
32	3.1	1.4
64	2.1	1.0
128	1.5	0.7
256	1.1	0.5

In Figure 2-22 the loss in the SNR for random signals when compared with a deterministic signal with flat amplitude spectrum is shown as a function of the number of processed blocks M. It shows that for small M the SNR increases very rapidly because dips in the averaged input power spectrum disappear. It also shows that four experiments are needed to guarantee that 95% of the measurement points have an SNR corresponding to that of a well-designed, deterministic excitation after only one period (SNR normalized at 0 dB). This is one of the reasons why we strongly advocate the use of periodic excitations.

The coherence $\gamma^2(\omega_k)$, as given in Eq. (2-39), can be used again to give an overall impression of the quality of the measurement. In practice, the variance on the FRF is estimated from the coherence using Eq. (2-42).

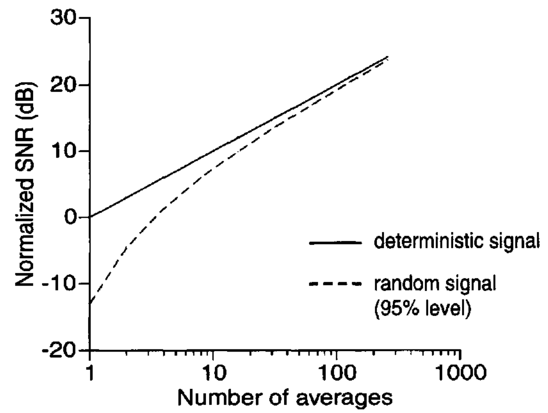


Figure 2-22. Loss in SNR for random excitations as a function of the number of processed blocks M.

2.6.3 Leakage Errors

In the previous section we assumed that it was possible to pass, easily, from a continuous-time signal $u(t)$ to its Fourier transform $U(j\omega)$. In practice, the DFT of random signals suffers from leakage errors (see Section 2.2.2). So even for undisturbed signals ($n_u = 0$, $n_y(t) = 0$, and $n_g(t) = 0$ in Figure 2-16) the FRF measurement is incorrect, that is, $G_0(j\omega_k) \neq Y_0(k)/U_0(k)$, where $U_0(k)$, $Y_0(k)$ denote the DFT of u_0, y_0. Ljung (1999) shows for discrete-time systems that the error on the FRF disappears as $O(N^{-1/2})$. This result can also be extended to FRF measurements of continuous-time systems. This is formulated precisely in the following theorem.

Theorem 2.6 (Leakage Errors on FRF Measurements of Continuous-Time Systems): Consider the signals $y(t)$ and $u(t)$ obeying Assumption 2.2 and related by the strictly stable system $G(j\omega) = F\{g(t)\}$ $(y(t) = g(t) * u(t))$. Let

$$U(k) = \frac{1}{\sqrt{N}}\sum_{n=0}^{N-1} u(nT_s)e^{-j2\pi kn/N}, \quad Y(k) = \frac{1}{\sqrt{N}}\sum_{n=0}^{N-1} y(nT_s)e^{-j2\pi kn/N} \tag{2-54}$$

be the DFT spectra of the sampled signals $u(nT_s)$ and $y(nT_s)$. If $u(nT_s)$ is uniformly bounded, filtered white noise, then

$$Y(k) = G(j\omega_k)U(k) + R_N(k) \tag{2-55}$$

with $R_N(k) = O(N^{-1/2})$ uniformly over the frequency k.

Proof. See Appendix 2.B. □

Remarks

(i) The DFT for random signals is defined with a scaling factor $1/\sqrt{N}$ so that the DTF spectrum behaves as $O(N^0)$.

(ii) If the excitation signal is a periodic signal and the number of data points is increased by repeating this signal (so that no additional frequencies are excited), the previous result can be formulated more strongly as $|R_N(k)| \leq O(N^{-1})$.

(iii) This theorem shows that the leakage error decreases with an increasing number of data, but it does not guarantee that the errors are small for finite N.

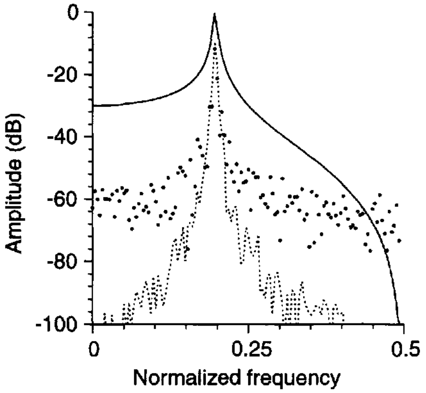


Figure 2-23. Illustration of the leakage effect. ——: $G_0(e^{-j\omega T_s})$, +: complex errors with a rectangular window, ---- complex errors with a Hanning window.

Example 2.7 (Leakage Errors on FRF Measurement): To illustrate the impact of the leakage effect, a simulation is made on a second-order discrete-time system with a narrow resonance peak of 30 dB. The system is driven with white normally distributed noise, without disturbing noise. The record is split into $M = 100$ subrecords of length 256 data points each. Next, the FRF is estimated using (2-43) and the results are shown in Figure 2-23 for a rectangular window and a Hanning window (Section 2.2.3). The errors can become very large, especially around the resonance frequency, where fast variations of the FRF occur. Replacing the rectangular window with a Hanning window reduces the errors significantly at most frequencies, but the problem at the resonance persists. Note also that these results are obtained after 100 averages. So the systematic errors dominate in these results, which shows that leakage not only increases random errors but also creates a bias. These errors are proportional to the second derivative $d^2G_0(j\omega)/d\omega^2$ (Bendat and Piersol, 1980). In Figure 2-24 the coherence calculated with Eq. (2-41) is shown. Although it is poor everywhere for the rectangular window, it is quite good for the Hanning window except around the resonance frequency. □

Conclusion: Leakage can jeopardize the quality of the FRF measurements significantly. Averaging reduces the random appearance, resulting in smoother measurements, but cannot eliminate the systematic errors. Using other windows makes it possible to reshape the leakage errors, but they remain large in the frequency bands with fast variations of the FRF. Often these bands carry most information (e.g., the resonance frequency). To avoid leakage, the best solution is use of periodic excitations and measurements of an integer number of periods. An alternative method is given in Section 2.6.4 which also gives more insight into the nature of leakage errors on FRF measurements. A last possibility is to use burst random excitations as explained in Section 2.2.4.

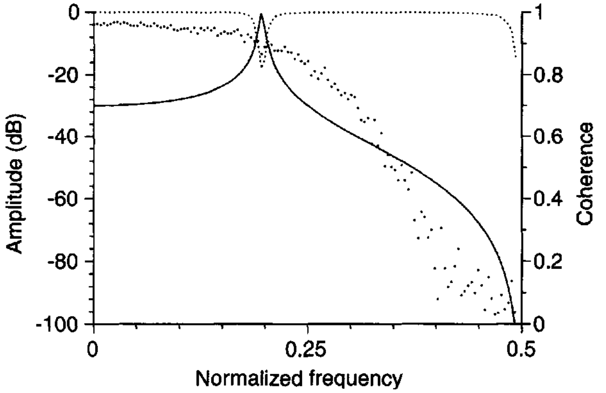


Figure 2-24. Coherence of the measurements in Figure 2-23. ——: $G_0(e^{-j\omega T_s})$, +: coherence with a rectangular window, ---- coherence with a Hanning window.

2.6.4 Improved FRF Measurements for Random Excitations

In the previous section it was illustrated that FRF measurement with random excitations are prone to increased systematic and stochastic errors. Using periodic excitations and measuring an integer number of periods eliminates the leakage errors completely. In this case also the bias errors are reduced significantly as shown in Eq. (2-28). *Hence, periodic excitations should be used to measure the FRF whenever possible.* Sometimes it is impossible, for technical or psychological reasons, to apply periodic excitations and we have to stick to noise excitations, causing the measurements of the input and output spectrum to be distorted by the leakage effect. Because the FRF is measured as the division of these spectra, it seems obvious that this result also suffers from leakage. However, a correct measurement of the input-output spectrum is a sufficient but not a necessary condition to get good FRF measurements. If the linear relation between the input and output signals is maintained, it should be possible to extract the exact FRF measurements. A detailed analysis shows that the underlying error mechanism is actually a transient phenomenon (see Chapter 5, Section 5.3.2). This is illustrated in Figure 2-25. The measurements in subrecord $[l]$ depend not only on the input signal $u^{[l]}(t)$ but also on the tail of the response to $u^{[l-1]}(t)$, while the tail of the response to $u^{[l]}(t)$ is added to the next subrecord $[l+1]$. If the system and the excitation are known, it is possible to calculate these tails and to compensate for their presence or absence. The basic idea of this method is to approximate the output of the system using an intermediate parametric transfer function model. Next, the FRF between the residuals (the difference between the modeled and the measured output) and the input is calculated using the classical methods (for example, using Eq. 2-43) and the final FRF estimate is obtained as the sum of the transfer function of the parametric model and the FRF of the residuals. This is explained subsequently in more detail.

Consider a discrete-time system $G_d(e^{-j\omega T_s})$ with impulse response $g(nT_s)$, approximating $G(j\omega)$ in the frequency band $[0, \omega_c]$, with $\omega_c \le \pi f_s$. Calculate an approximate output

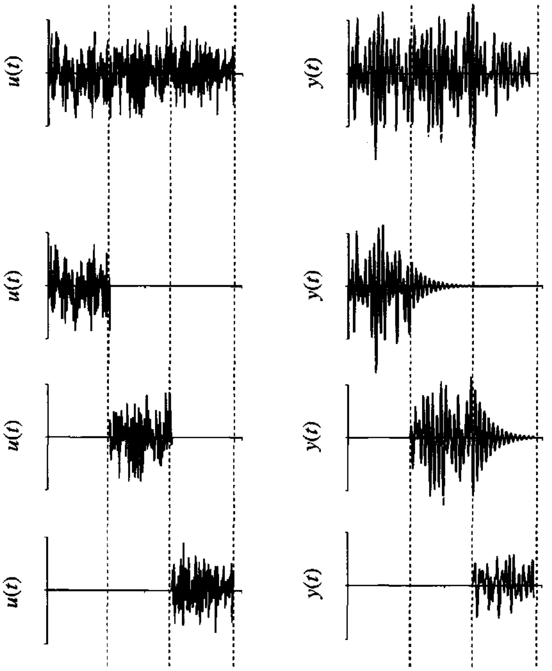


Figure 2-25. Interpretation of the leakage error as a transient effect.

$$\hat{y}(nT_s) = g(nT_s)*u(nT_s) \tag{2-56}$$

and consider the residuals

$$e(nT_s) = y(nT_s) - \hat{y}(nT_s) \tag{2-57}$$

The new FRF estimate is then given as

$$\tilde{G}(j\omega_k) = G_d(e^{-j\omega_k T_s}) + \hat{S}_{EU}(j\omega_k)/\hat{S}_{UU}(j\omega_k) \tag{2-58}$$

This algorithm allows us to shift the leakage problem to the choice of a good discrete-time approximate system $g(nT_s)$ in the frequency band of interest. A simple model $g(nT_s)$ in (2-56) is generated using an FIR filter with impulse response given by the inverse DFT (IDFT) of the FRF $\hat{G}(j\omega_k)$ obtained in (2-43):

$$g(nT_s) = \text{IDFT}(\hat{G}(j\omega_k)) \tag{2-59}$$

It is shown that the method provides better estimates of $G(j\omega_k)$ without increasing the noise sensitivity under the condition that the subrecords are long enough to capture the impulse response (e.g., 5 to 10 times the dominating time constant) (Schoukens et al., 1998c). By selecting the optimal length of the FIR filter $g(nT_s)$, the systematic errors (the FIR filter is too short) are balanced with the noise sensitivity (the FIR filter is too long). The algorithm is implemented in a dedicated routine, omitting all these choices and questions for the user. The practical implementation details of the method are given in Schoukens et al. (1998c).

As an illustration, the experimental results for a bandpass filter are presented. First, the filter is excited with a periodic excitation signal; next a binary random excitation is applied. Both signals have a peak value of 0.2 V and excite the full band (up to half the sampling frequency). The signals are generated and measured with a sampling frequency of 4.8828 kHz. $M = 9$ periods (1024 points/period) of the periodic signal and $M = 9$ blocks of the random signals (1024 points/block) are processed. The measurements were carried out with a VXI measurement setup (generator HP E1445A, acquisition HP1430A). All measurements were alias proof. The results in Figure 2-26 show a significant reduction of the leakage errors.

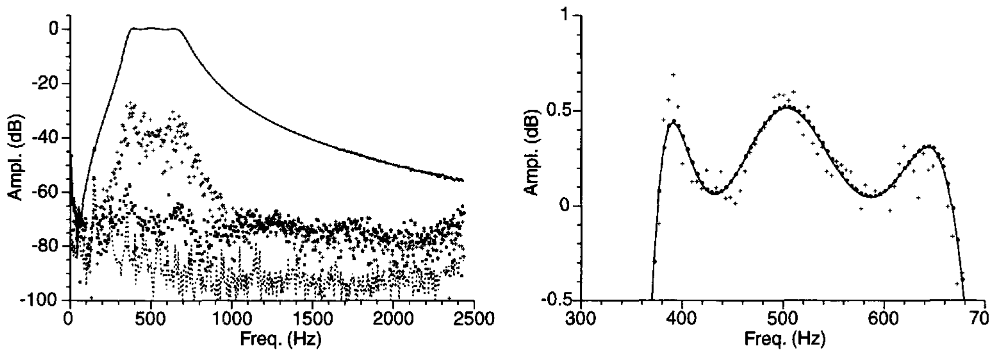


Figure 2-26. Experimental verification of the improved FRF measurements on a bandpass filter. Left: full band, ___ reference value from periodic data, ... standard deviation on the reference value, + complex error of the classical method with Hanning window, ▪ complex error for the new method; right: zoom on the passband, ___ reference value, + classical method with Hanning window, ▪ new method.

2.7 FRF MEASUREMENTS OF MULTIPLE INPUT, MULTIPLE OUTPUT SYSTEMS

The results that are presented in this chapter are also valid for multiple input, multiple output (MIMO) systems. Excitation signals that are suitable for single input, single output (SISO) systems also form a good basis to start MIMO measurements. However, additional precautions have to be taken because the FRF of a MIMO system is described by a matrix at each frequency:

$$G(j\omega_k) \in \mathbb{C}^{n_y \times n_u} \tag{2-60}$$

with n_u and n_y the numbers of inputs and outputs of the system. At least n_u different excitations are needed to extract G from the data. This can be done by cutting a random excitation record in n_u subrecords or by applying n_u different (combinations of) periodic excitations. The relation between the input and output is

$$\mathbf{Y}(k) = G(j\omega_k)\mathbf{U}(k) \tag{2-61}$$

with $\mathbf{U}(k) \in \mathbb{C}^{n_u \times n_u}$, $\mathbf{Y}(k) \in \mathbb{C}^{n_y \times n_u}$, and the entry $\mathbf{X}_{[p,q]}(k)$ corresponds to the pth input-output signal and the qth subrecord or periodic excitation. The estimate is then obtained from

$$\hat{G}(j\omega_k) = \mathbf{Y}(k)\mathbf{U}^{-1}(k) \tag{2-62}$$

It is clear that this puts a strong condition on the excitation design: the matrix $\mathbf{U}(k)$ should be regular, otherwise $\hat{G}(j\omega_k)$ is not identifiable for the given experiment. Also, the uncertainty on $\hat{G}(j\omega_k)$ depends strongly on $\mathbf{U}(k)$, and a careful design is necessary in order to avoid deterioration of the results. In case of two inputs ($n_u = 2$) it is shown that an optimal choice (maximizing $\det(\mathbf{U}(k))$) using periodic excitations is given by:

$$\mathbf{U}(k) = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} U(k) \tag{2-63}$$

with $U(k)$ the DFT spectrum of one excitation signal (Guillaume et al., 1996b). This means that in the first experiment, both inputs are excited with the same periodic excitation, while in the second experiment the sign of the second input is changed. This strategy can be generalized to more inputs. Sometimes the number of experiments is even higher than the number of inputs. In that case $\mathbf{U}^{-1}(k)$ is replaced by the Moore-Penrose pseudoinverse $\mathbf{U}^{+}(k)$, and although the previous strategy still results in good designs, the optimality cannot be shown anymore.

2.8 GUIDELINES FOR FRF MEASUREMENTS

The aim of this section is to condense the information from the previous section to a short list of guidelines. Following these guidelines does not always guarantee good measurements but at least ensures avoidance of a number of common mistakes.

2.8.1 Guideline 1: Use Periodic Excitations

We strongly advocate periodic excitations instead of random excitations because the former lead to consistent estimates, even in feedback (see Section 2.5.1), and allow estimation simultaneously with the (co-)variances of the noise. The following are recommended in order of importance: (i) measure multiple periods in one record; (ii) select a good synchronization; (iii) collect a number of single measurements. We ^dvise using random excitations only if there are strong contraindications against periodic excitations (which the authors are not aware of). The design of periodic and random excitations is discussed in detail in Chapter 4.

2.8.2 Guideline 2: Select the Best FRF Estimator

2.8.2.1 Periodic Excitations. Use $\hat{G}_{\mathrm{ML}}(j\omega_k)$ if multiple periods are measured or if repeated measurements with good synchronization are made (Section 2.5.2); otherwise, in case of poor or no synchronization, select $G_{\mathrm{H_{log}}}(j\omega_k)$ (2-37), $\hat{G}_{\mathrm{H1}}(j\omega_k)$ (2-43), or $\hat{G}_{\mathrm{H2}}(j\omega_k)$ (2-46) depending on the SNR of the measurements using Figure 2-27. Use a rectangular window in the DFT.

Remark. If it is impossible to measure an integer number of periods precisely (even after selecting a smaller number of samples), a Hanning window can be used to reduce the errors from $O(N^{-1})$ to $O(N^{-2})$ at the excited frequency lines (see Section 2.2.3) if at least four periods are captured.

2.8.2.2 Random Excitations. Select a Hanning window (Section 2.2.3) in the DFT to reduce the leakage errors. Use $\hat{G}_{\mathrm{H1}}(j\omega_k)$ (2-43) if the input SNR is best and $\hat{G}_{\mathrm{H2}}(j\omega_k)$ (2-46) if the output SNR output is best to estimate the FRF. Keep in mind that the measurements are biased if both input and output are prone to noise distortions. If the impulse response of the system is not longer than the window length, improved results can be obtained using the transient compensating method in Section 2.6.4.

2.8.3 Guideline 3: Pretreatment of Data

Before processing the data, we strongly advise effecting a visual inspection for anomalies such as (periodic) spikes, outliers, overload, drift, and offset. Some of these problems can also be detected automatically. A slow drift can be removed by use of polynomial regression

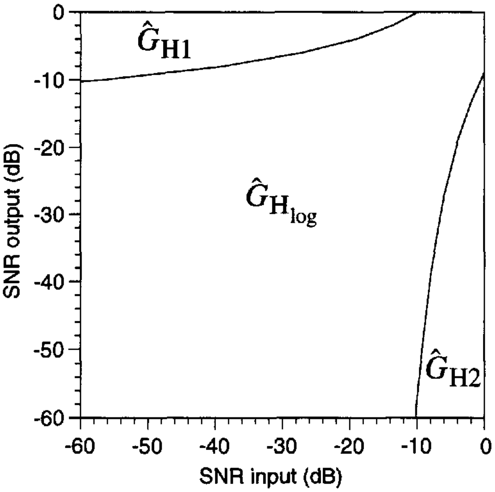


Figure 2-27. Selection between $G_{\mathrm{H_{log}}}(j\omega_k)$, $\hat{G}_{\mathrm{H1}}(j\omega_k)$, and $\hat{G}_{\mathrm{H2}}(j\omega_k)$ as a function of the SNR in case the repeated measurements are not well synchronized.

(McCormack et al., 1994a; Peirlinckx et al., 1996). Outliers can be detected using the periodic nature of the excitation by observing the variations from one period to the next. A sound solution is to perform a new experiment. If this is not possible, a simple alternative is to replace the erroneous data by the equivalent value of the neighboring periods.

2.9 CONCLUSION

FRF measurements give a great deal of information about the device or plant under test. Very often the FRF is easily accessible and it is strongly advised to take this intermediate step in the identification process. It provides not only much qualitative information about the complexity of the problem but also quantitative information about the plant and the measurement quality. This can be used to set up a measurement-driven weighting function for the identification step and also gives very valuable information for the model validation. The user has significant influence on the measurement quality by generating a good excitation and selecting the proper algorithms to process the raw measurement data. For these reasons, we strongly encourage the reader to take the time to understand the basic principles of FRF measurements. Good nonparametric measurements will simplify the task of building parametric models significantly.

2.10 EXERCISES

Remark. In these exercises (and also in the next chapters) we will use the Matlab® notation. Matlab® is a high-performance language for technical computing developed by Mathworks Inc. More information can be found at http://www.mathworks.com/

2.1. Calculate the signal $u_0(t) = \sum_{l=1}^{255} A_l \sin(2\pi f_0 l t T_s + \phi_l)$, $t = 0, \dots, N-1$ with $f_0 = 1$, $T_s = 1/1024$, and ϕ_l independently and uniformly distributed in, $[0, 2\pi[$. Calculate $U_0(k) = \text{DFT}(u_0(t))$ using the Matlab FFT instruction for $N = 1024, 1500, 4096, 5000$ and plot the amplitude spectrum in dB $U_{\text{dB}}(k) = 20\log_{10}|U_0(k)|$. Use for the first time a rectangular window and for the second time a Hanning window, and discuss your results. (What is the impact of leakage? What happens if a Hanning window is applied on a record consisting of an integer number of periods?)

Note that this routine also works for $N \neq 2^n$ but that it becomes significant more slowly.

2.2. In this exercise it is shown how a very fast calculation of a periodic signal is achieved, starting from its spectrum.

Define $\underset{\sim}{U} = \text{ZEROS}(1024, 1)$ %; this is a 1024×1 vector with all entries zero.

Set $\underset{\sim}{U}(2{:}256) = \exp(j\text{rand}(255, 1))$.

$u = 2{*}\text{REAL}(\text{IFFT}(\underset{\sim}{U}))$

Compare the computational effort of this approach with that of Exercise 2.1 Give an explanation of the algorithm.

Remark: In Matlab $\underset{\sim}{U}(1)$ contains the DC component and $\underset{\sim}{U}(k)$ the Fourier coefficient of the harmonic $(k-1)f_0$. The underscore $\underset{\sim}{\ }$ indicates that we refer to a spectrum in the Matlab notation: $\underset{\sim}{U}(k) = U(k+1)$.

2.3. Calculate one period of $u_0(t)$ with random phased components (with unit amplitude) at the frequencies lf_0, $l = 1, 2, \dots, 255$, $f_0 = 10$ Hz putting $N_p = 512$ points in one period. Use the method of Exercise 2.2

What is the sample period T_s that is needed to generate this signal?

Set up the signals $u_{0_l} \in \mathbb{R}^{lN_p \times 1}$ containing l successive periods using the REPMAT instruction for $l = 1, 2, 4$ and study the relation between the fundamental frequency, N_p, and the line number of the spectral components of the repeated signal.

2.4. Define the discrete-time system $G_0(z^{-1})$: [b,a] = CHEBY1(2, 10, 0.5) (this is a second-order system with resonance frequency at $0.25 f_s$).
Plot the amplitude of the transfer function of this system in dB (use the function FREQZ). Consider the signals u_{0_l} of Exercise 2.3 and calculate the responses $y_{0_l}(t) = g_0(t) * u_{0_l}(t)$ using the filter operation $y_{0_l} = \text{FILTER}(b, a, u_{0_l})$ of Matlab.
Estimate the FRF of the system at the excited frequency lines as $\hat{G}_l(z_r^{-1}) = \underset{\sim}{Y}_{0_l}(k_r)/\underset{\sim}{U}_{0_l}(k_r)$. The indices k_r should be properly chosen to select only the lines where the system is excited. Compare the measured FRF with the exact one and discuss the result. What is the origin of the errors?

2.5. Repeat the previous exercise for $l = 2$ but eliminate the first period in u_{0_2}, y_{0_2} before calculating the DFT spectra. Explain why the errors disappeared.

2.6. Generate an iid random signal with zero mean $u_0(t)$, $t = 1, \ldots, 512M$. Normalize the rms value of this signal to 1. Calculate $y_0 = g_0(t) * u_0(t)$ (Exercise 2.4) and estimate $\hat{G}_{\text{H1}}$ (2-43) for $M = 1, 4, 16, 64$. Discuss the results. Repeat the exercise but this time eliminate the transient effects using the technique of Exercise 2.5.

2.7. Generate an iid random signal with zero mean $u_0(t)$, $t = 1, \ldots, 512M$. Normalize the rms value of this signal to 1. Calculate $y(t) = g_0(t) * u_0(t) + n_y(t)$ (Exercise 2.4) with $n_y(t)$ iid normally distributed noise with zero mean and $\sigma_y = 0.1$. Estimate $\hat{G}_{\text{H1}}$ (2-43) for $M = 1, 4, 16, 64$. Discuss the results.

2.8. Generate an iid random signal with zero mean $u_0(t)$, $t = 1, \ldots, 512M$. Normalize the rms value of this signal to 1. Calculate $y_0 = g_0(t) * u_0(t)$ (Exercise 2.4). Generate $u(t) = u_0(t) + n_u(t)$, and $y(t) = y_0(t) + n_y(t)$, with $n_u(t), n_y(t)$ iid normally distributed noise with zero mean and $\sigma_u = 0.5$, $\sigma_y = 0.1$. Estimate $\hat{G}_{\text{H1}}(j\omega_k)$ (2-43) for $M = 1, 4, 16, 64$. Discuss the results. Can you suggest a better method?

2.9. Estimate the variance of $\hat{G}_{\text{H1}}(j\omega_k)$ (2-43) for the setup of Exercise 2.7 using the coherence $\gamma^2(\omega_k)$. Put $M = 16$. Repeat the simulation 50 times and calculate $\sigma_{\hat{G}}^2(k)$ from the repeated estimates. Compare both results.

2.10. Consider the signal $u_{0_{16}}$ of Exercise 2.4 and calculate the output for the input $u(t) = u_{0_{16}}(t) + n_g(t)$ where $n_g(t)$ is zero mean iid generator noise with $\sigma_{n_g} = 0.1$. Calculate the system output $y(t) = g(t) * u(t)$ and skip the first period to avoid transients. Calculate the FRF $\hat{G}_{\text{ML}}(j\omega_k)$ (2-27) and estimate the variance of the FRF for this setup using the coherence. Explain why the impact of generator noise on the variance is so small.

2.11 APPENDIXES

Appendix 2.A Asymptotic Behavior of Averaging Techniques

The proofs of the asymptotic ($M \to \infty$) properties of the averaging techniques (2-27), (2-43), and (2-46) follow the lines of Sections 14.13 (general theory) and 14.15 (application on the measurement of a resistance). To understand these proofs fully we advise reading Sections 14.13 and 14.15 first.

We will prove the results for the ML estimator (2-27); the proofs for the H_1 (2-43) and H_2 (2-46) methods follow exactly the same lines. The only difference is that the H_1 and H_2 methods require the existence of the fourth-order moments of the disturbing noise instead of the second-order moments for the ML method (2-27). This is due to the squaring operation of the noise in (2-43) and (2-46). We split the proof in two parts: (i) the data blocks (subrecords) are independent, Assumption 2.5(i), and (ii) the data blocks are correlated, Assumption 2.5(ii).

2.A.1 Independent Data Blocks. In (2-27) sums of the form

$$S(M)/M = \frac{1}{M}\sum_{l=1}^{M} N^{[l]}(k) \tag{2-64}$$

occur with $N^{[l]}(k)$ the DFT of $n_u^{[l]}(t)$ or $n_y^{[l]}(t)$, $t = 0, 1, \ldots, N_\mathrm{p} - 1$. Under Assumption 2.5(i, $P = 2$) the noise $N^{[l]}(k)$, $l = 1, 2, \ldots, M$, is independent over l and has finite second-order moments. Hence, $S(M)/M$ converges with probability one (w.p. 1) at the rate $O_\mathrm{p}(M^{-1/2})$ to its expected value (see Section 14.9, version 2 of the law of large numbers). The expected value of $S(M)$ is zero because

$$\mathcal{E}\{N^{[l]}(k)\} = N_\mathrm{p}^{-1/2}\sum_{t=0}^{N_\mathrm{p}-1} \mu_n^{[l]} e^{-j2\pi kt/N_\mathrm{p}} = 0 \text{ for } k \neq 0$$

where $\mu_n^{[l]} = \mathcal{E}\{n^{[l]}(t)\}$. Using the results of Sections 14.13.1 and 14.13.2 it follows directly that the estimate $\hat{G}_\mathrm{ML}(j\omega_k)$ (2-27) converges w.p. 1 (almost surely) at the rate $O_\mathrm{p}(M^{-1/2})$ to $G_0(j\omega_k)$.

Under Assumption 2.5(i, $P = 2 + \varepsilon$), the noise $N^{[l]}(k)$ is independent over l and has finite moments of order $2 + \varepsilon$. Hence, $S(M)/\sqrt{M}$ is asymptotically normally distributed (see Section 14.10, version 2 of the central limit theorem). Using the results of Section 14.13.4 it follows directly that $\hat{G}_\mathrm{ML}(j\omega_k)$ is asymptotically normally distributed and that its variance is asymptotically given by

$$\sigma_{\hat{G}}^2(k) = \frac{|G_0(j\omega_k)|^2}{M}\left(\frac{\sigma_{\hat{Y}}^2(k)}{|Y_0(k)|^2} + \frac{\sigma_{\hat{U}}^2(k)}{|U_0(k)|^2} - 2\mathrm{Re}\left(\frac{\sigma_{\hat{Y}U}^2(k)}{Y_0(k)\overline{U}_0(k)}\right)\right)$$

where $\sigma_{\hat{U}}^2(k)$, $\sigma_{\hat{Y}}^2(k)$, and $\sigma_{\hat{Y}U}^2(k)$ are the noise (co-)variances of one data block (subrecord).

2.A.2 Correlated Data Blocks. The proof follows the same lines of the previous section. The only difference is that other versions of the strong law of large numbers and the central limit theorem are used. The sum (2-64) can be written as

$$S(M)/M = \mathrm{DFT}(s(M)/M) \text{ where } s(M)/M = \frac{1}{M}\sum_{l=1}^{M} n^{[l]}(t) \tag{2-65}$$

with $n^{[l]}(t) = n_u(t)$ or $n_y(t)$. Under Assumption 2.5(ii, P) the disturbing noise $n(t)$ in (2-65) can be written as filtered white noise $e(t)$ with finite moments of order P, so that $n(t)$ is mixing over t of order P (see Example 14.6). Hence, the subrecord $n^{[l]}(t) = n(t + lN_\mathrm{p})$ is mixing over l of order P. We conclude that under Assumption 2.5(ii, $P = 2$), the sum $s(M)/M$ converges w.p. 1 at the rate $O_\mathrm{p}(M^{-1/2})$ to its expected value (see Section 14.9, version 3 of the law of large numbers), and that under Assumption 2.5(ii, $P = \infty$), $s(M)/\sqrt{M}$ is asymptotically normally distributed (see Section 14.10, version 3 of the central limit theorem). This is also valid for $S(M)/M$, with $\mathcal{E}\{S(M)\} = 0$, because the number of elements N_p in the DFT sum does not increase with M. Using the results of Sections 14.13.1, 14.13.2 and 14.13.4 it follows directly that the estimate $\hat{G}_\mathrm{ML}(j\omega_k)$ (2-27) converges w.p. 1 (almost surely) at the rate $O_\mathrm{p}(M^{-1/2})$ to $G_0(j\omega_k)$ and that $\hat{G}_\mathrm{ML}(j\omega_k)$ is asymptotically normally distributed.

Appendix 2.B Proof of Theorem 2.6 (On Decaying Leakage Errors)

Proof. For notational simplicity, we omit the subscript 0 to indicate the undisturbed variables. We have by definition $y(t) = \int_0^{\infty} g(\tau)u(t-\tau)d\tau$ so that

$$Y(k) = \frac{1}{\sqrt{N}}\sum\nolimits_{t=0}^{N-1} y(tT_s)e^{-j\omega_k tT_s}$$

$$= \int_0^{\infty} g(\tau)\frac{1}{\sqrt{N}}\sum\nolimits_{t=0}^{N-1} u(tT_s-\tau)e^{-j\omega_k tT_s}d\tau$$

$$= \int_0^{\infty} g(\tau)e^{-j\omega_k\tau}\frac{1}{\sqrt{N}}\sum\nolimits_{t=0}^{N-1} u(tT_s-\tau)e^{-j\omega_k(tT_s-\tau)}d\tau$$

with $\omega_k = 2\pi k f_s/N$. Define l such that $\tau = lT_s + \varepsilon$ with $0 \le \varepsilon < T_s$ and change variable $t - l \to t$

$$Y(k) = \int_0^{\infty} g(\tau)e^{-j\omega_k\tau}\frac{1}{\sqrt{N}}\sum\nolimits_{t=-l}^{N-1-l} u(tT_s-\varepsilon)e^{-j\omega_k(tT_s-\varepsilon)}d\tau$$

Calculating the difference $Y(k) - G(j\omega_k)U(k)$, using $G(j\omega_k) = \int_0^{\infty} g(\tau)e^{-j\omega_k\tau}d\tau$, gives

$$Y(k) - G(j\omega_k)U(k) = \int_0^{\infty} g(\tau)e^{-j\omega_k\tau}\left(\frac{1}{\sqrt{N}}\sum\nolimits_{t=-l}^{N-1-l} u(tT_s-\varepsilon)e^{-j\omega_k(tT_s-\varepsilon)} - U(k)\right)d\tau \quad (2\text{-}66)$$

The absolute value of (2-66) can be bounded above by

$$\begin{aligned} |Y(k) - G(j\omega_k)U(k)| &\le \int_0^{\infty}|g(\tau)|\left|\frac{1}{\sqrt{N}}\sum\nolimits_{t=-l}^{N-1-l} u(tT_s-\varepsilon)e^{-j\omega_k(tT_s-\varepsilon)} - U(k)\right|d\tau \\ &\le N^{-1/2}\left(C_1\int_0^{\infty}\tau|g(\tau)|d\tau + C_2\int_0^{\infty}|g(\tau)|d\tau\right) \\ &\le C_3N^{-1/2} \end{aligned} \quad (2\text{-}67)$$

where the second inequality in (2-67) is due to Lemma 2.8 (see below) and with C_3 independent of k. □

Lemma 2.8 For a strictly stable system excited by uniformly bounded ($|u(t)| \le C_u$ for any t) filtered white noise we have

$$\left|\frac{1}{\sqrt{N}}\sum\nolimits_{t=-l}^{N-1-l} u(tT_s-\varepsilon)e^{-j\omega_k(tT_s-\varepsilon)} - U(k)\right| \le N^{-1/2}(C_1\tau + C_2) \quad (2\text{-}68)$$

with C_1 and C_2 constants independent of k, ε, l, and N.

Proof. Using $\sum_{t=-l}^{N-1-l} = \sum_{t=-l}^{-1} + \sum_{t=0}^{N-1} - \sum_{t=N-l}^{N-1}$, the left-hand side of (2-68) is bounded above by

$$\begin{aligned} \left|\frac{1}{\sqrt{N}}\sum_{t=-l}^{N-1-l} u(tT_s-\varepsilon)e^{-j\omega_k(tT_s-\varepsilon)} - U(k)\right| &\le \left|\frac{1}{\sqrt{N}}\sum_{t=-l}^{-1} u(tT_s-\varepsilon)e^{-j\omega_k(tT_s-\varepsilon)}\right| \\ &+ \left|\frac{1}{\sqrt{N}}\sum_{t=0}^{N-1} u(tT_s-\varepsilon)e^{-j\omega_k(tT_s-\varepsilon)} - U(k)\right| + \left|\frac{1}{\sqrt{N}}\sum_{t=N-l}^{N-1} u(tT_s-\varepsilon)e^{-j\omega_k(tT_s-\varepsilon)}\right| \end{aligned} \quad (2\text{-}69)$$

The sum of the first and third terms of the right-hand side of (2-69) is bounded by

$$\left|\frac{1}{\sqrt{N}}\sum_{t=-l}^{-1} u(tT_s-\varepsilon)e^{-j\omega_k(tT_s-\varepsilon)}\right| + \left|\frac{1}{\sqrt{N}}\sum_{t=N-l}^{N-1} u(tT_s-\varepsilon)e^{-j\omega_k(tT_s-\varepsilon)}\right| \le \frac{2lC_u}{\sqrt{N}} \le \frac{\tau C_1}{\sqrt{N}} \tag{2-70}$$

where the last inequality stems from $l = (\tau-\varepsilon)/T_s$ with $C_1 = 2C_u/T_s$. So it remains to be shown that the second term of (2-69) can be bounded as

$$\left|\frac{1}{\sqrt{N}}\sum\nolimits_{t=0}^{N-1} u(tT_s-\varepsilon)e^{-j\omega_k(tT_s-\varepsilon)} - U(k)\right| = O(N^{-1/2}). \tag{2-71}$$

with $U(k) = \frac{1}{\sqrt{N}}\sum\nolimits_{t=0}^{N-1} u(tT_s)e^{-j\omega_k tT_s}$. Consider

$$E(k) = \frac{1}{\sqrt{N}}\sum\nolimits_{k=0}^{N-1} u(tT_s-\varepsilon)e^{-j\omega_k(tT_s-\varepsilon)} - \frac{1}{\sqrt{N}}\sum\nolimits_{t=0}^{N-1} u(tT_s)e^{-j\omega_k tT_s} \tag{2-72}$$

Then

$$\begin{aligned}\mathcal{E}\{\,|E(k)|^2\} &= \frac{2}{N}\sum\nolimits_{t,s=0}^{N-1} R_{uu}((t-s)T_s)e^{-j\omega_k(t-s)T_s} \\ &\quad - \frac{2}{N}\mathrm{Re}(\sum\nolimits_{t,s=0}^{N-1} R_{uu}((t-s)T_s-\varepsilon)e^{-j\omega_k(t-s)T_s}e^{j\omega_k\varepsilon})\end{aligned} \tag{2-73}$$

or (change variables $t-s = n$ and $t = m$)

$$\begin{aligned}\mathcal{E}\{\,|E(k)|^2\} &= 2\sum\nolimits_{n=-(N-1)}^{N-1}(1-|n|/N)R_{uu}(nT_s)e^{-j\omega_k nT_s} \\ &\quad - 2\ \mathrm{Re}(\sum\nolimits_{k=-(N-1)}^{N-1}(1-|n|/N)R_{uu}(nT_s-\varepsilon)e^{-j\omega_k nT_s}e^{j\omega_k\varepsilon})\end{aligned} \tag{2-74}$$

Because $u(t)$ can be modeled as uniformly bounded filtered white noise, we have

$$\sum\nolimits_{n=N}^{\infty}|R_{uu}(nT_s\pm\varepsilon)| \le O(e^{-C_4N}) \quad \text{and} \quad \sum\nolimits_{n=-\infty}^{\infty}|n||R_{uu}(nT_s\pm\varepsilon)| = O(N^0) \tag{2-75}$$

for $0 \le \varepsilon < T_s$. Using $\sum\nolimits_{n=-(N-1)}^{N-1} = \sum\nolimits_{n=-\infty}^{\infty} - \sum\nolimits_{n=-\infty}^{-N} - \sum\nolimits_{n=N}^{\infty}$ and (2-75), (2-74) can be bounded above as

$$\begin{aligned}\mathcal{E}\{\,|E(k)|^2\} &\le 2\left|\sum_{n=-\infty}^{\infty} R_{uu}(nT_s)e^{-j\omega_k nT_s} - \mathrm{Re}(\sum_{n=-\infty}^{\infty} R_{uu}(nT_s-\varepsilon)e^{-j\omega_k nT_s}e^{j\omega_k\varepsilon})\right| \\ &\quad + O(N^{-1}) - O(e^{-C_4N})\end{aligned} \tag{2-76}$$

Because the excitation signal $u(t)$ is band limited, its power spectrum $S_{UU}(j\omega) = F\{R_{uu}(t)\}$ is zero for $|\omega| \ge \omega_s/2$, and, hence, the Fourier transform of the discrete-time sequence $R_{uu}(nT_s-\varepsilon)$ becomes

$$\sum\nolimits_{n=-\infty}^{\infty} R_{uu}(nT_s-\varepsilon)e^{-j\omega nT_s} = F\{R_{uu}(t-\varepsilon)\} = S_{UU}(j\omega)e^{-j\omega\varepsilon} \text{ for } |\omega| < \omega_s/2$$

reducing (2-76) to $\mathcal{E}\{\,|E(k)|^2\} \le O(N^{-1})$.

3

Frequency Response Function Measurements in the Presence of Nonlinear Distortions

Abstract: In this book we deal with the measurement and identification of linear dynamic systems. However, in reality the linearity assumption is only approximately valid. Many systems that are assumed to be linear are disturbed by nonlinear distortions. The aim of this chapter is not to show how nonlinear systems should be modeled because this problem is beyond the scope of this book. The goal is to provide the reader with an insight into the impact of nonlinear distortions on FRF measurements. We will also look for tools to detect, qualify, and quantify the presence of nonlinear distortions. Finally, it will be shown how we can still use the linear framework under these conditions.

3.1 INTRODUCTION

The aim of this chapter is not to model nonlinear systems because this problem is beyond the scope of this book. The goal is to provide the reader with insight into the behavior of nonlinear distortions and their impact on frequency response function (FRF) measurements. This allows not only a better understanding of the error mechanism but also knowledge that can be used during the design of the experiment in order to get the best results under the imposed operational conditions. To do so, the user should clearly specify the goal of his measurements. In order to formalize this discussion, we use the general structure given in Figure 3-1. The measured output $y(t)$ consists of a linear $y_L(t)$ and a nonlinear $y_{NL}(t)$ contribution. For simplicity we assume that the linear contribution dominates the nonlinear one for sufficiently small inputs:

$$\lim_{u_{RMS} \to 0} \frac{(y_{NL})_{rms}}{(y_L)_{rms}} = 0 \qquad (3\text{-}1)$$

Under this assumption we have two basic options: (i) The goal of the measurement is to get the FRF of the underlying linear system, minimizing the impact of the NLS on the measurements. If (3-1) is not valid, the theory that is developed in this chapter is still applicable, but it

System Identification: A Frequency Domain Approach. Rik Pintelon and Johan Schoukens
Copyright © 2001 Institute of Electrical and Electronics Engineers. ISBN: 0-780-36000-1

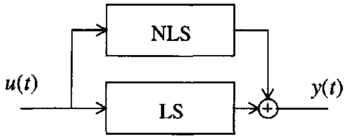


Figure 3-1. General setup of the nonlinear distortion.

is no longer possible to define an underlying linear system. (ii) Try to find the best linear approximation to the global system, including the NLS. The first option is the best choice if some underlying linear physical model exists and the user wants to identify it as well as possible. The second choice is preferred if the model will be used to describe the relation between input and output using a linear model. Then, the nonlinearity will be linearized around the operation point of the test. Both choices will be discussed in this chapter.

The chapter is structured along the following lines: a simple introduction to the behavior of nonlinear systems is given; the class of nonlinear distortions that is covered by this work is defined; detection techniques for nonlinear distortions are developed; and finally it is shown how the underlying linear system or the best linear approximation can be optimally measured.

3.2 INTUITIVE UNDERSTANDING OF THE BEHAVIOR OF NONLINEAR SYSTEMS

Consider the static nonlinear system $y = u + u^2 + u^3$ excited with a sine wave $u(t) = A\sin 2\pi f_0 t$. The response of this system is split into its linear, quadratic, and cubic contributions. The corresponding amplitude spectra are given in Figure 3-2. It shows that nonlinear systems create additional harmonics. On the one hand this allows the detection of nonlinear contributions, but it also shows that the FRF measurements are disturbed. The cubic subsystem also puts power at the original frequency f_0 that cannot be separated from the linear contributions using only a single sine measurement. More advanced methods that are beyond the scope of this book are needed to solve this problem (e.g., Bendat, 1998). In general, for a multiharmonic periodic signal, the frequencies of quadratic terms are found by looking for all combinations $f_i + f_j$ over the positive and negative frequencies of the signal. For the cubic terms triple sums $f_i + f_j + f_k$ should be considered, and in general n frequencies should be combined for a nonlinearity of degree n. This shows that for periodic signals having only odd frequency components (at $f_0, 3f_0, 5f_0, \ldots$), the even nonlinearities do not disturb the FRF measurements (the sum of two odd frequencies is always even) but it is impossible to avoid disturbances from the odd nonlinearities (e.g., $f_0 + f_0 - f_0 = f_0$).

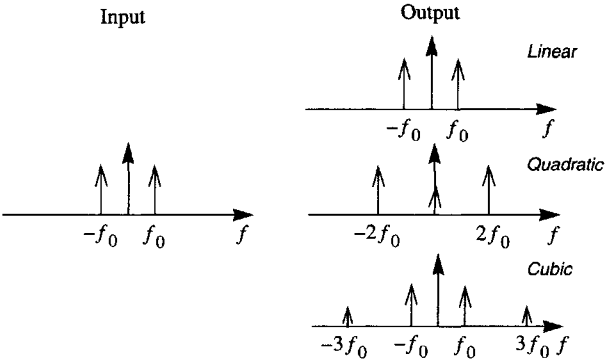


Figure 3-2. Impact of linear, quadratic, and cubic systems on the spectrum of a sine.

These results can be generalized using Volterra systems. A concise introduction to this technique is given in the book of Schetzen (1980). The basic idea is to extend the linear model to a nonlinear one using multidimensional convolutions, for example,

$$y(t) = \int_{-\infty}^{\infty} g_1(\tau)u(t-\tau)d\tau + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g_2(\tau_1, \tau_2)u(t-\tau_1)u(t-\tau_2)d\tau_1 d\tau_2 + \cdots \tag{3-2}$$

For static nonlinear systems this relation simplifies to a Taylor expansion:

$$y(t) = g_1 u(t) + g_2 u^2(t) + \cdots \tag{3-3}$$

The autocorrelation $R_{yu}(\tau)$ no longer depends on the second-order moments of u only but also on the higher order ones. Consequently, the nonlinear distortions of the FRF measurement also depend on the amplitude distribution of the excitation, for example, normally, uniformly, or binary distributed excitations. If the aim is to get the best linear approximation, it is important to use the same kind of excitations (power spectrum and amplitude distribution) as will be applied later on to the system, otherwise the linear approximation can become invalid.

For periodic excitations with N harmonics at frequencies $kf_{\max}/N$, $k = 1, \ldots, N$, relation (3-2) simplifies to a sum over all possible frequency combinations adding to the output Fourier coefficient Y_k at frequency $kf_{\max}/N$ (Chua and Ng, 1979):

$$Y_k = \sum_{\alpha=1}^{\infty} Y_k^{\alpha} \tag{3-4}$$

with Y_k^{α} the contribution of degree α

$$\begin{aligned} Y_k^{\alpha} &= \sum_{k_1, k_2, \ldots k_{\alpha-1} = -N}^{N} G^{\alpha}_{L_k, k_1, k_2, \ldots, k_{\alpha-1}} U_{k_1} U_{k_2} \ldots U_{k_{\alpha-1}} U_{L_k} \\ L_k &= k - \sum_{i=1}^{\alpha-1} k_i \end{aligned} \tag{3-5}$$

and U_r the input Fourier coefficient at frequency $rf_{\max}/N$ (see Section 2.3 for the relationship between the Fourier coefficient and the DFT spectrum of a periodic signal). $G^{\alpha}_{L_k, k_1, k_2, \ldots, k_{\alpha-1}}$ is the symmetrized frequency domain representation of the Volterra kernel of degree α (Schetzen, 1980) so that the order of the frequencies $L_k, k_1, k_2, \ldots, k_{\alpha-1}$ has no importance

$$G^{\alpha}_{k_1, k_2, \ldots, k_{\alpha}} = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} g_{\alpha}(\tau_1, \ldots, \tau_{\alpha}) e^{-j2\pi f_0 (k_1\tau_1 + \ldots k_{\alpha}\tau_{\alpha})} d\tau_1 \ldots d\tau_{\alpha} \tag{3-6}$$

The convergence of this sum is later guaranteed in Definition 3.5.

3.3 A FORMAL FRAMEWORK TO DESCRIBE NONLINEAR DISTORTIONS

Describing nonlinear systems is a tedious job because it is necessary to guarantee convergence of the Volterra series (3-4). Moreover, the limiting value also depends on the amplitude distribution of the excitation. A normally distributed excitation can result in a different limiting value than a uniform distribution, even if the power spectra of both excitation signals are the same. For these reasons it is necessary to state, precisely, the validity of these theories. This depends on the class of excitation signals and the class of nonlinear distortions that will be considered.

3.3.1 Class of Excitation Signals

As mentioned before, FRF measurements in the presence of nonlinear distortions depend on the class of excitation signals. We focus on random multisines. These are periodic random excitations with a user-defined amplitude spectrum. When an integer number of periods is measured, the amplitude spectrum is perfectly realized, which is not the case for a random excitation (see also Chapter 4 on excitation signals). All the results can be generalized easily to (periodic) random signals (random amplitude and random phase), at a price of taking an additional expectation with respect to the amplitudes in the expressions as is commented on after Theorem 3.7. This generalizes the results to the wider class of normally distributed random excitations. However, from the experimental point of view, we have a strong preference to use periodic excitations with well-controlled amplitude spectra as explained in the previous chapter.

Definition 3.1 (Random Multisine): A signal $u(t)$ is a random multisine if

$$u(t) = \sum\nolimits_{k=-N}^{N} U_k e^{j2\pi f_{\max} kt/N} \tag{3-7}$$

with $U_k = \overline{U}_{-k} = |U_k|e^{j\varphi_k}$, $f_{\max}$ the maximum frequency of the excitation signal, $N \in \mathbb{N}$ the number of frequency components, and the phases φ_k a realization of an independent distributed random process on $[0, 2\pi[$ such that $\mathcal{E}\{e^{j\varphi_k}\}=0$.

Remarks

(i) A possible choice for φ_k could be to select it as a uniformly distributed noise sequence, but other choices will also do. For example, φ_k can also be chosen to have a discrete distribution.

(ii) If the amplitude spectrum $|U_k|$ is random, then (3-7) equals periodic noise.

(iii) For simplicity U_0 is set to zero, considering the DC component as the operating point of the system. Also the output bias of the nonlinear system depends nonlinearly on the input. Consequently, linear models cannot describe the variations of the output bias as a function of the input. The DC information of the input and the output will not be used during the linear identification process.

(iv) It is strongly advised to use FFT techniques to calculate multisine signals, otherwise the computation time becomes very long (see Exercises 2.1 and 2.2).

We will study the asymptotic behavior of the nonlinear distortions for multisines with a growing number of harmonics. In order to keep excitations with a finite power for $N \to \infty$, the signals are scaled with $1/\sqrt{N}$. This leads finally to the class of normalized random multisines $\mathbb{E}_N$ and the class of periodic noise excitations $\mathbb{P}_N$ that we will use in this study.

Definition 3.2 (Normalized Random Multisine): The class of normalized random multisines $\mathbb{E}_N$ is given by the set of random multisines $u_N(t)$ (3-7) having a normalized amplitude spectrum: $|U_k| = \hat{U}(kf_{\max}/N)/\sqrt{N}$. The deterministic amplitudes $\hat{U}(kf_{\max}/N) \in \mathbb{R}^+$ are uniformly bounded, $\hat{U}(f) \leq M_U$, where the function $\hat{U}(f)$ has a finite number of discontinuities on the interval $[0, f_{\max}]$. The phases φ_k are the realization of an independent (over k) random process satisfying $E\{e^{j\varphi_k}\} = 0$. The DC component of the $u_N(t)$ is set to zero, $U_0 = 0$, and the frequency $f_{\max}$ is independent of N.

Definition 3.3 (Normalized Periodic Noise): The class of normalized periodic noise excitations $\mathbb{P}_N$ is given by the set of random multisines $u_N(t)$ (3-7) having a normalized random amplitude spectrum: $|U_k| = \hat{U}(kf_{\max}/N)/\sqrt{N}$. The amplitudes $\hat{U}(kf_{\max}/N) \in \mathbb{R}^+$ and the phases φ_k are the realization of independent (jointly, and over k) random processes satisfying the following conditions: $\hat{U}(kf_{\max}/N)$ has uniformly bounded moments of any order $\mathcal{E}\{\hat{U}^{\alpha}(f)\} \le M_{\hat{U}}^{\alpha}$, the function $\mathcal{E}\{\hat{U}^2(f)\}$ has a finite number of discontinuities on the interval $[0, f_{\max}]$, and $E\{e^{j\varphi_k}\} = 0$. The DC component of the $u_N(t)$ is set to zero, $U_0 = 0$, and the frequency $f_{\max}$ is independent of N.

Remark. In the sequel of this book reference will be made to U_k for $|k| > N$. In these cases we define $U_k = 0$ for $|k| > N$.

In the sequel of the book, a more general signal will be used. Because it is closely related to the concept of normalized multisines, we prefer to define it here. The ideas developed in this chapter can even be applied to this class of excitation signals, if some of the assumptions are modified (e.g., the convergence assumption in Definition 3.5). However, the reader should be aware that the limiting value of the measured FRF can depend on the specific signal in this generalized case.

Definition 3.4 (Normalized Periodic Signals): The class of normalized periodic signals is given by the set of periodic signals $u_N(t)$ (3-7) that have a normalized amplitude or power spectrum. For signals with a deterministic amplitude spectrum, we have $|U_k| = O(N^{-1/2})$. For signals with a random amplitude spectrum, the expected value $\mathcal{E}\{|U_k|^2\}$ is normalized: $\mathcal{E}\{|U_k|^2\} = O(1/N)$. For deterministic signals the peak value $(\max_t |u(t)| \le C < \infty$ for any t, including $t = \infty)$ should be bounded.

3.3.2 Selection of a Model Structure for the Nonlinear System

In this section we set up a mathematical description for the nonlinear distortions. Although we are not interested, at all, in extracting these models from the measurement, a formal description is needed in order to characterize and quantify the impact of the nonlinear distortions. One of the most general descriptions for nonlinear systems is the Volterra models (3-2) splitting the relation between input and output in different contributions of increasing degree of nonlinearity (Schetzen, 1980).

Convergence aspects are a central issue when dealing with these models. Uniform convergence requires that there exists an upper bound on the output error (= system output − model output) amplitude that is independent of the input and decreases to zero if the number of terms n_α in $\sum_{\alpha=1}^{n_\alpha} Y_k^{\alpha}$ goes to infinity. It can be shown only for a very restricted set of systems, e.g., the underlying nonlinear function is analytic for all considered inputs. The class of allowable systems is considerably extended if the uniform convergence is replaced by mean square convergence. In that case it is no longer necessary that the output converges everywhere in the domain of interest. Only the power (or root mean square value) of the error signal should converge to zero for a specified class of excitations. Thus, at a discrete set of isolated points the model does not necessarily converge (similar to the convergence of a Fourier series to a discontinuous function). Under mean square convergence relays, quantizers and other discontinuous nonlinear systems can be included in the model set. The reader should be aware that this set of systems is not complete; for example, bifurcations can still not be modeled within this concept. These ideas are very similar to the idea of Wiener series as explained by Schetzen (1980). Because the FRF measurements can be considered as the minimizers of a weighted least squares cost function, it is clear that the input-output relationship of the nonlinear distortions is approximated in least square sense. This motivates the following assumption:

Definition 3.5 (Class of Nonlinear Systems): $\mathbb{S}$ is the set of nonlinear systems such that for random multisines $u_N \in \mathbb{E}_N$ (see Definition 3.2) or periodic noise $u_N \in \mathbb{P}_N$ (see Definition 3.3)

$$\sum\nolimits_{\alpha=1}^{\infty} M_{G^\alpha} M_U^\alpha \le C_1 < \infty \tag{3-8}$$

with $M_{G^\alpha} = \max \left| G^\alpha_{L_k, k_1, k_2, \dots, k_{\alpha-1}} \right|$ and where M_U^α is defined in Definition 3.2 or Definition 3.3.

Under condition (3-8) there exists a uniformly bounded Volterra series whose output converges in mean square sense to the output of the nonlinear distortion for $u_N \in \mathbb{E}_N$. The FRF measurement $G(j\omega_k)$ at frequency f_k for nonlinear systems belonging to the set $\mathbb{S}$ excited with $u_N \in \mathbb{E}_N$ or $u_N \in \mathbb{P}_N$ is the sum of the nonlinear contributions of degree α, $G^\alpha(j\omega_k)$ (see Eq. (3-5)):

$$\begin{aligned} G(j\omega_k) &= \frac{Y_k}{U_k} = \sum\nolimits_{\alpha=1}^{\infty} G^\alpha(j\omega_k) \\ G^\alpha(j\omega_k) &= \frac{Y_k^\alpha}{U_k} = \sum_{k_1, \dots, k_{\alpha-1} = -N}^{N} G^\alpha_{L_k, k_1, k_2, \dots, k_{\alpha-1}} \frac{U_{k_1} U_{k_2} \dots U_{k_{\alpha-1}} U_{L_k}}{U_k} \end{aligned} \tag{3-9}$$

3.4 STUDY OF THE PROPERTIES OF FRF MEASUREMENTS IN THE PRESENCE OF NONLINEAR DISTORTIONS

In this section, profound insight is given into the impact of the nonlinear distortions on the FRF measurements for normalized random multisine excitations. It is shown in Appendix 3.A that the contributions to the FRF can be partitioned into two sets, the first one consisting of contributions that do not depend on the random phases of the excitation and the second one containing the contributions that depend on the random phases:

(i) Systematic contributions $G_B(j\omega_k)$: There exists a related linear dynamic system $G_R(j\omega_k)$ to which the expected value of the FRF estimate converges under weak conditions. It differs from the underlying linear system $G_0(j\omega_k)$ by the systematic contributions $G_B(j\omega_k)$ of the nonlinear distortions. We will show that for the class of normally distributed signals (including random multisines and noise excitations) the related linear dynamic system (RLDS) is the best linear approximation to the nonlinear system. The contributions of $G_R(j\omega_k)$ do not depend upon the random phases of the input.

(ii) Stochastic contributions $G_S(j\omega_k)$: Even for a very large number of frequencies and in the absence of disturbing noise, the FRF measurement is not smooth as a function of the frequency. It is scattered around its expected value, and these deviations do not converge to zero. They are called the stochastic nonlinear distortions. The contributions to $G_S(j\omega_k)$ depend on the random phases of the input.

These concepts are formalized below. For a system belonging to the set $\mathbb{S}$ and a normalized random multisine excitation $u_N \in \mathbb{E}_N$ (or normalized periodic noise $u_N \in \mathbb{P}_N$), the measured FRF consists of three parts:

$$G(j\omega_k) = G_R(j\omega_k) + G_S(j\omega_k) + N_G(k) \tag{3-10}$$

with $G_R(j\omega_k)$ the RLDS, $G_S(j\omega_k)$ the stochastic nonlinear contributions, and $N_G(k)$ the errors due to the output noise.

The related linear dynamic system $G_R(j\omega_k)$ consists of two parts:

$$G_R(j\omega_k) = G_0(j\omega_k) + G_B(j\omega_k) \tag{3-11}$$

with $G_0(j\omega_k)$ the underlying linear system and $G_B(j\omega_k)$ the bias or systematic errors due to the nonlinear distortions.

$G_S(j\omega_k)$ is called a stochastic contribution because it behaves as uncorrelated (over the frequencies) noise, although the reader should be aware that it is not a random signal once the excitation signal is fixed. Because of this noisy behavior, the presence of nonlinear distortions is often not recognized.

$N_G(k)$ describes the impact of the disturbing noise on the FRF measurement. For simplicity we assume that the input measurements are noise free (dominating output noise) resulting in a noise distortion $N_G(k)$ having the following properties:

Assumption 3.6 (Measurement Noise): The noise $N_G(k)$ on the FRF measurement has the following properties.

(i) $\mathcal{E}\{N_G(k)\} = 0$

(ii) $\mathcal{E}\{N_G(k)\overline{N}_G(l)\} = \sigma_G^2(k)\delta_{kl}$ and $\mathcal{E}\{|N_G(j\omega_k)|^2\} = \sigma_G^2(k)$

(iii) $\mathcal{E}\{N_G(l)|N_G(k)|^2\} = 0$ for $k, l \neq 0$

(iv) $\mathcal{E}\{(|N_G(k)|^2 - \sigma_G^2(k))(|N_G(l)|^2 - \sigma_G^2(l))\} = \begin{cases} 0 & k \neq l \\ O(N^0) & k = l \end{cases}$

The different contributions to the FRF are studied in more detail in the following for two situations. In the first case we look for the average value if the experiment is repeated for a constant number of harmonics in the excitation. The second case deals with the asymptotic behavior if the number of harmonics $N \to \infty$.

3.4.1 Study of the Expected Value of the FRF for a Constant Number of Harmonics

What happens if the FRF measurement is averaged for different realizations of a normalized random multisine excitation, keeping its amplitude spectrum constant? Or more formally: what is the expected value $\mathcal{E}\{G(j\omega_k)\}$ for N fixed? Thereto the mathematical expectation $\mathcal{E}\{G^{\alpha}(j\omega_k)\}$ is calculated with respect to the phases. This means that the measured frequency response function of the system is averaged over different realizations of the random multisine excitation, keeping the frequency grid and the amplitude of the Fourier coefficients U_k of the excitation signal $u(t)$ constant.

Theorem 3.7 (Response Nonlinear System): For a system belonging to the system set $\mathbb{S}$ (see Definition 3.5), excited with independent realizations of a normalized random multisine $u_N \in \mathbb{E}_N$ (see Definition 3.2) or normalized periodic noise $u_N \in \mathbb{P}_N$ (see Definition 3.3), we have:

1. The expected value of $G(j\omega_k)$ is given by

$$\mathcal{E}\{G(j\omega_k)\} = G_R(j\omega_k) = G_0(j\omega_k) + G_B(j\omega_k) \qquad \text{(3-12)}$$

with

$$G_B(j\omega_k) = \begin{cases} \sum_{\alpha=2}^{\infty} G_B^{2\alpha-1}(j\omega_k) & \text{for uniform continuous phase distributions} \\ \sum_{\alpha=2}^{\infty} G_B^{2\alpha-1}(j\omega_k) + O(N^{-1}) & \text{otherwise} \end{cases}$$

$$G_B^{2\alpha-1}(j\omega_k) = \mathcal{E}\{G^{2\alpha-1}(j\omega_k)\}.$$

2. The expected value of $G^{\alpha}(j\omega_k)$ is given by

$$\mathcal{E}\{G^{2\alpha-1}(j\omega_k)\} = c_{\alpha} \sum_{s_1, \ldots, s_{\alpha-1}=1}^{N} G^{2\alpha-1}_{k,-s_1,s_1,\ldots,-s_{\alpha-1},s_{\alpha-1}} \mathcal{E}_{\text{amp}}\{|U_{s_1}|^2 \ldots |U_{s_{\alpha-1}}|^2\} + O_{\alpha}(N^{-1}) \qquad \text{(3-13)}$$

$$\mathcal{E}\{G^{2\alpha}(j\omega_k)\} = \begin{cases} 0 & \text{for uniform continuous phase distributions} \\ O_{\alpha}(N^{-3/2}) & \text{otherwise} \end{cases}$$

with $c_{\alpha} = 2^{\alpha-1}(2\alpha-1)!!$, $\sum_{\alpha=2}^{\infty} O_{\alpha}(N^{-\beta}) = O(N^{-\beta})$, and where $\mathcal{E}_{\text{amp}}\{.\}$ denotes the expected value with respect to the random amplitudes of the periodic noise.

Proof. See Appendix 3.A. □

Remarks

(i) Note that from (3-13) it follows that $\mathcal{E}\{G^{2\alpha-1}(j\omega_k)\} = O(N^0)$ because in the sum $N^{\alpha-1}$ terms of $O(1/N^{\alpha-1})$ are added together ($|U_{s_1}| = O(N^{-1/2})$; see Definitions 3.2 and 3.3).

(ii) The related linear dynamic system depends on the number of frequencies N that are used in the random multisine (periodic noise). Therefore, it would be better to denote it as $G_{R,N}$. However, later on it will be shown that the limit for $N \to \infty$ exists: $\lim\limits_{N \to \infty} G_{R,N}(j\omega) = G_R(j\omega)$. For that reason we preferred not to overload the notation, leaving out the dependence on N.

(iii) Instead of $G_B^{\alpha}(j\omega_k)$ being considered as the expected value (see (3-12)), it can be interpreted as that part of the transfer function contribution of degree α that is independent of the random phase of the random multisine excitation. All the components that still depend on the random phase have a zero mean value because $\mathcal{E}\{e^{j\varphi}\} = 0$ and as such do not contribute to the bias term. Consequently, $G_B^{\alpha}(j\omega_k)$ is independent of the random phases of the excitation; in the contributing terms the random phases of the excitation cancel each other, resulting in a systematic contribution of the nonlinear distortion to the FRF. $G_S^{\alpha}(j\omega_k)$ depends on the random phases of the excitation so that it is a random component, modeling the stochastic contribution of the nonlinear distortion of degree α to the FRF.

(iv) A typical example of a discrete phase distribution is $\varphi \in \{0, \pi\}$. For discrete phase distributions, the even degree terms also have a bias contribution that disappears as an $O(N^{-1})$.

An important conclusion of this section is that only the odd terms $G^{2\alpha-1}(j\omega_k)$ contribute to the related linear dynamic system; it does (asymptotically) not depend on the even nonlinear distortions. This result will be used later on to formulate optimized measurement strategies. The theorem also gives a possibility to measure $G_R(j\omega_k)$. It can be obtained by averaging over a sufficient number of experiments with different realizations of the random multisine so that the stochastic nonlinear contributions are averaged to zero.

3.4.2 Asymptotic Behavior of the FRF if the Number of Harmonics Tends to Infinity

From the previous section we know that besides the disturbing noise $N_G(k)$, the measured FRF consists of two remaining components: a deterministic one $G_R(j\omega_k)$ and a stochastic one $G_S(j\omega_k)$. A first possibility to measure $G_R(j\omega_k)$ is to average over a large number of experiments so that the contribution $G_S(j\omega_k)$ is averaged to zero for a fixed number of frequency components N in the random multisine (periodic noise). Because in each realization we should calculate and load each time a new random multisine (periodic noise sequence) in the generator memory and wait until the transients in the measured signals disappear, it is tempting to stick to one experiment, but using a very dense ($N \to \infty$) multisine (periodic noise). One might hope that the resulting measurement of the FRF would become smooth because the stochastic nonlinear contributions would average to zero. It turns out that this is not the case. Neither of the contributions ($G_R(j\omega_k)$ and $G_S(j\omega_k)$) decreases if the number of frequencies N of the excitation increases; the FRF does not become smooth for $N \to \infty$. Also the bias contribution $G_B(j\omega_k)$ does not decrease when N increases because it is an $O(N^0)$. This is formalized in the next theorem.

Theorem 3.8 (Asymptotic Behavior of the Systematic and Stochastic Nonlinearities): Consider a system belonging to the system set $\mathbb{S}$, excited with a random multisine $u_N \in \mathbb{E}_N$ (see Definition 3.2) or periodic noise $u_N \in \mathbb{P}_N$ (see Definition 3.3). The systematic $G_B(j\omega_k)$ and stochastic $G_S(j\omega_k)$ contributions to the transfer function

$G(j\omega_k) = G_R(j\omega_k) + G_S(j\omega_k)$, with $G_R(j\omega_k) = G_0(j\omega_k) + G_B(j\omega_k)$, do not decrease to zero as $N \to \infty$: $G_B(j\omega_k)$ is an $O(N^0)$ and $G_S(j\omega_k)$ is an $O_{\text{m.s.}}(N^0)$.

Proof. See remarks in Section 3.4.1 on $G_B(j\omega_k)$ and Appendix 3.B on $G_S(j\omega_k)$. □

The stochastic behavior of $G_S(j\omega_k)$ can be further characterized, showing that its second-order properties are completely similar to those of the noise $N_G(k)$. This explains why it is difficult to distinguish between noise and nonlinear distortions. It is also the reason why nonlinear distortions are often not recognized.

Theorem 3.9 (Properties of Stochastic Nonlinearities): For a system belonging to the system set $\mathbb{S}$, excited with a random multisine $u_N \in \mathbb{E}_N$ (see Definition 3.2) or periodic noise $u_N \in \mathbb{P}_N$ (see Definition 3.3), the following properties are valid:

(i) $\mathcal{E}\{G_S(j\omega_k)\} = 0$

(ii) $\mathcal{E}\{G_S(j\omega_l)\bar{G}_S(j\omega_k)\} = O(N^{-1})$ if $k \neq l$ and $\mathcal{E}\{|G_S(j\omega_k)|^2\} \equiv \sigma^2_{G_S}(k) = O(N^0)$

(iii) $\mathcal{E}\{G_S(j\omega_l)|G_S(j\omega_k)|^2\} = O(N^{-1})$ for $k \neq l$

(iv) $\mathcal{E}\{(|G_S(j\omega_k)|^2 - \sigma^2_{G_S}(k))(|G_S(j\omega_l)|^2 - \sigma^2_{G_S}(l))\} = \begin{cases} O(N^{-1}) & k \neq l \\ O(N^0) & k = l \end{cases}$

Proof. See Appendix 3.B. □

Remark. These observations are in agreement with the classical result showing that the output of a nonlinear system can be split into two parts (Bendat, 1998; Forssell and Ljung, 2000b): a first part that is linearly related to the input (in our case leading to $G_R(j\omega_k)$) and a second part that is uncorrelated with the input (leading to $G_S(j\omega_k)$). Theorem 3.9 tells more about the second and higher order properties of the uncorrelated part.

In the previous theorem, the moments of the nonlinear contributions up to the fourth order were studied. In general, it is even possible to tell more about these nonlinear distortions. In the next theorem, it is shown that they are mixing (see also Section 14.4). Loosely speaking, this means that the dependence of the nonlinear contributions decreases fast enough to zero if the frequency distance between the contributions increases.

Theorem 3.10 (Mixing Property of Stochastic Nonlinearities): The nonlinear contributions for a system belonging to the system set $\mathbb{S}$, excited with a random multisine $u_N \in \mathbb{E}_N$ (see Definition 3.2) or periodic noise $u_N \in \mathbb{P}_N$ (see Definition 3.3), are mixing of order infinity.

Proof. See Appendix 3.C. □

Theorem 3.11 (Distribution of Stochastic Nonlinearities): For a system belonging to the system set $\mathbb{S}$, excited with a random multisine $u_N \in \mathbb{E}_N$ (see Definition 3.2) or periodic noise $u_N \in \mathbb{P}_N$ (see Definition 3.3), the stochastic nonlinearities are circular complex normally distributed.

Proof. See Appendix 3.E. □

3.4.3 Further Comments on the Related Linear Dynamic System

In this section a physical interpretation is given for the related linear dynamic system $G_R(j\omega)$. First, it will be shown that normally distributed random excitations and random multisines result in the same related linear dynamic system if both excitations are generated from the same power spectrum $\hat{U}^2(f)$ (see Definition 3.2 for the normalized random multisine). Next, it will be shown that $G_R(j\omega)$ corresponds to the best linear approximation, in least squares sense, of the nonlinear system; finally it will be shown that asymptotically $G_R(j\omega)$ is smooth.

3.4.3.1 Connecting the Random Multisine to Normally Distributed Noise. If the system is excited with Gaussian noise, the limit value of the estimated FRF (after averaging over an infinite number of blocks and neglecting leakage effects, see Chapter 2) is given by

$$\hat{G}(j\omega) = S_{YU}(j\omega)/S_{UU}(j\omega) \tag{3-14}$$

Splitting $Y(j\omega)$ into its contribution of degree α results in $Y(j\omega) = \sum_{\alpha=1}^{\infty} Y^{\alpha}(j\omega)$ and shows that the nonlinear contribution of degree α to $G_R(j\omega)$ should be calculated as: $\hat{G}^{\alpha}(j\omega) = S_{YU}^{\alpha}(j\omega)/S_{UU}(j\omega)$. To interpret $S_{YU}^{\alpha}(j\omega)$ higher order spectra can be used (Bendat and Piersol, 1980; Bendat, 1998; Billings, 1980; Brillinger, 1981; Mendel, 1991; Nikias and Mendel, 1993; Nikias and Petropulu, 1993). Because these higher order spectra depend not only on the power spectrum of the excitation noise but also on their higher order moments, it is clear that the value of $\hat{G}^{\alpha}(j\omega)$ also depends on its pdf. In the case of zero mean normally distributed noise, the higher order spectra can be calculated easily and the contribution of degree α is given by

$$\begin{aligned} \hat{G}^{2\alpha-1}(j\omega) &= c_{\alpha}\int_0^{\infty}\dots\int_0^{\infty} G^{2\alpha-1}_{f, f_1, -f_1, \dots, -f_{\alpha-1}} S_{UU}(j\omega_1)\dots S_{UU}(j\omega_{\alpha-1})df_1\dots df_{\alpha-1} \\ \hat{G}^{2\alpha}(j\omega) &= 0 \end{aligned} \tag{3-15}$$

with $c_{\alpha} = 2^{\alpha-1}(2\alpha-1)!!$ (see Appendix 3.G). This result allows us a better understanding of the RLDS as it was obtained for the random multisine: Eq. (3-15) is similar to (3-13). Integrals have to be considered over the continuous power spectrum of the noise instead of sums over discrete spectral components of the periodic signal. In Section 3.4.3.3 a formal statement is given on the asymptotic ($N \to \infty$) equivalence of $G_R(j\omega)$ for the three considered classes of excitations: random multisines, periodic noise, and normally distributed noise.

3.4.3.2 Interpretation of the Related Linear Dynamic System as the Best Linear Approximation. When a nonlinear system is approximated using a linear system, it is important to be sure that the best approximation is made. This is actually the case for $G_R(j\omega)$. This follows directly from the fact that Eq. (3-14) is shown to give the best linear approximation in least square sense (Eykhoff, 1974; Bendat and Piersol, 1980). The estimated impulse response (and the corresponding FRF) minimizes the mean square value of $e(t) = y(t) - g(t)*u(t)$ over the measurement interval. For periodic excitations, (3-14) boils down to $G(j\omega_k) = S_{YU}(j\omega_k)/S_{UU}(j\omega_k) = Y_k/U_k$, which is exactly the starting expression used in (3-9). So the related linear dynamic system is also the best linear approximation for the class of random multisine excitations. The reader should be aware that this approximation is a function of the power spectrum of the excitation.

3.4.3.3 Asymptotic Equivalences. The following theorem states that the asymptotic best linear approximation $G_R(j\omega)$ is the same for random phase multisines (Definition 3.2), periodic noise (Definition 3.3), and Gaussian noise with the same (power) spectra. Hence, $G_R(j\omega)$ can be used to predict the response of the nonlinear system to any signal belonging to these three classes. Note, however, that the prediction error is bounded below by the stochastic nonlinear contributions $y_s(t) = \lim_{N \to \infty} y_{s,N}(t)$ (the notation $y_{s,N}$ is used here to indicate explicitly the dependence on the number of components). If this error is too large for a particular application, then the only way to improve the prediction quality is to model also the nonlinear behavior of the system.

The advantage of using random phase multisines over periodic noise to measure $G_R(j\omega)$ is that additional averages over the random amplitudes are avoided. The advantage of using periodic noise over Gaussian noise to measure $G_R(j\omega)$ is that the leakage errors are avoided.

Assuming that FRF measurements with M different excitations signals are made, the asymptotic best linear approximation $G_R(j\omega_k)$ can be estimated as

$$\hat{G}_R(j\omega_k) = \sum_{m=1}^{M} Y^{[m]}(k)/U^{[m]}(k) \tag{3-16}$$

for random phase multisines (see Eq. 3-14) and as

$$\hat{G}_R(j\omega_k) = \frac{\sum_{m=1}^{M} Y^{[m]}(k)\bar{U}^{[m]}(k)}{\sum_{m=1}^{M} |U^{[m]}(k)|^2} \tag{3-17}$$

for periodic and Gaussian noise, where $U^{[m]}(k)$ and $Y^{[m]}(k)$ are the input and output DFT spectra of the mth FRF measurement.

Theorem 3.12 (Asymptotic Best Linear Approximation): Consider the following three classes of excitation signals: (i) random phase multisines (see Definition 3.2) with $\hat{U}^2(f) \equiv S_{\hat{U}\hat{U}}(f)$, (ii) periodic noise (see Definition 3.3) with $E\{\hat{U}^2(f)\} \equiv S_{\hat{U}\hat{U}}(f)$, and (iii) Gaussian noise with power spectrum $S_{UU}(j\omega) \equiv S_{\hat{U}\hat{U}}(f)/f_{\max}$ for $|f| < f_{\max}$ and zero elsewhere. For these three classes of excitation signals, the best linear approximations $G_{R,N}(j\omega)$ (H_1-FRF measurement) of a nonlinear system belonging to the class $\mathbb{S}$ (see Definition 3.5) converge (measurement time and $N \to \infty$) at the rate $O(N^{-1})$ to the same limit value $G_R(j\omega)$. If the joint second-order derivatives of the odd degree kernels $G^{2\alpha-1}_{f,-f_1,f_1,\ldots,-f_{\alpha-1},f_{\alpha-1}}$, $\alpha = 1, 2, \ldots, \infty$, w.r.t. $f, f_1, \ldots, f_{\alpha-1}$, are bounded for $f, f_1, \ldots, f_{\alpha-1} \in [0, f_{\max}]$, then $G_R(j\omega)$ is given by

$$\begin{aligned} G_R(j\omega) &= G_0(j\omega) + G_B(j\omega) = G_0(j\omega) + \sum_{\alpha=2}^{\infty} C_1^{\alpha}(j\omega) \\ C_1^{\alpha} &= \frac{c_\alpha}{f_{\max}^{\alpha-1}} \int_0^{f_{\max}} \ldots \int_0^{f_{\max}} G^{2\alpha-1}_{f,f_1,-f_1,\ldots,-f_{\alpha-1}} S_{\hat{U}\hat{U}}(f_1) \ldots S_{\hat{U}\hat{U}}(f_{\alpha-1}) df_1 \ldots df_{\alpha-1} \end{aligned} \tag{3-18}$$

with $c_\alpha = 2^{\alpha-1}(2\alpha-1)!!$.

Proof. See Appendix 3.H. □

From Theorem 3.12 it follows that the asymptotic best linear approximation $G_R(j\omega)$ depends only on the second-order moments $S_{\hat{U}\hat{U}}(f)$ of the input spectrum. Note also that (3-15), with $S_{UU}(j\omega) = S_{\hat{U}\hat{U}}(f)/f_{\max}$ for $|f| < f_{\max}$ and zero elsewhere, reduces to C_1^{α} in (3-18).

3.4.3.4 Smoothness. Additional assumptions are required to guarantee the smoothness of $G_R(j\omega)$. This restricts further the class of allowable nonlinear systems.

Assumption 3.13: For any $\omega \in [0, \omega_{\max}]$, the odd degree Volterra kernels $G^{2\alpha-1}_{f, f_1, -f_1, \ldots, -f_{\alpha-1}}$, $\alpha = 1, 2, \ldots$, are continuous functions of ω with continuous Pth order derivative w.r.t. ω.

For example, systems consisting of the cascade and parallel connection of linear systems and multipliers result in rational Volterra kernels for which Assumption 3.13 is satisfied (Schetzen, 1980).

Assumption 3.14: The series $\sum_{\alpha=2}^{Q} C_1^{\alpha}(j\omega)$, with C_1^{α} defined in (3-18), and its derivatives of order $1, 2, \ldots, P$ w.r.t. ω converge ($Q \to \infty$) uniformly in $\omega \in [0, \omega_{\max}]$ to their limit sum.

Note that Assumptions 3.13 and 3.14 do not exclude the possible nonuniform (point wise or mean square) convergence of the output of the Volterra series model $y_Q(t)$ to $y(t)$.

Theorem 3.15 (Smoothness Best Linear Approximation): Under the conditions of Theorem 3.12 and Assumptions 3.13 and 3.14, the asymptotic best linear approximation $G_R(j\omega)$ is a continuous function of $\omega \in [0, \omega_{\max}]$ with continuous Pth order derivative.

Proof. See Appendix 3.I. □

From this theorem it follows that $G_R(j\omega)$ and its higher order derivatives w.r.t. ω are continuous functions of ω. This explains why $G_R(s)$ can be approximated very well by a rational function of s of sufficiently high order.

3.4.3.5 Special Case: Wiener-Hammerstein Systems. In the case of a Wiener-Hammerstein system, consisting of a linear system with transfer function $R(j\omega)$, followed by a static nonlinearity, $v(t) = \sum_{k=0}^{\infty} a_k u^k(t)$, with $a_k \in \mathbb{R}$, and a second linear system $S(j\omega)$ (see Figure 3-3), the previous expressions can be simplified further. The Volterra kernel of degree α at frequency k is (Schetzen, 1980)

$$\begin{aligned} G^{2\alpha-1}_{k_1, \ldots, k_{2\alpha-1}} &= a_{2\alpha-1} S(j\omega_k) R(j\omega_{k_1}) \ldots R(j\omega_{k_{2\alpha-1}}) \\ G^{2\alpha}_{k_1, \ldots, k_{2\alpha}} &= 0 \end{aligned} \qquad (3\text{-}19)$$

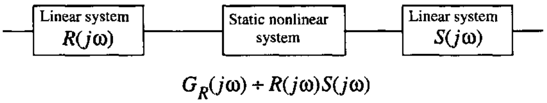


Figure 3-3. Nonlinear Wiener-Hammerstein system and its related linear dynamic.

with $\omega_k = \sum_{i=1}^{\beta} \omega_{k_i}$, $\beta = 2\alpha - 1$ or $\beta = 2\alpha$, and $a_{2\alpha-1}$ a constant independent of the frequencies and the input signal. Using Theorems 3.7 and 3.12, we find

$$\begin{aligned}
G_0(j\omega_k) &= a_1 R(j\omega_k) S(j\omega_k) \\
G_B^{2\alpha-1}(j\omega_k) &= a_{2\alpha-1}\, D_\alpha R(j\omega_k) S(j\omega_k) + O_\alpha(N^{-1}) \\
D_\alpha &= \frac{c_\alpha}{f_{\max}^{\alpha-1}} \int_0^{f_{\max}} \dots \int_0^{f_{\max}} |R(f_1)|^2 \dots |R(f_{\alpha-1})|^2 S_{\hat{U}\hat{U}}(f_1) \dots S_{\hat{U}\hat{U}}(f_{\alpha-1}) df_1 \dots df_{\alpha-1} \\
G_B^{2\alpha}(j\omega_k) &= 0
\end{aligned} \tag{3-20}$$

with $c_\alpha = 2^{\alpha-1}(2\alpha-1)!!$, $R(f) = R(j\omega)$, and $\sum_{\alpha=1}^{\infty} O_\alpha(N^{-1}) = O(N^{-1})$. Hence, the asymptotic ($N \to \infty$) related linear dynamic system is given by

$$G_R(j\omega) = C(U, R) R(j\omega) S(j\omega) \tag{3-21}$$

with $C(U, R) = \sum_{\alpha=1}^{\infty} a_{2\alpha-1} D_\alpha$. As a result, for Wiener-Hammerstein systems, the asymptotic best linear approximation $G_R(j\omega)$ equals the underlying linear system within a real frequency-independent scale factor $C(U, R)$ that depends on the excitation signal and the system $R(j\omega)$. Similar results were also reported (Billings and Fakhour, 1982; Nikias and Petropulu, 1993) for special classes of excitation signals such as white zero mean Gaussian noise.

Remark. Sometimes the structure in Figure 3-3 is called the "general model" (e.g., Billings and Fakhour, 1982).

3.4.4 Further Comments on the Stochastic Nonlinear Contributions

Also for the stochastic nonlinear contributions, the smoothness and the equivalence results can be obtained. Using (3-9) and (3-10) with $N_G(k) = 0$ (no disturbing noise), the relation between the input and output Fourier coefficients at frequency $f_k = kf_{\max}/N$ can be written as

$$Y_k = G_R(j\omega_k) U_k + Y_{Sk} \tag{3-22}$$

with $Y_{Sk} = G_S(j\omega_k) U_k$ the stochastic nonlinear contributions observed at the output of the system. Because $U_k = O(N^{-1/2})$ (see Definitions 3.2 and 3.3) and $|G_S(j\omega_k)| = O(N^0)$ (see Theorem 3.9) it follows that $Y_{Sk} = O(N^{-1/2})$. The following theorem studies the asymptotic ($N \to \infty$) behavior of the variance of $\sqrt{N} Y_{Sk}$ for random phase multisines, periodic noise, and Gaussian noise excitations.

Theorem 3.16 (Asymptotic Variance of Stochastic Nonlinear Contributions): Consider the following three classes of excitation signals: (i) random phase multisines (see Definition 3.2) with $\hat{U}^2(f) \equiv S_{\hat{U}\hat{U}}(f)$, (ii) periodic noise (see Definition 3.3) with $E\{\hat{U}^2(f)\} \equiv S_{\hat{U}\hat{U}}(f)$, and (iii) Gaussian noise with power spectrum $S_{UU}(j\omega) \equiv S_{\hat{U}\hat{U}}(f)/f_{\max}$ for $|f| < f_{\max}$ and zero elsewhere. For these three classes of excitation signals, the variances $\text{var}(\sqrt{N} Y_{Sk,N})$ of the stochastic nonlinear distortions

$\sqrt{N}Y_{Sk,N}$ of a nonlinear system belonging to the class $\mathbb{S}$ (see Definition 3.5) converge (measurement time and $N \to \infty$) at the rate $O(N^{-1})$ to the same limit value $\sigma_{\mathbb{S}}^2(f)$.

Note: We denoted explicitly the dependence of the results on the number of frequencies N by adding a subscript N.

Proof. See Appendix 3.J. □

The asymptotic variance ($N, M \to \infty$) of the FRF estimate $\hat{G}_R(j\omega_k)$ (3-16) and (3-17) due to the stochastic nonlinear distortions is given by

$$\frac{\sigma_{\mathbb{S}}^2(f)}{MS_{\hat{U}\hat{U}}(f)} \text{ with } \sigma_{\mathbb{S}}^2(f) = \lim_{N \to \infty} \mathrm{var}(\sqrt{N}Y_{Sk,N}) \tag{3-23}$$

(see Eqs. (2-25) and (2-52) with $\sigma_{\hat{U}}^2 = 0$, $\sigma_{\hat{Y}U}^2 = 0$, $|Y_0|^2/|G_0|^2 = S_{\hat{U}\hat{U}}$ and $S_{Y_0Y_0}/|G_0|^2 = S_{\hat{U}\hat{U}}$). From Theorem 3.16 it follows that the variance (3-23) of the FRF measurement (3-16) and (3-17) depends only on the second-order moments $S_{\hat{U}\hat{U}}(f)$ of the input spectrum. Hence, it is the same for random phase multisines, periodic noise, and Gaussian noise excitations.

It can also easily be shown that $\sigma_{\mathbb{S}}^2(f)$ in Theorem 3.16 is a smooth function of the frequency f (continuous and continuous high order derivatives). This motivates why $Y_{Sk,N}$ ($\mathrm{var}(Y_{Sk,N})$) can be modeled very well as a discrete-time, filtered white noise sequence $H(z_k^{-1})E(k)$ ($\sigma^2|H(z_k^{-1})|^2$), where $H(z^{-1})$ is a rational function in z^{-1}.

3.4.5 Extension to Discrete-Time Modeling

The results of Sections 3.4.1 to 3.4.4 were obtained for continuous-time systems. In this section we will show that these can be extended to discrete-time models. Some precautions should be taken because for the discrete-time domain, the frequency axis is finite: $\omega \in [-\pi, \pi)$. In the nonlinear operations, higher frequencies can be created (e.g., $k\omega$), but these are folded back to the previous interval by the modulo operation: $\omega_{\text{folded}} = [(\omega + \pi) \bmod 2\pi] - \pi$, so that new frequency combinations appear that were not present in the previous sections. We show subsequently that the folding operation does not change the nature of these components (systematic or stochastic contributions). To do so, we consider the unfolded frequency $\underset{\sim}{\omega}$, as it results from the frequency combinations in the nonlinear system. In the next theorem we show that for a nonlinear system, excited by a band-limited random multisine excitation ($U(j\omega) = 0$ for $|\omega| > \omega_{\max}$), its output components at frequencies $|\underset{\sim}{\omega}| > \omega_{\max}$ can only be stochastic contributions. This means that they cannot be combined with any component of the random multisine excitation to result in a phase-independent combination.

Remark. To formalize this result in a theorem, we have to consider discrete-time random multisines. These are obtained directly from Definition 3.2 by replacing t by the discrete-time variable k, with $k = 0, 2, \ldots, N-1$. The frequencies of a discrete-time random multisine are restricted to the grid $2\pi/N$ in order to get periodic discrete-time signals (see Oppenheim et al., 1997: not all frequencies result in a periodic signal in the discrete-time domain!).

Theorem 3.17 (Stochastic Behavior of the Out-of-Band Components): For a (discrete-time) system belonging to the system set $\mathbb{S}$ (see Definition 3.5), excited with independent realizations of a (discrete-time) normalized random multisine $u_N \in \mathbb{E}_N$ (see

Definition 3.2 and the previous note) or (discrete-time) normalized periodic noise $u_N \in \mathbb{P}_N$ (see Definition 3.3 and the previous note), with maximum angular frequency $\omega_{\max} = l_{\max}\omega_1$ $(\omega_1 = 2\pi f_1)$, we have for $\underset{\sim}{\omega}_L = L\omega_1$, $|L| > l_{\max}$:

$$\mathcal{E}\{ Y_L^{\alpha} e^{-j\angle U_l}\} = 0 \tag{3-24}$$

Proof. See Appendix 3.K. □

Note that this theorem is valid for continuous-time and discrete-time systems (using $\underset{\sim}{\omega}$). A direct result of this theorem is that all results of the previous sections can also be applied to discrete-time systems. Because none of the "out-of-band" components can create systematic contributions, the folding process does not change the nature of the output contributions of a nonlinear system, and, hence, the previous proofs remain valid.

3.4.6 Experimental Illustration

A nonlinear mechanical resonating system (mass, viscous damping, nonlinear spring) is simulated with an electrical circuit. The displacement $y(t)$ (output) is related to the force $u(t)$ (input) by the following nonlinear, second-order differential equation:

$$m\frac{d^2y(t)}{dt^2} + d\frac{dy(t)}{dt} + k(y(t))y(t) = u(t) \tag{3-25}$$

The nonlinear spring is described by a static but position-dependent stiffness

$$k(y) = a + by^2 \tag{3-26}$$

For small excitations, the spring becomes almost linear so that the underlying linear system consists of a second-order resonance system. A series of experimental results on this system are shown. First, the nonlinear behavior will be illustrated using stepped sine measurements. Next, the split of the transfer function into the underlying linear system $G_0(j\omega_k)$, the related linear dynamic system $G_R(j\omega_k)$, the stochastic nonlinear distortions $G_S(j\omega_k)$, and the noise contributions $N_G(k)$ are shown.

3.4.6.1 Visualization of the Nonlinearity Using Stepped Sine Measurements. To visualize the nonlinear behavior of the system, a stepped sine measurement is made (Figure 3-4). The frequency of the sine is first stepped upward until the maximum frequency is reached and then stepped down again. At each frequency a measurement is made over an integer number of periods. During the experiment we took care to have a continuous excitation signal; no discontinuities appeared at the frequency-changing instants. The nonlinear behavior of the system is clearly visible. The measured transfer function depends, strongly, on the amplitude of the sine excitation. Moreover, the measurements also show that the actual output of the system depends on the past inputs: the up-path differs from the down-path for large excitations. Such behavior cannot be described using Volterra-based descriptions. Nevertheless, we will still apply the previously developed theory to this system. This can be done because the bifurcation appears only for large excitations, injecting a lot of power close to the resonance frequency of the system. If we use normalized random multisines, only a fraction of the power is injected in this band so that the bifurcation problem does not disturb the measurements anymore.

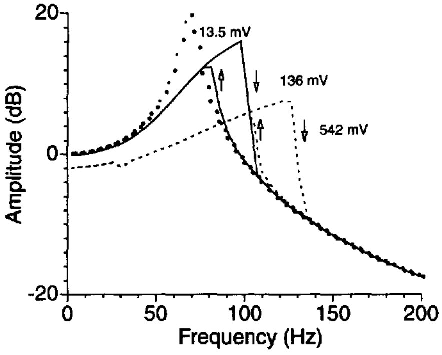


Figure 3-4. Stepped sine measurement at different amplitudes (rms values given). An up and down sweep is made. For the 13.5 mV measurement: black boxes up sweep, white boxes down sweep. For the others: ↓ up sweep, ↑ down sweep.

3.4.6.2 Measurement of the Related Linear Dynamic System. In a second step, the underlying linear system is measured using a normalized random multisine $(f_k = (2k+1)f_0$, $k = 0, 1, \ldots, 1340$ and $f_0 \approx 0.0745$ Hz$)$ with a small amplitude (rms value of 34.2 mV). The standard deviation $\sigma_{N_G}(k)$ is calculated from 10 consecutive periods. The results are shown in Figure 3-5.

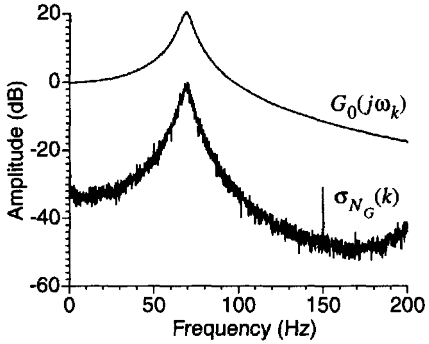


Figure 3-5. Measurement of the underlying linear system $G_0(j\omega_k)$ and its standard deviation.

The impact of the nonlinearity is made visible by increasing the excitation level of the normalized random multisine to an rms value of 127 mV. The measurement was repeated for 10 different realizations of the excitation signal so that $\sigma_{G_S}(k)$ could also be measured. The measurement results are shown in Figure 3-6. On the left side, the related linear dynamic system is compared with the underlying linear system. A number of observations can be made: the resonance frequency is shifted to the right, the peak value is decreased, and the measurement became more noisy.

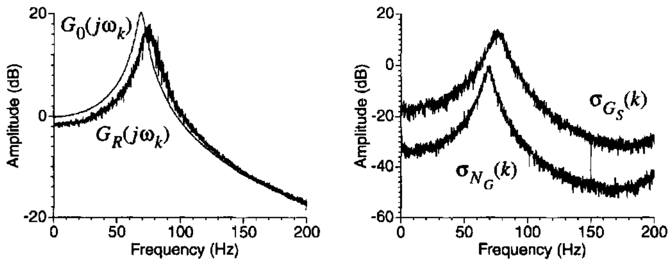


Figure 3-6. Comparison of the measured related linear dynamic system $G_R(j\omega_k)$ obtained from 10 realizations and the underlying linear system $G_0(j\omega_k)$.

The shift to the right of the resonance frequency is due to the nonlinear behavior of the hardening spring. For larger excursions, the average stiffness increases and so also does the resonance frequency. Note that if the $G_0(j\omega_k)$ measurement were not available, there would be no indication at all that this system is strongly nonlinear. This shows, clearly, why we need dedicated tools to detect the presence of nonlinear distortions. The difference between $G_R(j\omega_k)$ and $G_0(j\omega_k)$ is due to the systematic contributions $G_B(j\omega_k)$.

The increased noise level can be understood only from the previous, explained theory; they are due to the stochastic contributions $G_S(j\omega_k)$. Changing the excitation level did not change the disturbing noise, but $G_S(j\omega_k)$ became much larger. This is visualized on the left side of the figure. The standard deviation $\sigma_{G_S}(k)$ is obtained by measuring the FRF from 10 realizations of the normalized random multisine. For the small excitation level, it is completely dominated by the measurement noise $\sigma_{N_G}(k)$, whereas for the large excitation, $\sigma_{G_S}(k)$ dominates. This is also illustrated in Figure 3-7, where the evolution of the measured FRF is shown as a function of the excitation level. As can be seen, the stochastic contributions grow with the level while the measurement conditions (and, hence, the disturbing noise) remain the same. Again, it is very difficult to understand this result without the previously gained insight into the behavior of nonlinear systems. This also suggests a first test to detect the presence of nonlinear distortions. The standard deviation calculated from a set of consecutive periods (without changing the excitation signal) should be the same as that calculated from repeated measurements, using different realizations of the excitation signal.

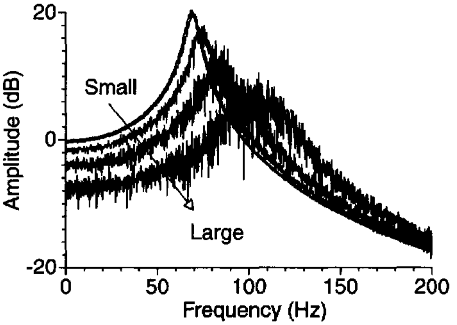


Figure 3-7. Evolution of the related linear dynamic system for growing excitation levels: rms values of 34 mV, 54 mV, 127 mV, 253 mV, and 507 mV.

3.5 DETECTION OF NONLINEAR DISTORTIONS

The ideal FRF-measurement method should provide the measured FRF, and at the same time the presence of nonlinear distortions should be detected, qualified (even or odd distortions), and quantified (the level of the distortions). Because the prime interest in these measurements is the FRF, it is unacceptable that most of the time would be spent on the detection of the nonlinear distortion at the cost of a reduced quality of the FRF measurement. This excludes most existing methods that require a series of dedicated measurements to make the nonlinearity test. In general, it is impossible to realize this ideal; however, when specially selected periodic excitations are applied, we can come close to it. This will be shown in the next section. Finally, in Section 3.5.3 some background information on the classical detection methods is given.

3.5.1 Detection of Nonlinear Distortions Using Periodic Excitations

The sine test is the simplest test characterizing, directly, the nonlinear behavior by verifying the generation of higher harmonics. However, this approach has a number of serious drawbacks. It is not only very slow (see Chapter 2), but as shown in the example of Section

3.4.6 it also does not measure the best linear approximation, except for very small excitations. This is due to the fact that it is not a random multisine excitation. This leads to the first conclusion that the excitation signals are restricted to broadband random multisines. The possibility to detect nonlinear distortions with these signals will be embedded by a careful selection of their amplitude spectrum, only a selected set of harmonics will be excited. This idea has already been suggested by Evans et al. (1994) and McCormack et al. (1994b). The odd-odd multisines that excite the system at the frequencies $(4k+1)f_0$, $k = 0, 1, \ldots, F$, are such a possibility. The linear system generates only an output at the excitation lines while the nonlinear distortions also hit the nonexcited harmonics. This allows their detection and characterization. From Section 3.2 it follows that:

At lines $4k+1$: the outputs consist of the linear contribution + odd nonlinear distortions.

At lines $4k+2$: only the even nonlinear distortions appear.

At lines $4k+3$: only odd nonlinear distortions appear.

So it is possible to detect and separate the even and the odd nonlinearities. The level of the distortions is indicated by the level at the detection lines. This can be extrapolated, with some care, to the measurement lines, although significant differences can still occur, especially when the low harmonics are filtered before arriving at the nonlinearity (e.g., a high-pass or bandpass input behavior of the system). For this reason, the results should be used as an indication and not as an absolute measure. In practice, it gives an underestimate because the level at the detection lines is usually below that at the measurement lines.

The test can be made more robust against these problems by using a modified multisine with components at kf_0, $k = 1, 3, 9, 11, 17, 19, \ldots$ (Vanhoenacker and Schoukens, 1999). In this case the even nonlinearities are detected at the even lines and the odd nonlinearities at the nonexcited, odd lines.

In many applications, the nonlinear distortions are of the same magnitude as the noise distortions. Consequently, it is necessary to separate them from the noise. This is again possible by exploiting the periodic nature of the signals. For each realization of the excitation signal, M periods are measured. Then a first, elegant method to distinguish between noise and distortions is to measure the "harmonic" coherence (McCormack et al., 1994b) at the nonexcited frequencies:

$$\gamma_H^2(\omega_k) = \left|\sum_{l=1}^{M} Y^{[l]}(k)\right|^2 \Big/ \sum_{l=1}^{M} |Y^{[l]}(k)|^2 \qquad (3\text{-}27)$$

This measure converges for $M \to \infty$ to $|Y_0(k)|^2/(|Y_0(k)|^2 + \sigma_Y^2(k))$, where $Y_0(k)$ is the DFT spectrum at a nonexcited frequency. If the nonlinear distortions are large compared with the noise, $|Y_0(k)| \gg \sigma_Y(k)$, it will be close to 1, while a small value indicates that the nonlinear contributions are below the noise level $|Y_0(k)| \ll \sigma_Y(k)$.

A second possibility is to calculate the sample variance over each block of M periods (for a single realization of the excitation) and to compare, directly, the measured distortion levels with the noise levels. The advantage of this approach is that a full characterization of the second-order moments of the noise is available at the end of the measurement.

In practice, some additional problems can occur during this test. The nonlinear interaction between generator and plant can also generate unwanted excitations at the detection frequencies, and it is no longer clear what part of the output should be assigned to the linear behavior and what part is due to the nonlinear distortions. In that case a first-order correction can compensate the output: $\tilde{Y}(k) = Y(k) - \tilde{G}(j\omega_k)U(k)$. The FRF estimate $\tilde{G}(j\omega_k)$ is obtained by linear interpolation of the FRF measurements at the excited frequencies (Vanhoenacker and Schoukens, 1999).

Conclusion. At the end of this simple experiment, the user gets a broadband measurement of the FRF, a detection, qualification, and rough quantification of the nonlinear distortions together with a noise analysis. The price to be paid is the loss in resolution, caused by the nonexcited lines.

3.5.2 Illustration on the Electrical Simulator

The experimental test setup of Section 3.4.6 is used again. This time the system is excited with an odd-odd random multisine, exciting the system at $(4k-1)f_0$, $k = 1, 2, \ldots, 128$ and $f_0 \approx 0.596$ Hz, with an rms value of 62.7 mV. In a first step, the standard deviation of the disturbing noise is extracted from a single input realization, measured during 10 periods. Next, 15 realizations of the random excitation are generated and each time the output is measured during one period, after waiting until the transients are negligible. The mean square average of the amplitude spectrum over these 15 realizations is shown in Figure 3-8. These results show that in one experiment it is possible to measure the FRF, the noise level, and the nonlinear distortions. In this case it is clear that the latter are the dominating error mechanism acting on the setup; the odd nonlinear distortions are 20 dB larger than the noise. This is very valuable information for the rest of the modeling process.

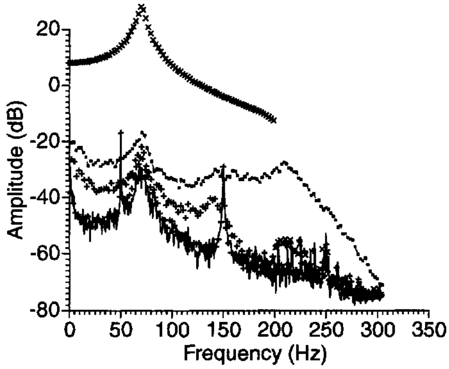


Figure 3-8. Detection of nonlinear distortions at the output of the nonlinear circuit using an odd-odd multisine. x: linear + odd nonlinear contributions; +: even nonlinear contributions; • : odd nonlinear contributions, __ σ_Y.

Remark. In practice it is not necessary to consider different realizations of the excitation. One experiment would do. However, later on in the experiment, we also wanted to measure the variance of the stochastic nonlinear contributions (see Figure 3-10) requiring more than one realization.

3.5.3 A Short Overview of Other Methods to Detect Nonlinear Distortions

The literature describes a series of other methods, different from that presented before. Here, we will touch only on a few of them; an extended list of references is available in Natke et al. (1988). Also Haber (1985) gives a brief review of nonlinearity tests. The simplest method is to scale the input $u(t) \to \alpha u(t)$ and verify if the output also scales with α, after taking care for the offsets. In practice, this method is less appealing. Two separate measurements are needed, and in many applications it is not simple to impose a scaled input due to the nonlinear load of the generator with the input impedance of the tested system. This problem is not disposed only in the special case where a discrete-time model is built between a signal in a computer memory and the output of the physical system (see Chapter 10). In this special situation the user has full control over the excitation signal. Moreover, the small nonlinearities have to be detected by taking the difference between two large, measured signals, making the method extremely sensitive to all possible measurement errors due to this indirect nature. Another popular test is to check the coherence. As pointed out before, this method does not allow separation of noise disturbances from nonlinearity problems and it fails completely for periodic excitations. Extending the test to higher order spectra by probing directly for higher order correlations that are typical for nonlinear systems may eliminate these drawbacks, but these methods are very time consuming, especially for random excitations. Also, Hilbert transform tests have been proposed (Tomlinson, 1987). Actually, these methods do not, directly, detect the nonlinear behavior itself. The authors check for a noncausality in the impulse response of the linear approximation (FRF) that might be induced by the nonlinear behavior, although there is no guarantee at all that there is a one-to-one relation between both effects. The method imposes significant constraints (e.g., only working on lowly damped systems) and a series of correction terms should be added because an FRF measurement can be made only in a restricted frequency band. For these reasons, we do not discuss these methods in detail and refer the reader to the available literature.

3.6 MINIMIZING THE IMPACT OF NONLINEAR DISTORTIONS ON FRF MEASUREMENTS

For clarity of the presentation, we first give a set of general guidelines so that the reader may keep a maximum overview over the problem. Next, we will go into a more detailed discussion and motivation of these guidelines, some of which are also illustrated in experiments or simulations.

3.6.1 Guidelines

In the previous sections it is shown that (see (3-10) and (3-11))

$$\begin{aligned} G(j\omega_k) &= G_R(j\omega_k) + G_S(j\omega_k) + N_G(k) \\ G_R(j\omega_k) &= G_0(j\omega_k) + G_B(j\omega_k) \end{aligned} \tag{3-28}$$

Depending upon the goal of the measurement, it is possible to select dedicated excitations. The signals needed to measure $G_0(j\omega_k)$, as well as possible, are different from those that should be used in order to get the best measurement of $G_R(j\omega_k)$. In each case, it is necessary to minimize the impact of the distortions $G_S(j\omega_k)$ and $N_G(k)$. In this section both problems are studied. To clarify the presentation, we give, first, the general overview and, next, a more detailed discussion and motivation.

3.6.1.1 Goal: Measurement of the True Underlying Linear System

- *First choice: odd-odd random multisine with the amplitude kept as small as possible*
  Advantage: This facilitates measurement of the FRF, together with its standard deviation $\sigma_{N_G}(k)$ due to the disturbing noise. Also the presence of nonlinear distortions is detected, qualified, and quantified. The impact of the nonlinear distortion on the uncertainty $\sigma_{G_S}(k)$ is minimized.
  Disadvantage: A loss in frequency resolution with a factor 4.

- *Second choice: odd random multisine with minimized crest factor*
  Advantage: This facilitates measurement of the FRF with its standard deviation $\sigma_{N_G}(k)$ due to the disturbing noise. The impact of the nonlinear distortion on the uncertainty $\sigma_{G_S}(k)$ is minimized (the same quality as in the first choice and the loss in frequency resolution is reduced to a factor 2).
  Disadvantage: It is no longer possible to detect the presence of odd nonlinearities.
- *Third choice: binary excitation, preferably with an odd spectrum*
  Advantage: The impact of the distortions is minimized.
  Disadvantage: Almost impossible to detect the presence of nonlinear distortions.

3.6.1.2 Goal: Measurement of the Best Linear Approximation. Advice: Use test signals with the same power spectrum and the same amplitude distribution as those that will be applied later on to the system.

- *First choice: use different realizations of an odd-odd (or odd) random multisine and average the FRF over these experiments.*
  Besides the advantages and disadvantages discussed under point 3.6.1.1, the major advantage is that the stochastic contributions $G_S(j\omega_k)$ are reduced in the averaging process. The major disadvantage is the increased measurement time because of the need for different realizations.
- *Second choice: use one realization of a very dense odd-odd (or odd) random multisine.*
  Advantage: Only one experiment is needed. It is still possible to smooth the FRF over small frequency bands.

3.6.2 Discussions and Illustrations

In this section the previous guidelines are commented on, motivated, and illustrated.

3.6.2.1 Goal: Measurement of the True Underlying Linear System. In this case, we try to measure the underlying linear system in such a way that all nonlinear influences should be minimized. From (3-2), it is clearly seen that the nonlinear contributions grow with the higher order moments. Therefore, the amplitude should be as small as possible. Minimizing the crest factor still maximizes the SNR of the measurements. Using an odd or odd-odd multisine, the even stochastic disturbances ($G_S(j\omega_k)$) are completely eliminated. This results in the first and second advice. It is still possible to reduce the nonlinear impact by choosing signals with an amplitude distribution (≠ power spectrum) that reduces the higher order moments in Eq. (3-2) for a fixed second-order moment (= total power in the signal). Binary distributions have the lowest ratio $\mathcal{E}\{(u/\sigma_u)^{2k}\} = 1$. This is the basis for the third advice. However, the reader should be aware that with a binary excitation it becomes very difficult to get any information about the nonlinear distortions, making it difficult to detect their presence. For the extreme situation of a static nonlinear system it becomes impossible to detect the nonlinearity. For this reason, it is strongly advised to select this solution only after making a preliminary nonlinearity test.

Example 3.18 (Static Nonlinear System): In order to visualize the impact of the crest factor and the power spectrum (consecutive, odd, and odd-odd multisines) on the nonlinear distortion, a simulation was made. The FRF of a static, nonlinear system $y = u + u^2/2.8 + u^3/15$ ($G_0(j\omega) = 1$) is measured using three different excitation signals with a flat power spectrum: a random noise excitation (zero mean normally distributed), an odd (50 frequencies), and a consecutive (100 frequencies: kf_0, $k = 1, 2, \ldots, 100$) multisine excitation, each with an rms value of 1. For one third of the random multisines, the crest factor was actively pushed down using a crest factor minimizing algorithm (see Chapter 4) to cover the interval [1.4, 2.4]. The power spectrum of all the signals was band-limited with $f_{\max} = 0.1 f_s$. The error

$$\frac{1}{N}\sum_{k=1}^{N} \left|\hat{G}(j\omega_k) - G_0(j\omega_k)\right| \tag{3-29}$$

with $G_0(j\omega) = 1$ and f_k an excited frequency, is plotted as a function of the crest factor for 1000 realizations in Figure 3-9. This figure clearly shows that an odd multisine is significantly better than the consecutive one or the normally distributed noise excitation. The odd-odd multisine has a similar behavior. The errors of the full multisine are also significantly smaller than those of the random excitation. This is due to a similar effect as explained in Chapter 2, where it was shown that at some frequencies the FRF measurements are extremely sensitive to distortions due to the dips that appear in the realized input power spectrum.

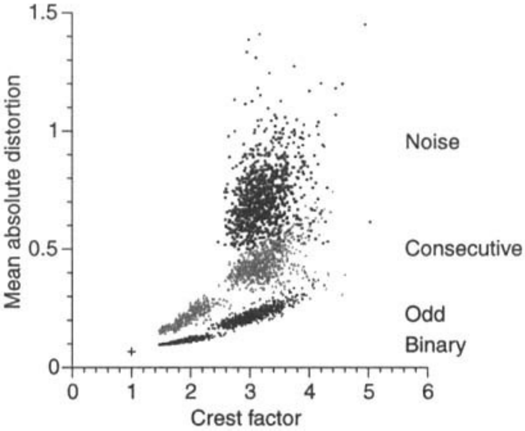


Figure 3-9. Mean absolute distortion for different excitation signals.

Also, a binary signal is created by applying the sign function on the random excitation. The impact of this operation on the power spectrum was studied by Schoukens et al. (1995). For a static nonlinearity, all the realizations result in exactly the same FRF with a very small error. If the nonlinearity is preceded by a dynamic part, the binary behavior will be partly lost and the results will be smeared. □

3.6.2.2 Goal: Measurement of the Best Linear Approximation Using Averaging. If the model is to be used to describe the input-output behavior of the system using a linear system, the related linear dynamic system $G_R(j\omega_k)$ should be measured. From Eq. (3-2) it is seen that it depends on the higher order moments and, so, also on the amplitude distribution of the signal. As a consequence, crest factor minimization is not allowed because this pushes the multisine to a binary behavior (see Chapter 4) and affects the measured $G_R(j\omega_k)$. The measurement of $G_R(j\omega_k)$ is disturbed by two stochastic errors: $N_G(k)$ and $G_S(j\omega_k)$. Both can be reduced by averaging over different realizations of the excitation. For random excitations, the H_1 method (2-43) is recommended, while for the periodic excitations the direct method (2-27) may be used. Again, the periodic excitations are preferred over the random excitations for exactly the same reasons as before: a lower uncertainty for a smaller number of averages.

The impact of the excitation signal on the uncertainty of the related linear dynamic system measurement is, again, illustrated on the nonlinear system. As explained before, in this case the stochastic nonlinear distortions dominate the disturbing noise. In Figure 3-10 the variability of $G_R(j\omega_k)$ is shown for the different, recommended excitation signals. Just as a reference, the disturbing noise level obtained from 10 repeated periods is also shown. On the left side there was no preload of the system, hence, the odd nonlinearity dominates, while on the right side a preload was added resulting in significant even nonlinear contributions. It can be observed that while the consecutive and the odd multisines result in about the same variability if there are no large even distortions (left side), the consecutive multisine has a much larger uncertainty when even distortions exist (right side). Note also that, in all cases, random noise excitations result in worst results and offer no additional advantages compared with random multisines.

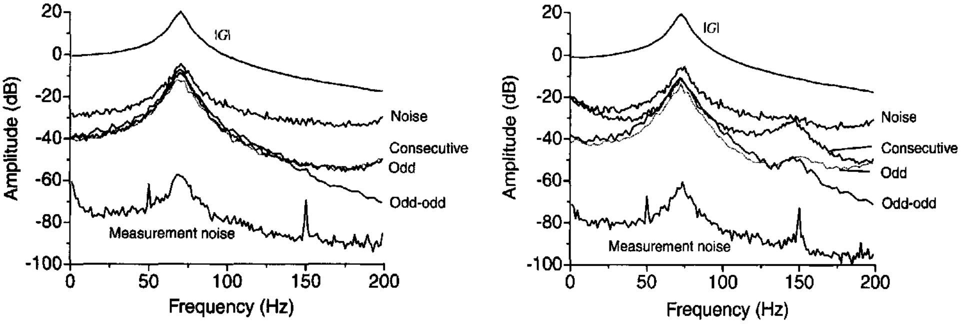


Figure 3-10. Impact of the excitation signal on the uncertainty of the RLDS. Left, no preload (odd nonlinearity), and right, preload (even and odd nonlinearity). Uncertainties on the mean value 30 measurements (noise, full, odd) and 15 measurements for the odd-odd.

Finally, the dependence of the measured $G_R(j\omega_k)$ on the nature of the excitation signal is also illustrated. The result obtained for a random multisine (as advised) is compared with that of a swept sine-like signal (a Schroeder multisine in this case, see Chapter 4). The measurement results are shown in Figure 3-11. Whereas the random multisine still results in an FRF measurement that is very similar to the small-signal results, the Schroeder multisine strongly deviates from it. Without prior knowledge, no second-order system is recognized

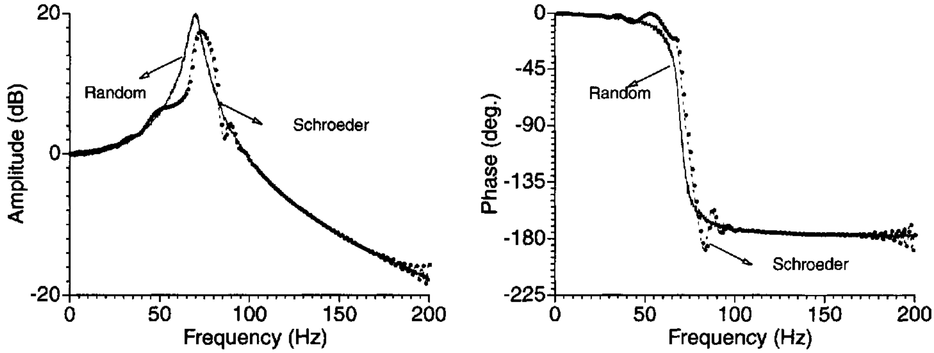


Figure 3-11. Impact of the phase of the multisine on the measured FRF (rms value of 54 mV). Random: odd random multisine with 1342 components. Schroeder: odd multisine with Schroeder phase. This signal acts like a swept sine.

anymore. This illustrates, again, that in the presence of nonlinear distortions, the choice of the excitation signal is crucial. A random multisine combines the advantages of random excitations and periodic excitations, resulting in fast measurements of $G_R(j\omega_k)$, the best linear approximation.

3.6.2.3 Goal: Measurement of the Best Linear Approximation without Averaging. In this case, it is still possible to recover $G_R(j\omega_k)$ using a very dense frequency grid (and hence, again, a long experiment) combined with a random multisine. If the density is very high compared with the variations of $G_R(j\omega_k)$ over ω_k, a local smoothing can be used to reduce the stochastic contributions $G_S(j\omega_k)$ so that improved estimates of $G_R(j\omega_k)$ are obtained at a sparser frequency grid. This technique is completely similar to the empirical transfer function estimate (ETFE) smoothing technique (Ljung, 1999) because $G_S(j\omega_k)$ behaves as noise.

3.7 CONCLUSION

FRF measurements give a great deal of information about the device or plant under test. Very often it is easily accessible and it is strongly advised to take this intermediate step in the identification process. It provides a lot of qualitative information about the complexity of the problem, as well as quantitative information about the plant and the measurement quality. This can be used to set up a measurement-driven weighting function for the identification step and also provides very valuable information for the model validation. The user has a large impact on the measurement quality by generating a good excitation and selecting the proper algorithms to process the raw measurement data. For these reasons, we strongly encourage the reader to take the time to understand the basic principles of FRF measurements. Good, nonparametric measurements will significantly simplify the task of building parametric models.

3.8 EXERCISES

3.1. Generate a random multisine $u_0(k)$ (see Exercise 2.2), with N_p points in one period, that excites the frequency lines, $4k+1$ for $k = 1, \ldots, \text{fix}(N_p/12)$ with equal power. Normalize the rms value of $u(t)$ to 1. Calculate

$$y_0(t) = u_0(t) + 0.1u_0^2(t) + 0.01u_0^3(t) \tag{3-30}$$

Calculate the output spectrum and discuss it. Observe the spectral behavior inside and outside the excited frequency band. Does the behavior depend on the value of N_p?

3.2. Repeat the previous exercise for $u_{\text{rms}} = 10$ and $u_{\text{rms}} = 100$. Discuss the behavior of the even and odd spectral lines.

3.3. Measure the FRF for $u_{\text{rms}} = 1, 10, 100$. Consider, for each situation, 100 realizations of the random multisine. Study the mean value and the standard deviation of the FRF. Extract $G_B(j\omega_k)$, $G_S(j\omega_k)$, $\sigma^2_{G_S}(k)$ and discuss your results.

3.4. Repeat Exercise 3.3 but, this time, use a random multisine that excites all spectral lines between 1 and $\text{FIX}(N_p/12)$. Compare both results and explain the different behavior.

3.5. Repeat Exercise 3.3 but use a zero mean random noise excitation that has approximately the same power spectrum as the excitation in Exercise 3.4.

3.6. Construct a discrete-time Wiener-Hammerstein system $y_0 = WH(u_0)$ (see Figure 3-3) with static nonlinearity: $z = x + 0.1x^2 + 0.01x^3$. Measure $G_R(j\omega_k)$ (make a motivated choice for the power spectrum of the excitation signal) for $u_{\text{RMS}} = 1, 10, 100$. Scale the gain of the first system so that the power of the contribution of degree 3 generates 1% of the linear output power for the first input amplitude. Discuss the results.

3.7. Consider the Wiener-Hammerstein system of Exercise 3.6 and add white, zero mean disturbing noise to the output.

$$y_0 = WH(u_0) \quad \text{and} \quad y(t) = y_0(t) + n_y(t) \tag{3-31}$$

Measure $G_R(j\omega_k)$ again (consider 100 realizations) and calculate $\sigma^2_{\hat{N}_G}(k)$ and $\sigma^2_{\hat{G}_S}(k)$. Use repeated periods to separate the measurement noise $n_y(t)$ from the nonlinear distortions.

3.9 APPENDIXES

Appendix 3.A: Bias and Stochastic Contributions of the Nonlinear Distortions

In this appendix we assume a deterministic amplitude and a uniform continuous phase distribution. The random amplitude, the discrete phase, and the nonuniform continuous phase distributions are commented on in Appendix 3.F.

In order not to overload the notations, the following simplifications are made in this appendix: $G(j\omega_k)$, $G_R(j\omega_k)$, and $G_S(j\omega_k)$ are denoted as $G(k)$, $G_R(k)$, and $G_S(k)$, respectively.

Consider the contribution of degree α to the FRF:

$$\begin{aligned} G^{\alpha}(k) &= \sum_{k_1, \ldots, k_{\alpha-1} = -N}^{N} G^{\alpha}_{L_k, k_1, k_2, \ldots, k_{\alpha-1}} \frac{U_{k_1} U_{k_2} \ldots U_{k_{\alpha-1}} U_{L_k}}{U_k} \\ &= \sum_{k_1, \ldots, k_{\alpha-1} = -N}^{N} \left| G^{\alpha}_{L_k, k_1, k_2, \ldots, k_{\alpha-1}} \right| \frac{|U_{k_1}||U_{k_2}| \ldots |U_{k_{\alpha-1}}||U_{L_k}|}{|U_k|} e^{j\phi(k_1, k_2, \ldots, k_{\alpha-1}, L_k)} \end{aligned} \tag{3-32}$$

with $k = L_k + \sum_{i=1}^{\alpha-1} k_i$, $\phi(k_1, \ldots, L_k) = \sum_{i=1}^{\alpha-1} \varphi_{k_i} + \varphi_{L_k} + \varphi_G - \varphi_k$, $\varphi_{k_i} = \angle U_{k_i}$ and $\varphi_G = \angle G^{\alpha}_{L_k, k_1, k_2, \ldots, k_{\alpha-1}}$. Define the disjoint sets

$$\mathbb{K}_{Bk} = \{(k_1, k_2, \ldots, k_{\alpha-1}) | \phi(k_1, k_2, \ldots, k_{\alpha-1}, L_k) \text{ is independent of } \phi\}$$

$$\mathbb{K}_{Sk} = \{(k_1, k_2, \ldots, k_{\alpha-1}) | \phi(k_1, k_2, \ldots, k_{\alpha-1}, L_k) \text{ depends on } \phi\}$$

with $\phi = \{\varphi_1, \ldots, \varphi_N\}$. The set $\mathbb{K}_{Bk}$ corresponds to the situation where all frequencies, but one (equal to k), can be grouped in pairs $(-l, l)$ so that their phases cancel. This results by definition in contributions to $G^{\alpha}_B(k)$, while this is not the case for the set $\mathbb{K}_{Sk}$ (the phases cannot cancel) so that, by definition, these contribute to $G^{\alpha}_S(k)$. Equation (3-32) becomes

$$\begin{aligned} G^{\alpha}(k) &= G^{\alpha}_S(k) + G^{\alpha}_B(k) \\ G^{\alpha}_S(k) &= \sum_{K \in \mathbb{K}_{Sk}} G^{\alpha}_{L_k, k_1, k_2, \ldots, k_{\alpha-1}} U_{k_1} U_{k_2} \ldots U_{k_{\alpha-1}} U_{L_k} / U_k \\ G^{\alpha}_B(k) &= \sum_{K \in \mathbb{K}_{Bk}} G^{\alpha}_{L_k, k_1, k_2, \ldots, k_{\alpha-1}} U_{k_1} U_{k_2} \ldots U_{k_{\alpha-1}} U_{L_k} / U_k \end{aligned} \tag{3-33}$$

with $K = (k_1, k_2, ..., k_{\alpha-1})$, and where $\sum_{K \in \mathbb{K}_{Sk}}$ and $\sum_{K \in \mathbb{K}_{Bk}}$ denote the sum over all combinations belonging to the sets $\mathbb{K}_{Sk}$ and $\mathbb{K}_{Bk}$, respectively.

In the second part of this appendix, we prove Eq. (3-13). From the definition of $\mathbb{K}_{Bk}$, it follows that the only contributions different from zero are those with α odd. For that reason we focus from now on to $G_B^{2\alpha-1}(k)$. The factor c_α in (3-13) is due to the fact that each of the terms in this sum appears multiple times in the original expression (3-33) or (3-9) where the frequency indices run from $-N$ to N. The number of appearances when starting the sums at zero will be different, and c_α compensated for that. The number of contributions to the sum (3-33) for a given frequency combination $\in \mathbb{K}_{Bk}$ depends on the fact that some of the paired frequencies are equal to each other or not. If some of the paired frequencies are equal to each other or equal to k, there remain less degrees of freedom (because not all paired frequency values can be freely chosen), and, hence, they contribute to the final result only as an $O_\alpha(N^{-p})$, $p \geq 1$ (see also the following Appendices) with $\sum_{\alpha=2}^{\infty} O_\alpha(N^{-p}) = O(N^{-p})$ (the Volterra series converges). Hence, we can focus completely on the situation where all paired frequencies are different from each other and from k. Each such frequency combination appears $(2\alpha - 1)!$ times in (3-33) for $G_B^{2\alpha-1}(k)$, keeping in mind the symmetrical Volterra kernels. In (3-13) each contributing combination appears only $(\alpha - 1)!$ times. Hence, a correction term

$$\frac{(2\alpha-1)!}{(\alpha-1)!} = 2^{\alpha-1}(2\alpha-1)!! \tag{3-34}$$

is needed.

Appendix 3.B: Study of the Moments of the Stochastic Nonlinear Contributions

In this appendix we assume a deterministic amplitude and a uniform continuous phase distribution. The random amplitude, the discrete phase, and the nonuniform continuous phase distributions are commented on in Appendix 3.F.

In order not to overload the notations, the following simplifications are made in this appendix: $G(j\omega_k)$, $G_R(j\omega_k)$, and $G_S(j\omega_k)$ are denoted as $G(k)$, $G_R(k)$, and $G_S(k)$, respectively.

In this appendix, the moments of the stochastic nonlinear contributions $G_S(k)$ are calculated for nonlinear systems belonging to the set $\mathbb{S}$ (Definition 3.5), assuming a normalized random multisine excitation (Definition 3.2). From (3-33) it follows that the stochastic nonlinear contributions to the measurement, at frequency k, are given by multidimensional sums with $(k_1, k_2, \ldots, k_{\alpha-1}) \in \mathbb{K}_{Sk}$, for which it is not possible to partition all the frequencies, but one, in pairs $(-l, l)$. As a consequence, these terms have a random phase such that $\mathcal{E}\{e^{j\varphi}\}=0$. It follows directly that $\mathcal{E}\{G_S^\alpha(k)\} = 0$, and, hence, $\mathcal{E}\{G_S(k)\} = \mathcal{E}\{\sum_{\alpha=2}^{\infty} G_S^\alpha(k)\} = 0$. The study of the higher order moments is much more complicated. The basic idea is first to prove that

$$\left|\mathcal{E}\{G_S^{r_1}(k_1)G_S^{r_2}(k_2)...G_S^{r_n}(k_n)\}\right| \leq O(N^{-p})M_U^{-n}\prod_{i=1}^{n} M_{G^{r_i}} M_U^{r_i} \tag{3-35}$$

for arbitrary n, where p depends on the actual situation. $G_S^{r_i}(k_i)$ stands for the stochastic nonlinear contribution of degree r_i at frequency k_i. Next, using (3-35) and Definition 3.5, we find

$$\begin{aligned}\left|\mathcal{E}\{\, G_S(k_1)G_S(k_2)\ldots G_S(k_n)\}\right| &\leq \sum\nolimits_{r_1=2}^{\infty}\ldots\sum\nolimits_{r_n=2}^{\infty}\left|\mathcal{E}\{\, G_S^{r_1}(k_1)G_S^{r_2}(k_2)\ldots G_S^{r_n}(k_n)\}\right| \\ &\leq O(N^{-p})M_U^{-n}\prod\nolimits_{i=1}^{n}\sum\nolimits_{r_i=2}^{\infty}M_{G^{r_i}}M_U^{r_i} \\ &\leq O(N^{-p})M_U^{-n}C_1^n\end{aligned} \tag{3-36}$$

so that $\mathcal{E}\{\, G_S(k_1)G_S(k_2)\ldots G_S(k_n)\}$ converges to zero, at least, as an $O(N^{-p})$.

Lemma 3.19 (number of nonzero contributions): Consider a system belonging to the set $\mathbb{S}$, excited with a random multisine $u_N \in \mathbb{E}_N$. Under Definitions 3.2 and 3.5, the expected value $\mathcal{E}\{\, G_S^{r_1}(k_1)G_S^{r_2}(k_2)\ldots G_S^{r_n}(k_n)\}$ is bounded by

$$\left|\mathcal{E}\{\, G_S^{r_1}(k_1)G_S^{r_2}(k_2)\ldots G_S^{r_n}(k_n)\}\right| \leq O(N^{-\nu})M_U^{-n}\prod\nolimits_{i=1}^{n}M_{G^{r_i}}M_U^{r_i} \tag{3-37}$$

with $\nu = \text{int}((n-2m+1)/2)$ and where m is the number of pairs $(k_i, k_j{=}-k_i)$ that can be formed in the set $\{k_1, k_2, \ldots, k_n\}$.

Note: If the number of unpaired frequencies k_i is odd, then $\nu = 1$, while $\nu = 0$ if it is even.

Proof. The basic idea is to count the number of nonzero contributions in $\mathcal{E}\{\, G_S^{r_1}(k_1)G_S^{r_2}(k_2)\ldots G_S^{r_n}(k_n)\}$ as a function of N. Note that each of the terms in the product $G_S^{r_i}(k_i)$ consists of a multiple sum over the frequencies; see Eq. (3-33). The terms in the product $G_S^{r_1}(k_1)G_S^{r_2}(k_2)\ldots G_S^{r_n}(k_n)$ that have a nonzero expected value are those where all phases of the participating frequency components cancel each other. This means that we have to look for frequency pairs $(l, -l)$ having a zero phase contribution.

Consider the frequencies that contribute to $G_S^{r_i}(k_i)$, $i = 1, \ldots, n$:

$$\begin{array}{ll} -k_1 & (l_1(k_1), l_2(k_1), \ldots, l_{r_1-1}(k_1), l_{r_1}(k_1)) \\ -k_2 & (l_1(k_2), l_2(k_2), \ldots, l_{r_2-1}(k_2), l_{r_2}(k_2)) \\ & \ldots \\ -k_n & (l_1(k_n), l_2(k_n), \ldots, l_{r_n-1}(k_n), l_{r_n}(k_n)) \end{array} \tag{3-38}$$

with

$$l_{r_i}(k_i) = k_i - \sum\nolimits_{p=1}^{r_i-1} l_p(k_i),\ i = 1, \ldots, n \tag{3-39}$$

The frequency $-k_i$ (called denominator frequencies) comes from the denominator in (3-33), where the minus sign accounts for the negative phase contribution of the denominator term; $(l_1(k_i), l_2(k_i), \ldots, l_{r_i}(k_i))$ are the frequencies in the numerator of (3-33) (called numerator frequencies) and their sum should be equal to k_i in order to get a contribution at frequency k_i. The total number of numerator frequencies participating in the sums is $F_a = \sum_{i=1}^{n} r_i$. Equation (3-39) imposes n constraints, so that the total number of degrees of freedom at this moment is $F_a - n$.

The only nonzero contributions to the expected value (3-35) are those where all frequencies (numerator frequencies and unpaired denominator frequencies) can be grouped in pairs $(-l, l)$ such that their phase contributions are canceled. This pairing process will be imposed step by step (first on the denominator frequencies k_i, next on the remaining numerator frequencies $l_j(k_h)$), and the additional constraints on the free frequencies in (3-38) will be checked.

3.B.1 Denominator Frequencies

(i) First pair the denominator frequencies $(-k_i = k_j)$. Assume there are m such pairs.

(ii) All remaining $n - 2m$ unpaired denominator frequencies k_i should be paired with one of the numerator frequencies $l_j(k_h)$, $h = 1, \ldots, n$ and $j = 1, \ldots, r_h$. Because the denominator frequencies have fixed values (no summing over k_i), this fixes $n - 2m$ numerator frequencies.

(iii) Eventually, after pairing all denominator frequencies, the number of free frequencies is $F_{\mathrm{a}} - n - (n - 2m) = F_{\mathrm{a}} - 2n + 2m$. Note that the worst case situation appears when m is maximized because this leaves the maximum number of numerator frequencies free.

3.B.2 Numerator Frequencies. Next the remaining numerator frequencies should be paired. These can be partitioned in two groups: the free numerator frequencies $(F_{\mathrm{a}} - 2n + 2m)$ and the (n) dependent frequencies $l_{r_i}(k_i)$. We impose pairs only on the free frequencies, assuming that the dependent frequencies are then automatically paired. This is again the worst case situation (the largest number of free frequencies), since in the other case additional constraints would be imposed. Note also that pairing is a worst case phase canceling process: grouping four or more frequencies together is a stronger restriction than making pairs of two frequencies. Two situations will be considered: n is even or n odd.

(i) n is even (F_{a} is even, otherwise there would always remain an unpaired frequency and these terms have zero mean): all free frequencies can be paired, resulting in $(F_a - 2n + 2m)/2$ pairs where the frequency can be freely chosen. So the maximum number of zero phase terms in $G_S^{r_1}(k_1)G_S^{r_2}(k_2)\ldots G_S^{r_n}(k_n)$ is an $O(N^{\nu_0})$ with $\nu_0 = (F_a - 2n + 2m)/2$. From Definitions 3.2 and 3.5 and (3-9), it follows that each term in the sum of $G_S^{r_i}(k_i)$ is an $O(N^{\nu_i})M_{G^{r_i}}M_U^{r_i - 1}$, with $\nu_i = (1 - r_i)/2$, and, hence, the expected value is bounded by

$$\left|\mathcal{E}\{\, G_S^{r_1}(k_1)G_S^{r_2}(k_2)\ldots G_S^{r_n}(k_n)\,\}\right| \le O(N^{-\nu})M_U^{-n}\prod\nolimits_{i=1}^{n} M_{G^{r_i}}M_U^{r_i} \qquad (3\text{-}40)$$

with $\nu = -\nu_0 - \sum_{i=1}^{n} \nu_i = (n - 2m)/2$ for n even.

(ii) n is odd (F_{a} is odd, otherwise there would always remain an unpaired frequency). In this case, not all the free numerator frequencies $(F_{\mathrm{a}} - 2n + 2m)$ can be paired since they are odd in number. So $(F_a - 2n + 2m - 1)/2$ pairs of free frequencies can be formed, and there remains one unpaired free frequency that should be combined with one depended frequency. Again we can assume that the other $n - 1$ (an even number) depended frequencies are then automatically paired (worst case). So the question is whether the last pairing step (independent frequency = - dependent frequency) creates a new constraint. To answer this question, it is important to note that not all numerator frequencies, but one, in a row of (3-38) can be paired to each other, because this would be a systematic contribution (see Appendix 3.A). As a consequence, the depended frequency $l_{r_i}(k_i)$ (3-38) cannot be paired with another frequency in its own row. This would either impose a new constraint in this row (put $l_{r_i}(k_i) = -l_p(k_i)$ for a given p in (3-39)) or create a systematic contribution. So the last pair (independent frequency, depended frequency) should be formed over two different rows (connected to $k_i, k_j,\ i \neq j$). Because the constraints (3-39) are active only row by row (they combined frequencies of the same row), this creates an additional constraint, and, hence, the frequency of the last pair is fixed by this constraint. So the number of free pairs is an $O(N^{\nu_0})$ with $\nu_0 = (F_a - 2n + 2m - 1)/2$. Because each contribution in the sum of $G_S^{r_i}(k_i)$ is an $O(N^{\nu_i})$, with $\nu_i = (1 - r_i)/2$, it is clear that the expected value is bounded by

$$\left|\mathcal{E}\{\, G_S^{r_1}(k_1)G_S^{r_2}(k_2)...G_S^{r_n}(k_n)\}\right| \leq O(N^{-\nu})M_{\bar{U}}^{n}\prod\nolimits_{i=1}^{n} M_{G^{r_i}}M_U^{r_i} \qquad (3\text{-}41)$$

with $\nu = -\nu_0 - \sum_{i=1}^{n}\nu_i = (n - 2m + 1)/2$ for n odd.

The bound in the results, (3-40) and (3-41), can be written as $O(N^{-\text{int}((n-2m+1)/2)})$, which proves the lemma. □

Theorem 3.20 (Moments Stochastic Nonlinear Contributions): Consider a system belonging to the set $\mathbb{S}$ (see Definition 3.5), excited with a random multisine $u_N \in \mathbb{E}_N$ (see Definition 3.2). The expected value $\mathcal{E}\{\, G_S(k_1)G_S(k_2)...G_S(k_n)\}$ is bounded by $\left|\mathcal{E}\{\, G_S(k_1)G_S(k_2)...G_S(k_n)\}\right| \leq O(N^{-\nu})M_{\bar{U}}^{n}C_1^n$, with $\nu = \text{int}((n - 2m + 1)/2)$.

Proof. The proof follows directly from Lemma 3.19, by the fact that ν is independent of r_i, $i = 1, 2, ..., n$. Hence, Lemma 3.19 can be directly generalized to Theorem 3.20. □

Theorem 3.21 (Properties Stochastic Nonlinear Contributions): Consider a system belonging to the set $\mathbb{S}$ (Definition 3.5), excited with a random multisine $u_N \in \mathbb{E}_N$ (Definition 3.2). The stochastic nonlinear contributions $G_S(k)$ have the following properties:

1. $\mathcal{E}\{\, G_S(k)\bar{G}_S(l)\} = O(N^{-1})$ for $k \neq l$
2. $\mathcal{E}\{\, |G_S(k)|^2\} \equiv \sigma_S^2(k) = O(N^0)$
3. $\mathcal{E}\{\, G_S(k)|G_S(l)|^2\} = O(N^{-1})$
4. $\mathcal{E}\{\, (|G_S(k)|^2 - \sigma_S^2(k))(|G_S(l)|^2 - \sigma_S^2(l))\} = O(N^{-1})$ for $k \neq l$
5. $\mathcal{E}\{\, G_S(k)\bar{G}_S(k+m)\bar{G}_S(l)\bar{G}_S(l+m)\} = O(N^{-2})$ for $k \neq l, -k \neq l+m, -l \neq k+m$, $m \neq 0$ (all frequencies differ from each other)
6. $\mathcal{E}\{\, |G_S(k)|^2|G_S(k+m)|^2\} = O(N^0)$ for $m \neq 0$

Proof. The proof consists of a straightforward application of Theorem 3.20. Note that ν is maximal if the number of paired numerator frequencies is maximized.

1. $\mathcal{E}\{\, G_S(k)\bar{G}_S(l)\} = G_S(k)G_S(-l)\} = O(N^{-1})$ for $k \neq l$.
   If $k \neq l$, then $m = 0$ and, hence, $\nu = \text{int}((n - 2m + 1)/2) = \text{int}((2 - 0 + 1)/2) = 1$.
2. $\mathcal{E}\{\, |G_S(k)|^2\} = \mathcal{E}\{\, G_S(k)G_S(-k)\} = O(N^0)$.
   In this case $m = 1$, $n = 2$ and $\nu = \text{int}((n - 2m + 1)/2) = \text{int}((2 - 2 + 1)/2) = 0$.

3. $\mathcal{E}\{ G_S(k)|G_S(l)|^2\} = \mathcal{E}\{ G_S(k)G_S(l)G_S(-l)\} = O(N^{-1})$.
   $m = 1$, $n = 3$ and $\nu = \text{int}((n-2m+1)/2) = \text{int}((3-2+1)/2) = 1$.
4. $\mathcal{E}\{ (|G_S(k)|^2 - \sigma_S^2(k))(|G_S(l)|^2 - \sigma_S^2(l))\} = O(N^{-1})$ for $k \neq l$.

Here, some precautions have to be taken. In order to simplify the proof, the expected value is rewritten as

$$\mathcal{E}\{ (|G_S(k)|^2 - \sigma_S^2(k))(|G_S(l)|^2 - \sigma_S^2(l))\} = \mathcal{E}\{ |G_S(k)|^2|G_S(l)|^2\} - \sigma_S^2(k)\sigma_S^2(l) \tag{3-42}$$

We study the first term in (3-42)

$$\mathcal{E}\{ |G_S(k)|^2|G_S(l)|^2\} = \mathcal{E}\{ G_S(k)G_S(-k)G_S(l)G_S(-l)\}$$

Here, two disjoint situations can be considered. In the first situation (A), all denominator frequencies are paired $(k, -k)$ and $(l, -l)$ so that $\nu = \text{int}((4-4+1)/2) = 0$, while in the second situation (B), at least one of the pairs $(k, -k)$ or $(l, -l)$ is not created in the pairing process so that $m \leq 1$, and $\nu = 1$. The expected value can be split over these two types of contributions.

$$\mathcal{E}\{ |G_S(k)|^2|G_S(l)|^2\} = \mathcal{E}\{ |G_S(k)|^2|G_S(l)|^2\}_A + \mathcal{E}\{ |G_S(k)|^2|G_S(l)|^2\}_B \tag{3-43}$$

(i) First, situation (A) is studied. Again two possibilities exist: 1) some pairs link the k-lines to the l-lines; 2) no such links appear.
First we deal with situation 1: From claims 2 and 3 in Appendix 3.E, it follows that such combinations are an $O(N^{-1})$, so these terms do not act as the dominating contributions. Situation 2: Here the k-lines are not lined to the l-lines. Because the combinations no longer depend on the phase (sum of the phases is zero), they are deterministic contributions and, hence,

$$\begin{aligned}&\mathcal{E}\{ |G_S(k)|^2|G_S(l)|^2\}_A \\ &= \mathcal{E}\{ |G_S(k)|^2\}\mathcal{E}\{ |G_S(l)|^2\} + O(N^{-1}) = \sigma_S^2(k)\sigma_S^2(l) + O(N^{-1})\end{aligned} \tag{3-44}$$

Clearly (3-44) cancels the second term in (3-42).

(ii) In set (B), we have that $\nu = 1$, and, hence, it has again an $O(N^{-1})$ contribution to (3-42).

We conclude that (3-42) is an $O(N^{-1})$.

5. The proofs of 5 and 6 are completely similar to any one of the previously studied situations.

Appendix 3.C: Mixing Property of the Stochastic Nonlinear Contributions

In this appendix we assume a deterministic amplitude and a uniform continuous phase distribution. The random amplitude, the discrete phase, and the nonuniform continuous phase distributions are commented on in Appendix 3.F.

In this appendix, the proof of Theorem 3.10 is given: Consider a system belonging to the set $\mathbb{S}$ (Definition 3.5), excited with a random multisine $u_N \in \mathbb{E}_N$ (Definition 3.2). The (stochastic) nonlinear contributions $G_S(k)$ are mixing of arbitrary order n.

Proof. We prove the mixing property for the nonlinear contributions $G_B(k) + G_S(k)$. Because $G_B(k)$ is deterministic, the mixing property of $G_S(k)$ follows immediately. We show that $G^{r_1}(k_1)G^{r_2}(k_2)...G^{r_n}(k_n)$ are mixing, for an arbitrary n. The theorem follows then from Definition 3.5 and the linearity property of mixing variables (Lemma 14.4). $G^{r_1}(k_1)G^{r_2}(k_2)...G^{r_n}(k_n)$ is mixing if

$$\max_{k_n} \sum_{k_1, k_2, \ldots, k_{n-1} = -N}^{N} \left|\text{cum}(G^{r_1}(k_1), G^{r_2}(k_2), \ldots, G^{r_{n-1}}(k_{n-1}), G^{r_n}(k_n))\right| \le c_n < \infty \tag{3-45}$$

for any N, infinity included, with c_n independent of N. Using Lemma 14.4 and Definition 3.5, it turns out that it is sufficient to prove that

$$\max_{k_n} \sum_{k_1, \ldots, k_{n-1} = -N}^{N} \sum_{l_i(k_i) = -N}^{N} \left|\text{cum}\left(U_{k_1}^{-1} \prod_{i=1}^{r_1} U_{l_i(k_1)}, \ldots, U_{k_n}^{-1} \prod_{i=1}^{r_n} U_{l_i(k_n)}\right)\right| \le c_n < \infty \tag{3-46}$$

for any N, infinity included, with c_n independent of N. In this expression $\sum_{l_i(k_i) = -N}^{N}$ stands for the sum over all numerator frequencies $l_1(k_1)$, $l_2(k_1)$, ..., $l_{r_1}(k_1)$, $l_1(k_2)$, ..., $l_{r_2}(k_2)$, ..., $l_{r_n}(k_n)$ (see Appendix 3.B) appearing in $G^{r_1}(k_1)G^{r_2}(k_2)...G^{r_n}(k_n)$. To calculate the cumulant we have to set up a table with all participating input Fourier coefficients (characterized by their frequency) and consider next all indecomposable sets in this table (see Appendix 14.A). The table is given by (see also 3-38)

$$\begin{array}{ll} -k_1 & (l_1(k_1), l_2(k_1), \ldots, l_{r_1-1}(k_1), l_{r_1}(k_1)) \\ -k_2 & (l_1(k_2), l_2(k_2), \ldots, l_{r_2-1}(k_2), l_{r_2}(k_2)) \\ & \ldots \\ -k_n & (l_1(k_n), l_2(k_n), \ldots, l_{r_n-1}(k_n), l_{r_n}(k_n)) \end{array} \tag{3-47}$$

with

$$k_i - \sum_{p=1}^{r_i - 1} l_p(k_i) - l_{r_i}(k_i) = 0, \; i = 1, \ldots, n \tag{3-48}$$

All frequencies but one (k_n) appear as a summation index in (3-46). We will count again the number of nonzero cumulants over the indecomposable partitions that appear in the sum. To do so, we have to determine the maximum number of degrees of freedom, taking into account all restrictions that will appear. The following constraints will be considered:

(i) The $\text{cum}(U_{j_1}, U_{j_2}, \ldots, U_{j_s})$ is different from zero only if $|j_1| = |j_2| = \ldots = |j_s|$ and the terms are paired. Hence, only cumulants over sets with an even number of elements can be different from zero. The sum of the frequencies in such a set is zero.

(ii) All the row constraints (3-48) should be respected.

(iii) All frequencies are different from zero $j_i \neq 0$.

(iv) Only indecomposable partitions are considered.

The constraint (3-48) can also be written as

$$A_1 J_a = 0 \tag{3-49}$$

where J_a is a vector containing all frequencies that participate in (3-47). The entries of A_1 are 1, −1, or 0 depending on how the corresponding frequency in J_a contributes to the corresponding row. Note that the "indecomposability" property is completely preserved in A_1.

PARTITIONING. Consider an indecomposable partition of (3-47) and select the partitions that have nonzero cumulants. On each subset of such partition we can associate one frequency (see condition 1 above). All these frequencies are put in the vector J_p, and we replace the set of equations (3-49) by

$$AJ_p = 0 \tag{3-50}$$

Some of the subsets will combine only frequencies belonging to one row. Because the sum over all these frequencies in such a subset is zero (see condition 1 above), their entry in A is zero. So only subsets that combine frequencies from different rows can have an entry in A that is different from zero. If such a subset (over different rows) has zero entries in A, it can be split in smaller subsets with nonzero entries (the partition remains indecomposable). This is a worst case situation because a smaller number of frequencies are linked to each other, and, hence, a larger number of free frequencies remains. For example,

$$\begin{bmatrix} l & -l \\ -l & l \end{bmatrix} \rightarrow \begin{bmatrix} l_1 \\ -l_1 \end{bmatrix} \begin{bmatrix} l_2 \\ -l_2 \end{bmatrix} \tag{3-51}$$

With these replacements, the structure of A and A_1 is the same with respect to the indecomposable partitions: A is indecomposable $\Leftrightarrow$ A_1 is indecomposable. So from now on we focus completely on A.

Note that the entries corresponding to a given frequency in J_p appear at most in one column in A.

A can also have subsets with an odd number of entries (e.g., three). However, because each subset covers an even number of frequencies, such a subset corresponds to a subset in A_1 with an even number of entries, e.g.,

$$\begin{bmatrix} 1 \\ 1 \\ -2 \end{bmatrix} \text{ in } A \quad \leftrightarrow \quad \text{corresponds for example to } \begin{bmatrix} 1 & \\ 1 & \\ -1 & -1 \end{bmatrix} \text{ in } A_1 \tag{3-52}$$

Such a set can always be broken into

$$\begin{bmatrix} 1 \\ \\ -1 \end{bmatrix} \begin{bmatrix} \\ 1 \\ -1 \end{bmatrix} \tag{3-53}$$

without changing the indecomposable structure. Again, this is a worst case situation. So, we should consider only subsets with an even number of entries in A.

Indecomposable Partitions. Only the indecomposable partitions are considered. It is possible to select a submatrix in A, A_{ind}, that is indecomposable. After rearranging the order of the columns, A can be written as $A = [A_{\text{ind}} A_{\text{rest}}]$.

No Zero Frequencies. None of the frequencies in J_p (3-50) may be equal to zero. So every row in A should contain at least two entries that are different from zero, otherwise (3-50) forces at least one frequency to be zero. Hence, it is possible to form a matrix $\tilde{A}$ by extending A_{ind} with additional columns of A_{rest}, such that each row of $\tilde{A}$ contains at least two nonzero entries.

Structure of $\tilde{A}$. We study the structure of $\tilde{A}$ in more detail in Appendix 3.D, where it is shown that $\tilde{A}$ can always be rearranged (some columns might be shifted back to A_{rest}) to a matrix with $\sum_{k=1}^{k_{\max}} 2kN_{2k} \geq 2n$ entries grouped in $\sum_{k=1}^{k_{\max}} N_{2k}$ columns, and $\text{rank}(\tilde{A}) = \sum_{k=1}^{k_{\max}} N_{2k} - 1$ (N_{2k} is the number of subsets in $\tilde{A}$ with $2k$ elements). Because $\text{rank}(A) \geq \text{rank}(\tilde{A})$ it follows that at most one frequency can be freely chosen.

Number of Degrees of Freedom. J_p contains at most $F_a + n - 2$ free frequencies, with $F_a = \sum_{i=1}^{n} r_i$, because at least one frequency is paired with $-k_n$. Each entry of $\tilde{A}$ corresponds to at least one free frequency in (3-47), so $\sum_{k=1}^{k_{\max}} 2kN_{2k} \geq 2n$ frequencies of (3-47) are used in $\tilde{A}$ while at most one is free (see above). The maximum number of degrees of freedom (worst case) appears when all remaining free frequencies (in A_{rest}) $F_a + n - 2 - \sum_{k=1}^{k_{\max}} 2kN_{2k} \leq F_a + n - 2 - 2n$ are grouped in pairs. So the free number of frequencies (including the free one of $\tilde{A}$) is given by

$$F_{\text{free}} \leq (F_a - n - 2)/2 + 1 \leq (F_a - n)/2 \tag{3-54}$$

Each of these frequencies can be freely chosen out of the $2N$ input frequencies. The number of degrees of freedom is thus an $O(N^{(F_a - n)/2})$.

Mixing. Because $U_l = O(N^{-1/2})$, each cumulant in the sum (3-46) is an $O(N^{(n - F_a)/2})$. Each cumulant in (3-46) is calculated as the sum over all indecomposable partitions of table (3-47), which reduces the number of free frequencies in the sums (3-46) to $(F_a - n)/2$ (see (3-54)). Hence, (3-46) is an $O(N^{(F_a - n)/2})O(N^{(n - F_a)/2}) = O(N^0)$, which proves the theorem.

Appendix 3.D: Structure of the Indecomposable Sets

In this appendix we assume a deterministic amplitude and a uniform continuous phase distribution. The random amplitude, the discrete phase, and the nonuniform continuous phase distributions are commented on in Appendix 3.F.

The matrix $\tilde{A}$ contains an indecomposable set extended with additional columns such that each row contains at least two nonzero entries. These additional columns might create additional links between the rows so that it might be possible to "break" larger subsets to smaller ones, the smallest ones corresponding to pairs, while the number of degrees of freedom is not decreased (so the worst case is maintained). The breaking process can be continued until all subsets are reduced to pairs, or the remaining subset is an "essential" set S_e with $2k$ ($k \geq 2$) entries in $\tilde{A}$ that cannot be broken without losing the indecomposability of $\tilde{A}$. This leads to the following definition.

Definition 3.22: The subset S_e with $2k$ $(k \geq 2)$ entries in $\tilde{A}$ is an essential subset if it is possible to define a partition on the entries of $\tilde{A}$, $\{S_e, \tilde{A}_1, \ldots, \tilde{A}_{2k}\}$, where each of the subsets $\tilde{A}_i$ is indecomposable and linked to only one element of S_e.

Lemma 3.23: Consider a subset S_i with $2k$ $(k \geq 2)$ entries in $\tilde{A}$. Either it is possible to brake it into two subsets S_{i1} and S_{i2}, without losing the indecomposability of $\tilde{A}$, or S_i is an essential subset.

Proof. The lemma follows directly from the definition. If S_i is not an essential subset, there is a partitioning in $\tilde{A}$, where at least one of the subsets is linked to two elements of S_i. Hence, S_i can be broken into two parts, each containing one of these elements, without losing the indecomposability of the partition. □

After repetitively applying Lemma 3.23, the matrix $\tilde{A}$ is partitioned in pairs and essential subsets. Consider, for example, a situation with one essential subset:

$$\left[\begin{array}{cccccc}
x & x & & & & \\
x & & & x & & \\
x & & & & & \\
x & & & & & \\
 & x\ \{\tilde{A}_1\} & & & & \cdots \\
 & & \cdots & & & \\
 & & & x\ \{\tilde{A}_2\} & & \\
 & & & & \cdots &
\end{array}\right] \tag{3-55}$$

with x a nonzero entry in $\tilde{A}$, and $\tilde{A}_i$ indecomposable sets consisting of pairs. Hence, their structure can always be written as

$$\left[\begin{array}{cccccc}
x & & & & & \\
x & x & & & & y \\
 & x & x & \cdots & & \\
 & & x & & & \\
 & & \cdots & & & \\
 & & & & x & \\
 & & & & x & x
\end{array}\right] \tag{3-56}$$

The entry y in the last column can appear at any of the rows but the last one. It is clear that the rank of this square matrix is the number of columns −1 because the sum of all entries in one column is zero. So only one frequency is free. This is the frequency that is linked to the essential set, so that no free frequency remains. This idea can be further extended to situations with multiple essential sets or no essential set (where one of the pairs can be considered as a special case of essential set). The conclusion is that the rank of $\tilde{A}$ is the number of columns −1.

Note. During the breaking process, additional depended columns might appear. These are shifted back from $\tilde{A}$ to A_{rest}.

Lemma 3.24: The matrix $\tilde{A}$ can be reduced using the breaking and column-removing process to a matrix with $\text{rank}(\tilde{A}) = \sum_{k=1}^{k_{\max}} N_{2k} - 1$, with N_{2k} the number of sets with $2k$ entries.

Appendix 3.E: Distribution of the Stochastic Nonlinearities

In this appendix we assume a deterministic amplitude and a uniform continuous phase distribution. The random amplitude, the discrete phase, and the nonuniform continuous phase distributions are commented on in Appendix 3.F.

In this appendix, the proof of Theorem 3.11 is given: for a system belonging to the system set $\mathbb{S}$, excited with a random multisine $u_N \in \mathbb{E}_N$, the stochastic nonlinearities are circular normally distributed. The frequency index k is sometimes omitted for notational simplicity.

The proof consists of the following steps.

- The Volterra series can be written as the sum of contributions up to degree M (G_S^+) plus a rest term, G_S^-, which is an $O(\varepsilon)$.
- Each of the M terms is normally distributed, and their variances are an $O(\varepsilon^0)$. Also the variance of G_S^+ is an $O(\varepsilon^0)$, while the variance of the rest term is an $O(\varepsilon^2)$. So, G_S converges in distribution to G_S^+, which is a finite sum of circular normally distributed variables. So, also G_S^+ is circular normally distributed.

(i) $G_S(k) = G_S^+(k) + G_S^-(k)$, with $\sigma_{G^-}^2 = O(\varepsilon^2)$, $\sigma_{G^+}^2 = O(\varepsilon^0)$, and ε arbitrary small.

Proof. $G_S(k) = \sum_{\alpha=2}^{\infty} G_S^{\alpha}(k)$, with $\sum_{\alpha=1}^{\infty} M_{G^\alpha} M_U^{\alpha} \le C_1 < \infty$ (Definition 3.5). So,

$$\forall \varepsilon, \exists\, M \text{ s.t. } \sum_{\alpha=M+1}^{\infty} M_{G^\alpha} M_U^{\alpha} < \varepsilon \Rightarrow |G_S^-(k)| = \left|\sum_{\alpha=M+1}^{\infty} G_S^{\alpha}(k)\right| = O(\varepsilon)$$

The variance of $G_S^-(k)$ can be bounded above by

$$\sigma_{G^-}^2(k) = \mathcal{E}\{ G_S^- \overline{G_S^-} \} = \sum_{\alpha_1, \alpha_2 = M+1}^{\infty} \mathcal{E}\{ G_S^{\alpha_1} \overline{G}_S^{\alpha_2} \} \le \sum_{\alpha_1, \alpha_2 = M+1}^{\infty} \left|\mathcal{E}\{ G_S^{\alpha_1} \overline{G}_S^{\alpha_2} \}\right| \tag{3-57}$$

From Lemma 3.19 $(n = 2, m = 1 \rightarrow \nu = 0)$, it follows that

$$\left|\mathcal{E}\{ G_S^{\alpha_1} \overline{G}_S^{\alpha_2} \}\right| \le O(N^0) M_U^{-2} M_U^{\alpha_1 + \alpha_2} M_{G^{\alpha_1}} M_{G^{\alpha_2}} \tag{3-58}$$

Combining (3-57) and (3-58) gives

$$\sigma_{G^-}^2(k) \le \frac{O(N^0)}{M_U^2} \left(\sum_{\alpha_1 = M+1}^{\infty} M_{G^{\alpha_1}} M_U^{\alpha_1}\right) \left(\sum_{\alpha_2 = M+1}^{\infty} M_{G^{\alpha_2}} M_U^{\alpha_2}\right) \le O(\varepsilon^2) \tag{3-59}$$

Similarly, it is shown that $\sigma_{G^+}^2(k) = O(\varepsilon^0)$.

(ii) All odd moments $\mathcal{E}\{ (G_S^{\alpha}(k))^{2p+1} \} = O(N^{-1})$.

Proof. Using Lemma 3.19, with $n = 2p + 1$, $m = p$, it follows that $\nu = 1$, which proves the statement.

(iii) Study of the even moments $\mathcal{E}\{|G_S^{\alpha}(k)|^{2p}\}$

We use again the notation of Appendix 3.C. In (3-47) we put $k_{2i-1} = k$ $(i = 1, \ldots, p$, and $k_{2i} = -k$, and define the set of equations

$$K = BJ_p\text{, with } K = (k, -k, \ldots, k, -k)^T \tag{3-60}$$

From Appendix 3.C, we know that the worst case (maximum number of combinations) is given if the denominator frequencies are paired with each other, because this leaves the largest number of frequencies free. Hence, the numerator frequencies should be partitioned s.t. the phases cancel each other. Just as in Appendix 3.C, the subsets can be each time restricted to depend on only one frequency (otherwise they can be broken into smaller subsets without changing their contribution). Next we prove a number of additional properties on the grouping process.

Claim 1: *Partitions that contain subsets linking more than two rows in (3-50) give only $O(N^{-\nu})$, $\nu \geq 1$ contributions.*

Proof. Consider the set of equations (3-50). Each row in B has more than one entry different from zero, because otherwise it would be a systematic contribution instead of a stochastic one (all frequencies, but one, are paired). So there is a submatrix $\tilde{B}$ in B, after rearranging the columns, that contains at least $4p$ entries. Using the definitions of Appendix 3.C, the number of entries in $\tilde{B}$ is $\sum_{k=1}^{k_{\max}} 2kN_{2k} \geq 4p$, and the number of columns (set frequencies) is $\sum_{k=1}^{k_{\max}} N_{2k}$, where, for the same reason as explained in (3-52), only subsets with an even number of entries are considered. So after pairing, the total number of independent frequencies is

$$\frac{(F_a - 2p)_1 - \left(\sum_{k=1}^{k_{\max}} 2kN_{2k}\right)_2}{2} + \left(\sum_{k=1}^{k_{\max}} N_{2k}\right)_3 = \frac{F_a - 2p}{2} - \sum_{k=1}^{k_{\max}} (k-1)N_{2k} \tag{3-61}$$

with $(\ )_1$ the total number of independent frequencies after imposing the row constraints (3-39), $(\ )_2$ the number of entries used in $\tilde{B}$, and $(\ )_3$ the number of set frequencies in $\tilde{B}$. Each of these combinations is an $O(N^{(2p - F_a)/2})$. If $\exists\ k > 1 N_{2k} \neq 0$, then the second part in (3-61) is negative, and consequently the claim is proved.

Conclusion. Only pairs should be considered.

Claim 2: *Partitions, using pairs as subsets, that link more than two rows $(k, -k)$ in (3-50) give only $O(N^{-\nu})$, $\nu \geq 1$ contributions.*

Proof. For such a partition, keeping in mind that each row should contain at least two entries different from zero, B should contain at least the following submatrix $\tilde{B}$:

$$\begin{array}{c} k \\ -k \\ k \\ -k \end{array} \quad \begin{bmatrix} 0 & \mathrm{x} & 0 & \mathrm{x} & 0 \\ \mathrm{x} & \mathrm{x} & 0 & \mathrm{x} & 0 \\ \mathrm{x} & 0 & \mathrm{x} & 0 & \mathrm{x} \\ 0 & 0 & \mathrm{x} & 0 & \mathrm{x} \end{bmatrix} \tag{3-62}$$

where $\mathrm{x} = \pm 1$. It is clear that $\tilde{B}$, consisting of q columns, has rank 3 and uses $2q$ entries. Assuming that the row conditions for the corresponding lines are automatically met, we get that the number of free frequencies in $\tilde{B}$ is $q-3$. The remaining $2p-4$ row conditions should still be met, so that there are $2p-4$ dependent variables. Hence, the number of free pairs is

$$(F_a - (2p-4) - 2q)/2 + q - 3 = (F_a - 2p)/2 - 1 \tag{3-63}$$

Because $U_l = O(N^{-1/2})$, each term in the sums of $\mathcal{E}\{\,|G_S^{\alpha}(k)|^{2p}\}$ is an $O(N^{-(F_a - n)/2})$. The number of free summation variables in $\mathcal{E}\{\,|G_S^{\alpha}(k)|^{2p}\}$ is given by (3-63). Hence,

$$\mathcal{E}\{\,|G_S^{\alpha}(k)|^{2p}\} = O(N^{(F_a - 2p)/2 - 1})O(N^{-(F_a - n)/2}) = O(N^{-1})$$

since $n = 2p$.

Claim 3: *Partitions that link pairs of rows (k, l), $l \neq -k$ in (3-50) give only $O(N^{-\nu})$, $\nu \geq 1$ contributions.*

Proof. For such a partition, keeping in mind that each row should contain at least two entries different from zero, B should contain at least the following submatrix $\tilde{B}$:

$$\begin{array}{c} k \\ l \end{array} \begin{bmatrix} 1 & 1 \\ -1 & -1 \end{bmatrix} \quad \text{or} \quad \begin{array}{c} k \\ l \end{array} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \text{ or similar} \tag{3-64}$$

Because the rank of $\tilde{B}$ is 1, and the rank of the augmented matrix

$$\begin{bmatrix} k & \tilde{B} \\ l & \end{bmatrix} \tag{3-65}$$

is 2, this set has no solution. Hence, at least an additional link with another row is needed to increase the rank of $\tilde{B}$ to 2. Claim 3 then follows from the previous Claim 2.

Note that pairing (k, k) is a special case of this claim.

Claim 4: *Partitions that contain rows that are not linked to another row do not exist.*

Proof. Because each row corresponds to a stochastic contribution, it is clear that not all the frequencies in one row can be paired within this row.

From Claims 1 to 4, it follows that the only contributions of $O(N^0)$ to $\mathcal{E}\{\,|G_S^{\alpha}(k)|^{2p}\}$ are those where the partitions link all the rows per two with the denominator frequencies of the form $(k, -k)$. $\mathcal{E}\{\,|G_S^{\alpha}(k)|^{2p}\}$ is given, within an $O(N^{-1})$, by the sum of all these contributions

$$\mathcal{E}\{\,\left|G_S^{\alpha}(k)\right|^{2p}\} = \sum_{\text{all distinct combinations of pairs}} \mathcal{E}\{\,G_S^{\alpha}(k)G_S^{\alpha}(-k)\}\ldots\mathcal{E}\{\,G_S^{\alpha}(k)G_S^{\alpha}(-k)\} \qquad \text{(3-66)}$$

In this expression "all distinct combinations of pairs" indicates all permutations that can be formed over the rows (3-60) such that distinct products of pairs $(k, -k)$ are formed. For example, if we have four rows (1,2,3,4) with frequencies $k, -k, k, -k,$ we should consider $(1, 2)(3, 4)$; $(1, 4)(2, 3)$. The combination $(1, 3)(2, 4)$ forms pairs (k, k) and does not contribute. From Picinbono (1993, p. 112, Eq. (4.95)) it follows that this corresponds to the moments of a circular, normal distribution. As convergence in the moments implies convergence in distribution (see Lemma 14.11), it follows that $G_S^{\alpha}(k)$ is normally distributed, which proves the theorem.

Appendix 3.F: Extension to Random Amplitudes and Nonuniform Phases

Because the random amplitude has uniformly bounded moments of any order and is independently distributed of the phase, we can calculate the expected value w.r.t. the phase, independently of the amplitude distribution. Hence, all previous proofs in Appendix 3.A to 3.E remain valid for random amplitudes.

The basic reason that a discrete phase or nonuniform continuous distribution needs special attention is that $\mathcal{E}\{\,U_k U_k\}$ can be different from zero, e.g., $\varphi_k \in \{0, \pi\}$. However, a careful check shows that all previous proofs in Appendix 3.A to 3.E remain valid if (l, l) is also considered as a paired frequency. Notice that for such a pair the sum of the frequencies is no longer zero (no major impact on the proofs). A second difference is the fact that such a pair is represented by one element in the A and B matrices, but notice that there are still two frequencies linked to this single element. The $\varphi_k \in \{0, \pi\}$ distribution is a worst case. Discrete distributions with more elements link more frequencies to generate a nonzero expected value.

The major difference is the expected value $\mathcal{E}\{\,G^{\alpha}\}$. Additional $O(N^{-1})$ terms appear, also for the even nonlinearities.

A typical odd degree bias contribution for a discrete phase distribution $\varphi \in \{0, \pi\}$ would be $-k \qquad l_1, l_1, \ldots, l_e, l_e, m_1, -m_1, \ldots, m_o, -m_o, k$. It is important to notice that $\sum_{i=1}^{e} l_i = 0$ in order to meet the frequency constraint and, hence, an additional frequency constraint becomes active. Using arguments similar to those in the previous appendices, the sum of all these contributions is an $O(N^{-1})$.

An example of an even degree systematic for a nonuniform phase contribution is $-k \qquad l_1, l_1, l_1, l_2, l_2, k$ with $3l_1 + 2l_2$. Note that in this case at least three frequencies are linked in one "pair" so that an $O(N^{-3/2})$ results.

Appendix 3.G: Response of a Nonlinear System to a Gaussian Excitation

For noise excitations, the FRF is measured using the H_1 method (2-43), and its limit value is given by

$$\hat{G}^{\alpha}(j\omega) = \frac{\mathcal{E}\{\,Y^{\alpha}(j\omega)\bar{U}(j\omega)\}}{\mathcal{E}\{\,U(j\omega)\bar{U}(j\omega)\}} = \frac{S_{YU}(j\omega)}{S_{UU}(j\omega)} \qquad \text{(3-67)}$$

The cross-spectrum $S_{YU}(j\omega)$ is the Fourier transform of the cross-correlation $R_{yu}(\tau)$ between the input and the output and depends on higher order spectra. In the case of zero mean normal distributed noise, these higher order spectra can easily be calculated. Consider the contribution of degree α:

$$\begin{aligned} y^{\alpha}(t) &= \int_{-\infty}^{+\infty}\dots\int_{-\infty}^{+\infty} g_{\alpha}(\tau_0, \dots, \tau_{\alpha-1})u(t-\tau_0)\dots u(t-\tau_{\alpha-1})d\tau_1\dots d\tau_{\alpha} \\ R_{y^{\alpha}u}(\tau_0) &= \mathcal{E}\{\, y^{\alpha}(t)u(t-\tau_0)\} \\ &= \int_{-\infty}^{+\infty}\dots\int_{-\infty}^{+\infty} g_{\alpha}(\tau_1, \dots, \tau_{\alpha})\,\mathcal{E}\{\, u(t-\tau_0)u(t-\tau_1)\dots u(t-\tau_{\alpha})\}d\tau_1\dots d\tau_{\alpha} \end{aligned} \tag{3-68}$$

For zero mean jointly normally distributed noise, the higher order moments are given by (Schetzen, 1980, p. 218):

$$\mathcal{E}\{\, u_1 u_2 \dots u_M\} = \begin{cases} 0 & \text{if M is odd} \\ \Sigma\pi\,\mathcal{E}\{\, u_i u_j\} & \text{if M is even} \end{cases} \tag{3-69}$$

The $\Sigma\pi$ stands for the summation over all distinct ways of partitioning the M random variables into products of averages of pairs. It is shown that there are $(M-1)!!$ such combination for M even (Schetzen, 1980) and zero if M is odd. Hence, $R_{y^{2\alpha}u}(\tau_0) = 0$ and

$$G_B^{2\alpha}(j\omega) = \frac{S_{Y^{2\alpha}U}(j\omega)}{S_{UU}(j\omega)} = \frac{F\{R_{y^{2\alpha}u}(\tau_0)\}}{S_{UU}(j\omega)} = 0$$

From here on, it is assumed that α is odd so that an even number of input terms appear.

Using Eq. (3-69), the expected value in Eq. (3-68) becomes

$$\mathcal{E}\{\, u(t-\tau_0)u(t-\tau_1)\dots u(t-\tau_{\alpha})\}=\Sigma\pi R_{uu}(\tau_i - \tau_j) \tag{3-70}$$

Using the relationship between the autocorrelation and the power spectrum of the input,

$$R_{uu}(\tau) = \int_{-\infty}^{+\infty} S_{UU}(j\omega)e^{j\omega\tau}df \tag{3-71}$$

Eq. (3-68) can be rewritten as

$$\begin{aligned} &R_{y^{2\alpha-1}u}(\tau_0) = \\ &\int_{-\infty}^{+\infty}\dots\int_{-\infty}^{+\infty} g_{2\alpha-1}(\tau_1, \dots, \tau_{2\alpha-1})\Sigma\pi S_{UU}(j\omega_r)e^{j\omega_r(\tau_i-\tau_j)}d\tau_1\dots d\tau_{2\alpha-1}df_1\dots df_{\alpha} \end{aligned} \tag{3-72}$$

In order to calculate this expression, the contribution of one term of $\Sigma\pi$ is analyzed in detail for the partition $(\tau_0, \tau_1), (\tau_2, \tau_3), \dots, (\tau_{2\alpha-2}, \tau_{2\alpha-1})$:

$$\int_{-\infty}^{+\infty}\dots\int_{-\infty}^{+\infty} g_{2\alpha-1}(\tau_1, \dots, \tau_{2\alpha-1})\prod_{r=1}^{\alpha} S_{UU}(j\omega_r)e^{j\omega_r(\tau_{2r-2}-\tau_{2r-1})}d\tau_1\dots d\tau_{2\alpha-1}df_1\dots df_{\alpha}$$

Define

$$G^{2\alpha-1}_{f_1,-f_2,f_2,\ldots,-f_\alpha,f_\alpha} =$$

$$\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty} g_{2\alpha-1}(\tau_1,\ldots,\tau_{2\alpha-1})e^{-j\omega_1\tau_1}\prod_{r=2}^{\alpha} e^{j\omega_r(\tau_{2r-2}-\tau_{2r-1})}d\tau_1\ldots d\tau_{2\alpha-1}$$

Because $G^{2\alpha-1}_{f_1,-f_2,f_2,\ldots,-f_\alpha,f_\alpha}$ is a symmetrical kernel, it does not depend on the order of its arguments. So, all possible terms in the partitioning give the same result, thus Eq. (3-72) becomes

$$R_{y^{2\alpha-1}u}(\tau_0) =$$

$$(2\alpha-1)!!\int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty} G^{2\alpha-1}_{f_1,-f_2,f_2,\ldots,-f_\alpha,f_\alpha}\prod_{r=2}^{\alpha} S_{UU}(j\omega_r)S_{UU}(j\omega_1)e^{j\omega_1\tau_0}df_1\ldots df_\alpha \tag{3-73}$$

Note that the power spectrum of $R_{y^{2\alpha-1}u}(\tau_0)$ is given by

$$S_{Y^{2\alpha-1}U}(j\omega) = \int_{-\infty}^{\infty} R_{y^{2\alpha-1}u}(\tau_0)e^{-j\omega\tau_0}d\tau_0 \tag{3-74}$$

Applying (3-74) to (3-73) results in

$$S_{Y^{2\alpha-1}U}(j\omega) =$$

$$S_{UU}(j\omega)(2\alpha-1)!!\int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty} G^{2\alpha-1}_{f,-f_2,f_2,\ldots,-f_\alpha,f_\alpha}S_{UU}(j\omega_2)\ldots S_{UU}(j\omega_\alpha)df_2\ldots df_\alpha$$

Dividing $S_{Y^{2\alpha-1}U}(j\omega)$ by $S_{UU}(j\omega)$ gives $G^{2\alpha-1}_B(j\omega)$:

$$\begin{aligned}
G^{2\alpha-1}_B(j\omega) &= \frac{S_{Y^{2\alpha-1}U}(j\omega)}{S_{UU}(j\omega)} \\
&= (2\alpha-1)!!\int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty} G^{2\alpha-1}_{f,-f_2,f_2,\ldots,-f_\alpha,f_\alpha}S_{UU}(j\omega_2)\ldots S_{UU}(j\omega_\alpha)df_2\ldots df_\alpha \\
&= (2\alpha-1)!!2^{\alpha-1}\int_0^{\infty}\ldots\int_0^{\infty} G^{2\alpha-1}_{f,-f_2,f_2,\ldots,-f_\alpha,f_\alpha}S_{UU}(j\omega_2)\ldots S_{UU}(j\omega_\alpha)df_2\ldots df_\alpha
\end{aligned}$$

where in the last step the double-sided spectra are replaced by single sided spectra.

Appendix 3.H: Proof of Theorem 3.12

Note: In this appendix, we denote explicitly the dependence of the results on the number of frequencies N by adding a subscript N.

We elaborate the first term in the right-hand side of (3-13):

$$c_\alpha \sum_{k_1,\ldots,k_{\alpha-1}=1}^{N} G^{2\alpha-1}_{k,-k_1,k_1,\ldots,-k_{\alpha-1},k_{\alpha-1}}\mathcal{E}\{|U_{k_1}|^2\ldots|U_{k_{\alpha-1}}|^2\} \tag{3-75}$$

Splitting the sums in (3-75) as $\sum_{k_i} = \sum_{\text{all } k_i \text{ different}} + \sum_{\text{not all } k_i \text{ different}}$ and using

$$\mathcal{E}\{|U_{k_1}|^2 \dots |U_{k_{\alpha-1}}|^2\} = \prod\nolimits_{i=1}^{\alpha-1} \mathcal{E}\{|U_{k_i}|^2\} \text{ for all } k_i\text{'s different} \tag{3-76}$$

makes it possible to rewrite Eq. (3-75) as

$$\begin{aligned} c_\alpha\Bigg(&\sum_{\text{all } k_i \text{ different}} G^{2\alpha-1}_{k,-k_1,k_1,\dots,-k_{\alpha-1},k_{\alpha-1}} \prod\nolimits_{i=1}^{\alpha-1} |U_{k_i}|^2 \\ + &\sum_{\text{not all } k_i \text{ different}} G^{2\alpha-1}_{k,-k_1,k_1,\dots,-k_{\alpha-1},k_{\alpha-1}} \mathcal{E}\{|U_{k_1}|^2 \dots |U_{k_{\alpha-1}}|^2\}\Bigg) \end{aligned} \tag{3-77}$$

Adding and subtracting $\sum_{\text{not all } k_i \text{ different}}$ in the first summation of (3-77) gives, using $|U_{k_i}|^2 = S_{\hat{U}\hat{U}}(f_{k_i})/N$,

$$C^\alpha_{1N} + C^\alpha_{2N} \tag{3-78}$$

where

$$\begin{aligned} C^\alpha_{1N} &= \frac{c_\alpha}{N^{\alpha-1}} \sum_{k_1,\dots,k_{\alpha-1}=1}^{N} G^{2\alpha-1}_{k,-k_1,k_1,\dots,-k_{\alpha-1},k_{\alpha-1}} \prod\nolimits_{i=1}^{\alpha-1} S_{\hat{U}\hat{U}}(f_{k_i}) \\ C^\alpha_{2N} &= c_\alpha \sum_{\text{not all } k_i \text{ different}} G^{2\alpha-1}_{k,-k_1,k_1,\dots,-k_{\alpha-1},k_{\alpha-1}} \Delta_{\alpha-1} \\ \Delta_{\alpha-1} &= \mathcal{E}\{|U_{k_1}|^2 \dots |U_{k_{\alpha-1}}|^2\} - \prod\nolimits_{i=1}^{\alpha-1} |U_{k_i}|^2 \end{aligned} \tag{3-79}$$

Because $S_{\hat{U}\hat{U}}(f_{k_i})$ and $|G^{2\alpha-1}_{k,-k_1,k_1,\dots,-k_{\alpha-1},k_{\alpha-1}}|$ are uniformly bounded (see Definitions 3.2 to 3.4), $|U_{k_i}|^2 = O(N^{-1})$ (see Definitions 3.2 and 3.3), and the sum $\sum_{\text{not all } k_i \text{ different}}$ contains at most $\alpha - 2$ independent k_i's, we find

$$|C^\alpha_{1N}| \le \frac{c_\alpha}{N^{\alpha-1}} O(N^{\alpha-1}) = O(N^0) \quad \text{and} \quad |C^\alpha_{2N}| \le \frac{c_\alpha}{N^{\alpha-1}} O(N^{\alpha-2}) = O(N^{-1}) \tag{3-80}$$

Collecting (3-12), (3-13), (3-78), and (3-80), we get

$$G_{B,N}(s_k) = \sum\nolimits_{\alpha=2}^{\infty} G^{2\alpha-1}_{B,N}(s_k) = \sum\nolimits_{\alpha=2}^{\infty} C^\alpha_{1N} + O(N^{-1}) \tag{3-81}$$

Because $S_{\hat{U}\hat{U}}(f_{k_i})$ is by assumption the same for the three classes of excitation signals, it follows from (3-79) and (3-81) that for these three classes $G_{B,N}(s_k)$ converges ($N \to \infty$) at the rate $O(N^{-1})$ to the same limit value $G_B(s_k)$. Under some additional assumptions on the odd degree kernels it is possible to calculate an explicit expression for $G_B(s_k)$.

Because the joint second-order derivatives of

$$G^{2\alpha-1}_{k,-k_1,k_1,\dots,-k_{\alpha-1},k_{\alpha-1}} = G^{2\alpha-1}_{f_k,-f_{k_1},f_{k_1},\dots,-f_{k_{\alpha-1}},f_{k_{\alpha-1}}}$$

w.r.t. f_{k_1}, f_{k_2}, ... $f_{k_{\alpha-1}}$ and f_k are bounded for $f_{k_1}, f_{k_2}, \dots, f_{k_{\alpha-1}}, f_k \in [0, f_{\max}]$, the Riemann sum

$$C_{1N}^{\alpha} = \frac{c_{\alpha}}{N^{\alpha-1}} \sum_{k_1, \ldots, k_{\alpha-1}=1}^{N} G_{k, -k_1, k_1, \ldots, -k_{\alpha-1}, k_{\alpha-1}}^{2\alpha-1} \prod\nolimits_{i=1}^{\alpha-1} S_{\hat{U}\hat{U}}(f_{k_i}) \tag{3-82}$$

where $f_{k_i} - f_{k_{i-1}} = f_{\max}/N$, converges to C_1^{α}

$$C_1^{\alpha} = \frac{c_{\alpha}}{f_{\max}^{\alpha-1}} \int_0^{f_{\max}} \ldots \int_0^{f_{\max}} G_{f_k, -f_{k_1}, f_{k_1}, \ldots, f_{k_{\alpha-1}}}^{2\alpha-1} S_{\hat{U}\hat{U}}(f_1) \ldots S_{\hat{U}\hat{U}}(f_{\alpha-1}) df_1 \ldots df_{\alpha-1}$$

at the rate $O(N^{-2})$ (Ralston and Rabinowitz, 1984; midpoint rule (4.10-10)). Together with (3-81) it shows that $\lim_{N \to \infty} G_{R,N}(j\omega) = G_R(j\omega)$, with $G_R(j\omega)$ defined in (3-18).

Appendix 3.I: Proof of Theorem 3.15

The sum of a uniformly convergent series of continuous functions is continuous (see Kaplan, 1993, Theorem 31). Hence, under Assumptions 3.13 and 3.14, the sum $G_B(j\omega) = \sum_{\alpha=2}^{\infty} C_1^{\alpha}(j\omega)$, and its derivatives of order $1, 2, \ldots, P$ w.r.t. ω, are continuous functions of $\omega \in [0, \omega_{\max}]$.

Appendix 3.J: Proof of Theorem 3.16

Note: In this appendix, we denote explicitly the dependence of the results on the number of frequencies N by adding a subscript N.

From (3-9), (3-10), and (3-22), it follows that the stochastic nonlinear contributions $Y_{Sk,N}$ are given by

$$\begin{aligned} \sqrt{N} Y_{Sk,N} &= \sum\nolimits_{\alpha=2}^{\infty} \sqrt{N} Y_{Sk,N}^{\alpha} \\ \sqrt{N} Y_{Sk,N}^{\alpha} &= \sqrt{N} \sum_{k_1, \ldots, k_{\alpha-1}=-N}^{N} G_{k_1, \ldots, k_{\alpha}}^{\alpha} U_{k_1} \ldots U_{k_{\alpha}} \end{aligned} \tag{3-83}$$

with constraints

$$k = \sum\nolimits_{i=1}^{\alpha} k_i, \ \sum\nolimits_{i=1}^{\alpha} \varphi_{k_i} \neq \varphi_k, \ k \neq 0, \text{ and } k_i \neq 0 \text{ for } i = 1, \ldots, \alpha \tag{3-84}$$

and where $G_{k_1, \ldots, k_{\alpha}}^{\alpha} = G_{f_{k_1}, \ldots, f_{k_{\alpha}}}^{\alpha}$ with $f_{k_i} = k_i f_{\max}/N$. The variance of $\sqrt{N} Y_{Sk,N}$ equals

$$\operatorname{var}(\sqrt{N} Y_{Sk,N}) = N \sum_{\alpha,\beta=2}^{\infty} \mathcal{E}\{ Y_{Sk,N}^{\alpha} \bar{Y}_{Sk,N}^{\beta} \} = \sum_{\alpha,\beta=2}^{\infty} C_N^{\alpha,\beta} \tag{3-85}$$

with

$$C_N^{\alpha,\beta} = N \sum_{\substack{k_1, \ldots, k_{\alpha-1} = -N \\ l_1, \ldots, l_{\beta-1} = -N}}^{N} G_{k_1, \ldots, k_{\alpha}}^{\alpha} \bar{G}_{l_1, \ldots, l_{\beta}}^{\beta} \mathcal{E}\{ U_{k_1} \ldots U_{k_{\alpha}} \bar{U}_{l_1} \ldots \bar{U}_{l_{\beta}} \} \tag{3-86}$$

Because $U_k = N^{-1/2}\hat{U}(f_k)e^{j\varphi_k}$ with $f_k = kf_{\max}/N$, $\mathcal{E}\{e^{j\varphi_k}\} = 0$ and φ_k independent of $\hat{U}(f_k)$, it follows that

$$\mathcal{E}\{U_{k_1}\ldots U_{k_\alpha}\bar{U}_{l_1}\ldots\bar{U}_{l_\beta}\} \neq 0 \Leftrightarrow \textstyle\sum_{i=1}^{\alpha}\varphi_{k_i} = \sum_{i=1}^{\beta}\varphi_{l_i} \tag{3-87}$$

Taking into account the constraints (3-84), the phase condition in (3-87) can be met only if the frequencies are paired as $(m_j, -m_j)$ with $m_j \in \{k_1, \ldots, k_\alpha, -l_1, \ldots, -l_\beta\}$ and where not all m_j should be different. The maximum number of terms in the sums (3-86) is obtained by maximizing the number of independent m_j's (number of independent pairs). Because $|U_{k_i}|^2 = S_{\hat{U}\hat{U}}(f_{k_i})/N$ and the maximum number of independent pairs equals $\gamma = (\alpha+\beta)/2 - 1$, (3-86) can be written as

$$C_N^{\alpha,\beta} = N^{-\gamma}\sum_{m_1,\ldots,m_\gamma=-N}^{N}\left(\sum_{k_i,l_i}G^{\alpha}_{k_1,\ldots,k_\alpha}\bar{G}^{\beta}_{l_1,\ldots,l_\beta}\right)S_{\hat{U}\hat{U}}(f_{m_1})\ldots S_{\hat{U}\hat{U}}(f_{m_\gamma}) + O(N^{-1}) \tag{3-88}$$

where the sum $\sum_{k_j,l_i}$ extends over the choices of $m_j \in \{k_1, \ldots, k_\alpha, -l_1, \ldots, -l_\beta\}$ resulting in γ independent pairs $(m_j, -m_j)$, and where the second term stems from the nonzero contributions in (3-86) containing at most $\gamma - 1$ independent m_j's. Note that the first term in (3-88) is an $O(N^0)$ and that it is the same for random phase multisines, periodic noise, and Gaussian noise with the same (power) spectra $S_{\hat{U}\hat{U}}(f)$. Collecting (3-85) and (3-88) gives

$$\operatorname{var}(\sqrt{N}Y_{Sk,N}) = \sigma^2_{S,N}(k) + O(N^{-1}) \tag{3-89}$$

where $\sigma^2_{S,N}(k) = O(N^0)$ is the same for the three classes of excitation signals. Taking the limit $N \to \infty$ of (3-89) proves the theorem with $\sigma_S^2(f) = \lim_{N\to\infty}\sigma^2_{S,N}(k)$.

Appendix 3.K: Proof of Theorem 3.8

Consider the contributions to Y_L^{α}, $|L| > l_{\max}$ (see Eq. 3-5). These are of the form

$$G^{\alpha}_{L,k_1,k_2,\ldots,k_{\alpha-1}}U_{k_1}U_{k_2}\ldots U_{k_{\alpha-1}}U_{k_\alpha} \text{ with } k_\alpha = L - \textstyle\sum_{i=1}^{\alpha-1}k_i \tag{3-90}$$

To get systematic contributions, α should be odd because even nonlinearities cannot create systematic contributions.

Assume that $\exists\ l$ s.t. the phase of $U_{-l}(U_{k_1}U_{k_2}\ldots U_{k_{\alpha-1}}U_{k_\alpha})$ is zero (these combinations create the systematic contributions). We will check whether such combinations can exist.

The only possibility to get zero phase is that U_{-l} is paired with one of the components U_{k_i}. In that case there exists a k_i s.t. the phase $U_{-l}U_{k_i}$ is zero. After rearranging the order of the components, we can put U_{k_i} in the last place. Also the components that pair are put together, and eventually the contributions can be written as

$$U_{r_1}U_{r_2}\ldots U_{r_{2\beta}}(U_{s_1}U_{-s_1})\ldots(U_{s_p}U_{s_{-p}})(U_{k_\alpha}U_{-l}) \tag{3-91}$$

with U_{r_i} the unpaired components. Now there are two possibilities:

(i) There are no unpaired components left, $\beta = 0$. In that case, the combination in (3-90) contributes to the frequency $L = l = k_\alpha$ which is by definition in the excitation band (k_α is an excitation frequency). This violates that $|L| > l_{\max}$.

(ii) There are unpaired components ($\beta \neq 0$). In that case not all frequencies in (3-90) are paired, and, hence, the phase is not zero. So, also this situation cannot result in systematic contributions.

This proves the theorem. □

4

Design of Excitation Signals

Abstract: Good experiments are the best guarantee to build good models. The selection of good excitation signals is an important step in the design of the experiment. In this chapter we explain how to get such signals. In the first part, three classes of excitations are considered:

General purpose signals that can be applied without any optimization. The only parameters to be selected are the bandwidth of the excitation signal and the frequency resolution of the measurement.

Optimized test signals: these facilitate excitation of the system with a user-specified power spectrum, for example, a semilogarithmic distributed spectrum.

Dedicated test signals: these are signals with optimized characteristics for special situations; for example, the signal and its derivative do not exceed a user-specified peak value.

The second part of this chapter deals briefly with the design of optimal power spectra so that the available power is used at the frequencies where it contributes most to the knowledge of the system.

4.1 INTRODUCTION

In most system-analysis applications, the dynamic behavior of the system is derived from measurements of the input and output signals. In some situations the input signal is imposed by the environment and it is impossible to excite the device under test with an arbitrarily chosen input (for example, in biological systems, where the choice of excitation is very limited). In other situations, only binary signals may be applicable. However, in a wide variety of cases, the only restriction on input signals is that of a limitation in the permitted amplitude range.

A very common method used in transfer function measurements, until the end of the 1960s, was that of the combination of a slowly swept sine with a tracking filter. Since the development of advanced digital signal processing algorithms, and especially since the efficient

System Identification: A Frequency Domain Approach. Rik Pintelon and Johan Schoukens
Copyright © 2001 Institute of Electrical and Electronics Engineers. ISBN: 0-780-36000-1

implementation of the discrete Fourier transform (DFT) with the fast Fourier transform (FFT), it became possible to use more complex input signals. Instead of exciting the unknown system frequency by frequency, sophisticated waveforms with a broadband spectrum are generated, enabling collection of all the required spectral information from a single measurement. This can result in a considerable reduction of the measurement time but also in an undesired loss of accuracy if no precautions are taken. We will analyze the trade-off between accuracy and measurement time, but before starting we must choose between a nonparametric and a parametric modeling approach. In the nonparametric representation the system is characterized by measurements of the frequency response at a large number of frequencies, whereas in a parametric model the system is described by a mathematical transfer function model with a limited number of parameters. It is precisely these parameters that have to be estimated in the parametric modeling approach. The optimum spectrum of the excitation in the parametric case will be different from that in the nonparametric case: this is principally because the parametric model combines the information available from all frequencies in only a few parameters. In a direct nonparametric frequency response measurement, there is no relation between the measurements at the various frequencies and the excitation should be designed to achieve a predefined accuracy in the frequency bands of interest: for example, maximizing the absolute or relative accuracy of the measurements. In a parametric approach, the energy will be concentrated at the frequencies where it contributes most to the knowledge about the model parameters.

To design an optimized excitation signal, it is necessary to specify the final goal. For the nonparametric case, we will look for signals that *maximize the minimum accuracy obtained in a fixed measurement time for a specified maximum peak value of the excitation:*

$$\min(\max_{k \in \mathbb{F}} \sigma_G^2(k)) \text{ with } \max_t |u(t)| \le u_{\max} \tag{4-1}$$

where $\mathbb{F}$ is the set of frequencies at which the frequency response is measured. In the parametric case, the determinant of the information matrix will be maximized, as discussed later.

In this text we first focus our attention on the design of excitation signals for nonparametric measurements. The parametric modeling approach will be studied in the second part.

4.2 GENERAL REMARKS ON EXCITATION SIGNALS FOR NONPARAMETRIC FREQUENCY RESPONSE MEASUREMENTS

In this part the nonparametric frequency response function (FRF) measurement problem is studied. It should be clear to the reader that signals that provide good FRF measurements are also very well suited for use in a parametric identification step, which gives this section more general value.

Before starting a detailed comparison of some candidate excitation signals, we first introduce two quality measures for excitation signals. In general, these measures depend on the actual measured FRF and on the properties of the disturbing noise (e.g., its power spectrum). However, in order to simplify the discussion, we assume that we deal with flat systems (the amplitude of the FRF is a constant) in the presence of white noise. If necessary, we will indicate how the conclusions should be modified to the general situation of arbitrary systems with colored noise distortions.

4.2.1 Quantifying the Quality of an Excitation Signal

In Chapter 2 it was shown that the uncertainty on the FRF at ω_k after M averages is

$$\sigma_{\hat{G}}^2(k) \approx \frac{|G_0(j\omega_k)|^2}{M}\left(\frac{\sigma_{\hat{Y}}^2(k)}{S_{Y_0Y_0}(k)} + \frac{\sigma_{\hat{U}}^2(k)}{S_{U_0U_0}(k)} - 2\mathrm{Re}\left(\frac{\sigma_{\hat{Y}U}^2(k)}{S_{Y_0U_0}(k)}\right)\right) \qquad (4\text{-}2)$$

The uncertainty is inversely proportional to the total power of the excitation signal and also to the shape of its power spectrum. In order to have a constant variance $\sigma_{\hat{G}}^2(k)$ at all frequencies, the power distribution should be proportional to the impact of the disturbing noise. This leads to the definition of two characteristics for excitation signals: the crest factor and the time factor.

Definition 4.1 (Crest Factor): The crest factor $Cr(u)$ of a signal $u(t)$ is given by the ratio of the peak value u_{peak} of the signal to its rms value u_{RMSe} in the frequency band of interest

$$Cr(u) = \frac{u_{\mathrm{peak}}}{u_{\mathrm{RMSe}}} = \frac{\max\limits_{t \in [0,\,T]} |u(t)|}{u_{\mathrm{RMS}}\sqrt{P_{\mathrm{i}}/P_{\mathrm{T}}}} \quad \text{with } u_{\mathrm{RMS}}^2 = \frac{1}{T}\int_0^T u^2(t)dt$$

with T the measurement time, u_{RMS} the rms value of the signal, P_{T} the total power of the signal, and P_{i} the power in the frequency band of interest.

The crest factor gives an idea of the compactness of the signal. Signals with an impulsive behavior (having a large crest factor) inject much less power into the system than signals having the same peak value and a small crest factor. The effective rms value u_{RMSe} is used to express that only the power in the frequency band of interest contributes to the knowledge of the system.

The time factor of an excitation signal also accounts for the power distribution of the signal over the frequencies. If this is unequally distributed with respect to the noise, some FRF points will be poorly measured. We will require that the worst measurements still reach a minimum quality. For the sake of general conclusions, we make the following simplifying assumptions: $|G_0(j\omega_k)|^2$, $\sigma_{\hat{U}}^2(k)$, $\sigma_{\hat{Y}}^2(k)$, $\sigma_{\hat{Y}U}^2(k)$ are constant. Expression (4-1) reduces to

$$\sigma_{\hat{G}}^2(k) \sim \frac{1}{M|U(k)|^2} \qquad (4\text{-}3)$$

and the number of averages to reach a specified variance is proportional to $M \sim 1/|U(k)|^2$. The total measurement time T is proportional to the required number of averages M. Also notice that $Cr^2(u) = u_{\mathrm{peak}}^2/u_{\mathrm{RMSe}}^2$ and $u_{\mathrm{RMSe}}^2 = 2FU_{\mathrm{RMSe}}^2$ with F the number of frequencies in the frequency band of interest and, $U_{\mathrm{RMSe}}^2 = \sum_{k=1}^{F} |U(k)|^2/F)$. Then

$$T \sim \max_k \frac{1}{|U(k)|^2} \sim \max_k \frac{U_{\mathrm{RMSe}}^2}{|U(k)|^2 U_{\mathrm{RMSe}}^2} \sim \max_k \frac{Cr^2(u)}{\frac{|U(k)|^2}{U_{\mathrm{RMSe}}^2}\frac{u_{\mathrm{peak}}^2}{F}} \qquad (4\text{-}4)$$

and the required measurement time per frequency line for a specified peak value u_{peak} becomes proportional to

$$\frac{T}{F} \equiv Tf(u) \sim \max_k \frac{Cr^2(u)U_{\text{RMSe}}^2}{|U(k)|^2} \tag{4-5}$$

The proportionality factor is fixed by normalizing $Tf(u) \equiv 1$ for a sine wave. Thus, the time factor indicates the required measurement time per frequency point that is needed to guarantee a minimum SNR on the FRF measurement, and this is compared with a stepped sine excitation.

Definition 4.2 (Time Factor): The time factor $Tf(u)$ of a signal $u(t)$ is given by

$$Tf(u) = \max_{k \in \mathbb{F}} 0.5Cr^2(u)U_{\text{RMSe}}^2/|U(k)|^2$$

This result can be generalized for situations with frequency-dependent noise levels and varying transfer functions. However, it is still impossible to make general comparisons on the excitation signals. The ability of the excitation signals to deal with these situations depends on their flexibility to impose a user-specified power spectrum.

4.2.2 Stepped Sine versus Broadband Excitations

In this chapter we use the time factor of the sine as a reference to qualify the broadband excitations. However, the reader should be aware that this quality measure deals only with the SNR properties of the signal. In practice, other aspects also influence the total measurement time. To make this clear, the measurement time of a stepped sine experiment (consisting of a series of single sine measurements at the desired frequencies) is compared with the measurement time with a broadband excitation having the same time factor. Two extreme situations are considered, assuming a very good SNR the first time and a very poor SNR the second time. Finally, the intermediate situation is analyzed.

4.2.2.1 Very Good SNR. For the stepped sine, the measurement time is determined by two elements. At least one period of the sine should be measured and, after each frequency step, a waiting time $T_w(k)$ should be included, allowing the transients (of the plant and the measurement system) to disappear. For highly damped systems, these transients are short, but they are very long for lightly damped systems, as they appear, for example, in many mechanical applications. For simplicity, we assume that $T_w(k)$ is a frequency-independent constant. Under these conditions, the total measurement time is $T_{ss} = \sum_{k=1}^{F} 1/f_k + FT_w$ for the stepped sine and $T_{bs} = 1/\Delta f + T_w$ for the broadband measurement, where Δf is the frequency resolution (one period of the broadband measurement equals $1/\Delta f$). If $f_k = kf_0$ and $\Delta f = f_0$, these expressions become

$$T_{ss} = \frac{1}{f_0}\sum_{k=1}^{F} 1/k + FT_w \approx \frac{1}{f_0}(0.58 + \ln F) + FT_w \quad \text{and} \quad T_{bs} = \frac{1}{f_0} + T_w$$

This shows that a significant gain in measurement time is obtained using the broadband excitation.

4.2.2.2 Very Poor SNR. When the SNR is poor, the measurement time needed to get an acceptable uncertainty is much larger than the waiting time T_w, and it is proportional to $\max_k 1/|U(k)|^2$ (if we assume for simplicity a constant noise level on the measurements). A broadband signal distributes the power over F frequencies, while a stepped sine measurement keeps all power focused on one line at each partial measurement: $|U_{ss}(k)|^2 = F|U_{bs}(k)|^2$. Hence, to get the same SNR, the measurement time at one frequency will be F times smaller for the single sine measurement compared with the broadband excitation measurement. However, for a single sine measurement, F measurements should be made, one after another, while all information is captured at once for the broadband measurement, so that, eventually, the total measurement time is the same.

4.2.2.3 Intermediate Situation: Balancing the Transient Errors versus the Noise Errors. In general, the user faces a tricky situation with measurements of medium quality (for example, an SNR of 40 dB). In that case, (4-6) gives a rough rule of thumb for estimating the required waiting time so that the equivalent output noise errors equal the transient errors (Schoukens et al., 2000):

$$T_w = \frac{\tau}{2}\ln\left(\frac{\tau}{2T}\mathrm{SNR}^2\right) \text{ with } \mathrm{SNR}^2 = \frac{S_{YY}(j\omega)}{\sigma_Y^2(k) + \sigma_U^2(k)|G(\Omega_k)|^2 - 2\mathrm{Re}(\sigma_{YU}^2(k)\bar{G}(\Omega_k))} \tag{4-6}$$

with τ the dominating time constant of the system in the considered frequency band, T the length of the time record, and SNR expressed as the ratio of the output power to the equivalent output noise. For example, for $T = 10$ s, $\tau = 1$ s, and an SNR of 40 dB ($\mathrm{SNR} = 100$), the waiting time becomes at least 3 s after each frequency change.

Conclusion. The total measurement time required for step sine measurements will always be larger than that of broadband measurements, provided that we can design the latter excitations with a time factor close to 1. As the damping of a system decreases (time constants increase), the SNR where the stepped sine becomes competitive increases. For most practical situations, the broadband measurement results in a significantly reduced measurement time. For this reason, we focus completely on broadband excitations.

4.3 STUDY OF BROADBAND EXCITATION SIGNALS

The excitation signals are split into three classes: general purpose signals (no optimization involved), optimized test signals (passing through a fully automatic optimization procedure), and, finally, advanced test signals that have some very dedicated properties to deal with specific situations, for example, optimizing not only the signal but also its first and second derivative (such as displacement, velocity, and acceleration).

4.3.1 General Purpose Signals

In this section we study and compare the properties of some general purpose excitation signals. This means that no special optimization is performed to deal with specific situations. These signals should be able to excite the system with an almost flat power spectrum in a user-specified frequency band. From the previous section, we know already that an optimum signal should have a low crest and time factor. Besides these two conditions, it is also important that no leakage appears during the analysis of the measurements, as explained in Section 2.2.3. Therefore, we are strongly in favor of periodic excitations. Leakage errors cannot be avoided if aperiodic signals are used, and it will be necessary to average over a large number of measurements, even if a nonuniform time window is used. This considerably increases the measurement time required for a specified accuracy. Bursts, or time-limited signals, are exceptions to this rule: the continuous spectra of these signals are correctly sampled with the

DFT if the amplitude spectrum is sufficiently band limited for the aliasing effect to be neglected (see Section 2.2.4). Six general purpose signals are considered: swept sine, also called periodic chirp; multisine excitation; maximum length binary sequences; white noise; burst white noise; and pulse testing. At the end of this section, the signals are compared with each other in an example.

4.3.1.1 Swept Sine

Definition 4.3 (Swept Sine): A swept sine (also called periodic chirp) is a sine sweep test, where the frequency is swept up and/or down in one measurement period, and this is repeated in such a way that a periodic signal is created (Brown et al., 1977).

$$u(t) = A\sin((at+b)t) \qquad 0 \le t < T_0 \tag{4-7}$$

with T_0 the period, $a = \pi(k_2-k_1)f_0^2$, $b = 2\pi k_1 f_0^2$, $f_0 = 1/T_0$, $k_2 > k_1 \in \mathbb{N}$, and $k_1 f_0$, $k_2 f_0$ the lowest and the highest frequency, respectively.

Properties

- Periodic signal with period $T_0 = 1/f_0 \quad \rightarrow \quad$ no leakage.
- Frequency resolution $1/T_0$.
- Most of the power is equally distributed in the user-selected frequency band $[k_1, k_2]f_0$ with $k_2 > k_1 \in \mathbb{N}$.
- Crest factor typically 1.45, time factor typically between 1.5 and 4.

Discussion. A swept sine has a low crest factor (comparable to the crest factor of a sine wave) but the amplitude spectrum is not actually flat (see Figure 4-5 on page 125). This introduces frequency components with a lower SNR, resulting in a longer measurement time for a given accuracy. Although a swept sine can create band spectra, it is not possible to generate a signal with an arbitrary amplitude spectrum. A second drawback is that not only are the frequency lines of interest excited, but also a number of other spectral lines appear. This is unimportant with linear systems, but it can be very disturbing in systems with nonlinear behavior.

4.3.1.2 Schroeder Multisine

Definition 4.4 (Schroeder Multisine): A Schroeder multisine is a sum of harmonically related sine waves

$$u(t) = \sum_{k=1}^{F} A\cos(2\pi f_k t + \phi_k)$$

with Schroeder phases $\phi_k = -k(k-1)\pi/F$ and $f_k = l_k f_0$ with $l_k \in \mathbb{N}$.

Properties

- Periodic signal with period $T_0 = 1/f_0 \quad \rightarrow \quad$ no leakage.
- Frequency resolution $1/T_0$.

- All the power at the user-selected frequencies that can be chosen without restriction on the discrete grid kf_0.
- Crest factor typically 1.7, time factor typically 1.5.

Discussion. For the general purpose signal we selected a flat amplitude spectrum $A_k = A$ for the harmonic components of the multisine. However, in general, the user can make an arbitrary choice.

For simplicity, we also used the Schroeder phases (Schroeder, 1970). Although these are not optimal, they give good results for flat amplitude spectra of multisines where a successive set of frequencies is excited. Smaller crest factors can be found by optimizing the phases by using a numerical optimization routine. Dedicated methods are discussed in the next section on optimized test signals, reducing the crest factor from, typically, 1.7 to about 1.4.

Remark. It is strongly advised to use FFT techniques to calculate multisine signals, otherwise the computation time becomes very long (see Exercises 2.1 and 2.2).

4.3.1.3 Pseudorandom Binary Sequence

Definition 4.5 (Pseudorandom Binary Sequence): A pseudorandom binary sequence (PRBS) is a deterministic, periodic sequence of length N that switches between one level (e.g., +1) to another level (e.g., −1). The switches can occur only on a discrete time grid at multiples of the clock period T_c ($k_l T_c$, $k_l \in \mathbb{N}$) and are chosen such that the autocorrelation is as given in Figure 4-1 (Godfrey, 1993a, 1993b).

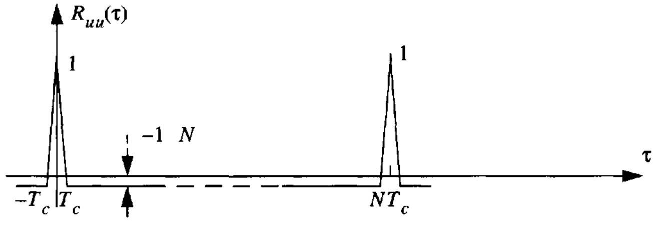


Figure 4-1. Autocorrelation function of a PRBS of length N, switching between ±1.

Properties

- Periodic signal with period $T = NT_c$ → no leakage.
- Frequency resolution $1/T$.
- Most of the power below $0.4f_c = 0.4/T_c$ (see Figure 4-3). Optimal choice of the clock frequency $f_c = 2.5f_{\max}$, with $f_{\max}$ the maximum frequency of interest.
- Crest factor is 1 if all power is considered, time factor typically 1.5.

Discussion. The PRBS has a spectrum whose components decrease in inverse proportion to the frequency. The amplitude $A(k)$ of the Fourier coefficient U_k of a PRBS is given by

$$A(0) = \frac{a}{N} \quad \text{and} \quad A(k) = a\frac{\sqrt{N+1}}{N}\text{sinc}(k\pi/N) \text{ for } k = 1, 2, \ldots, N-1 \qquad (4\text{-}8)$$

with $\text{sinc}(x) = \sin(x)/x$, $2a$ the peak-to-peak amplitude of the sequence, and $f_k = k(f_c/N)$.

It is not possible to find a binary sequence for every arbitrary length N. However, there are a number of possibilities to generate these sequences, hence, there is still much freedom in choosing N.

A first possibility is to use quadratic residue code methods (Godfrey, 1993b). This method generates a PRBS with length $N = 4k - 1$ where N should also be a prime number (e.g., $N = 3, 7, 11, 19, 23, 31, \ldots$). The signals can be generated by the following Matlab™ code:

```
x = –ones(N, 1); x(mod([1:N].^2, N) + 1) = 1; x(1) = 1
```

These sequences can easily be generated, nowadays, using arbitrary waveform generators.

Second possibility: For a long time (in the 1960s and 1970s), it was technically not possible to generate the previous sequences and for that reason other PRBS signals such as the maximum length binary sequences (MLBS) were preferred. These can be generated with the setup shown in Figure 4-2 (Godfrey, 1969, 1980, 1993b; Eykhoff, 1974; Norton, 1986). From all possible binary sequences that can be generated with a fixed register length, the MLBS has the longest period and the shortest correlation length. This means that the spectrum is as flat as possible. The feedback choice determines whether a sequence with the maximum period

$$T_{max} = (2^R - 1)T_c \tag{4-9}$$

is generated. Here, R is the register length and T_c is the clock period.

Because the length N (in clock cycles) does not equal 2^n samples, it is not possible to apply a straightforward FFT analysis. Instead, the chirp-z transform can be used, which permits efficient calculation of the DFT for an arbitrary number of data points (Rabiner and Gold, 1975; Oppenheim and Schafer, 1975). Most numerical packages can calculate the DFT for arbitrary lengths.

In Figure 4-3, details of the first lobe of the amplitude spectrum are given for an MLBS generated with lengths $N = 15, 31,$ and 63. The amplitude of the individual components decreases with increasing length. The crest factor varies as a function of the spectral band$(0 < f \le f_{\max})$ in use, decreasing to 1 as the bandwidth increases to infinity. However, the time factor has a different behavior, as seen in Figure 4-3(b): it decreases for low frequencies but increases to infinity if $f_{\max}$ approaches f_c, as the amplitudes decrease to zero at this frequency. The time factor is less than 1.5 if the upper limit of the frequency band is taken between 0.2 and 0.6 of the PRBS generator clock frequency. The optimal value of the upper fre-

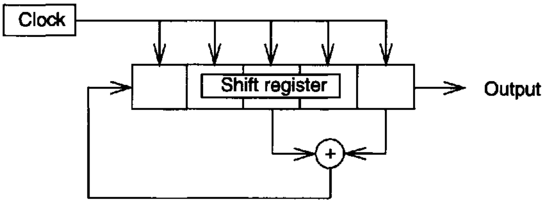


Figure 4-2. Generation of a maximum length binary signal with a shift register (can be initialized with an arbitrary nonzero code).

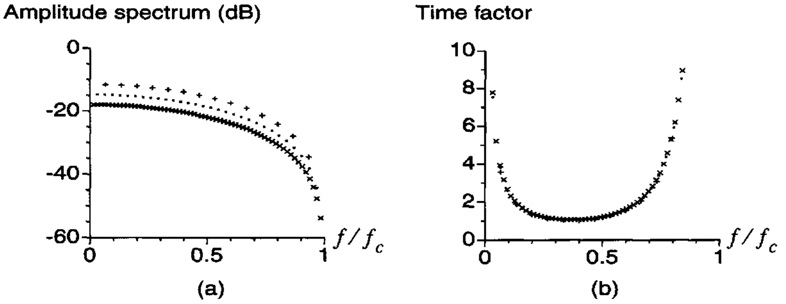


Figure 4-3. (a) Part of the amplitude spectrum and (b) the time factor of an MLBS as function of the bandwidth used ($0 \to f_{\max}$), lengths $N = 15, 31,$ and 63.

quency limit is around $0.4 f_c$, resulting in a time factor of 1.1 corresponding to a clock frequency f_c equal to 2.5 times the maximum frequency of interest.

4.3.1.4 Random Noise

Definition 4.6 (Random Noise): Random noise is a noise sequence whose power spectrum can be influenced by digital filters (Brown et al., 1977; Van Brussel, 1975).

Properties

- Random excitation → leakage problems.
- Equivalent frequency resolution $1/T$.
- Shaping of the power spectrum using a digital filter.
- Crest factor, typically 2-3, and time factor 4.5.

DISCUSSION. In practice, the extreme values of the random signal are clipped (for example, outside the 2 sigma interval) to avoid excessive peak values. The major disadvantages of random excitations are the leakage problems and the drops in the amplitude spectrum if only one realization is processed. In Section 2.6.2, we explain how to deal with these problems.

To maximize the power injected into the system, it is advantageous to use binary noise. This is done by retaining only the sign of the original noise signal (Schoukens et al., 1995). In order to maintain the binary nature, all prefiltering should be done before the sign operation. Because the sign operation is a nonlinear operation, it distorts the power spectrum. Consequently, it is impossible to keep full control over the power spectrum and the crest factor at the same time. This is illustrated in Figure 4-4. A white noise sequence is filtered, and then only the sign is retained so that a binary sequence is generated. The actual, realized spectrum

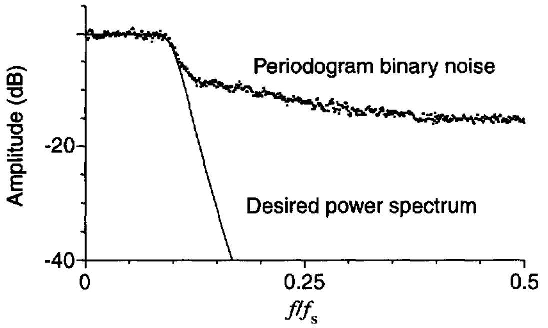


Figure 4-4. Comparison of the spectrum of a filtered noise sequence before and after the sign function.

is compared with the desired power spectrum. As can be seen, the power spectrum is only partly under control. Most of the power is injected in the frequency band of interest, but there is still a large fraction generated outside this band.

4.3.1.5 Random Burst

Definition 4.7 (Random Burst): A random burst is a noise sequence that is imposed on the system during the first part of the measurement sequence, and a zero input is applied for the rest of the measurement period (Herlufsen, 1984).

$$u(t) = w(t)r(t)$$

$$w(t) = \begin{cases} 1 & 0 \le t < T_1 \\ 0 & T_1 \le t < T \end{cases}$$

with $r(t)$ a random variable and $w(t)$ a window function.

Properties

- Random excitation, no leakage if the system response becomes negligible before the end of the measurement window (T).
- Equivalent frequency resolution $1/T$.
- Shaping of the power spectrum using a digital filter.
- Crest factor, typically, $3(T/T_1)^{1/2}$, minimum time factor $\ge 4.5T/T_1$.

DISCUSSION. The crest factor of a random burst sequence is equal to that of the random sequence multiplied by $\sqrt{T/T_1}$. For systems with low damping factors, the relative width T_1/T of the burst must be very small, resulting in a high crest factor. The biggest advantage of using a random burst is that there are no leakage errors (a uniform window should be used to calculate the DFT). The power spectrum of a random burst is a random variable, as it is for a periodic noise sequence, and so the same restrictions are valid as those mentioned for periodic noise.

4.3.1.6 Pulse-Impact Testing

Definition 4.8 (Pulse): The impulse response is measured directly in the time domain by exciting the plant with a short pulse (Halvorsen and Brown, 1977). For example, for a single pulse,

$$u(t) = \begin{cases} A & 0 \le t < T_1 \\ 0 & T_1 < t \le T \end{cases}$$

with T_1 the pulse width and T the measurement period.

Properties

- Deterministic excitation, no leakage if the system response becomes negligible before the end of the measurement window (T).
- Equivalent frequency resolution $1/T$.
- Shaping of the power spectrum by modifying the pulse shape.
- Optimal choice $T_1 = 1/(2.5f_{\max})$.
- Minimum crest factor $\sqrt{T/T_1}$, minimum time factor is T/T_1.

Discussion. The autocorrelation of the impulse response is the same as that of the MLBS, so their amplitude spectra are the same. To get the same input energy, the amplitude must be increased by a factor of $\sqrt{T/T_1}$. The minimum time factor is reached for the same upper frequency limit as for the MLBS. More sophisticated impulse generation techniques are given by Halvorsen and Brown (1977), but the general characteristics remain the same. In mechanical testing, the impulse (or hammer) excitation is still popular because it can be applied very simply: no shakers or other expensive equipment are needed to create the input.

4.3.1.7 Example: Comparison of the General Purpose Excitations. In order to get a better understanding of the behavior of the general purpose signals, they are compared with each other in this section. The aim is to excite a frequency band between 1 and 42 Hz, using signals with a length of 256 samples and a sampling frequency of 256 Hz. The resulting signals and their amplitude spectrum are shown in Figure 4-5.

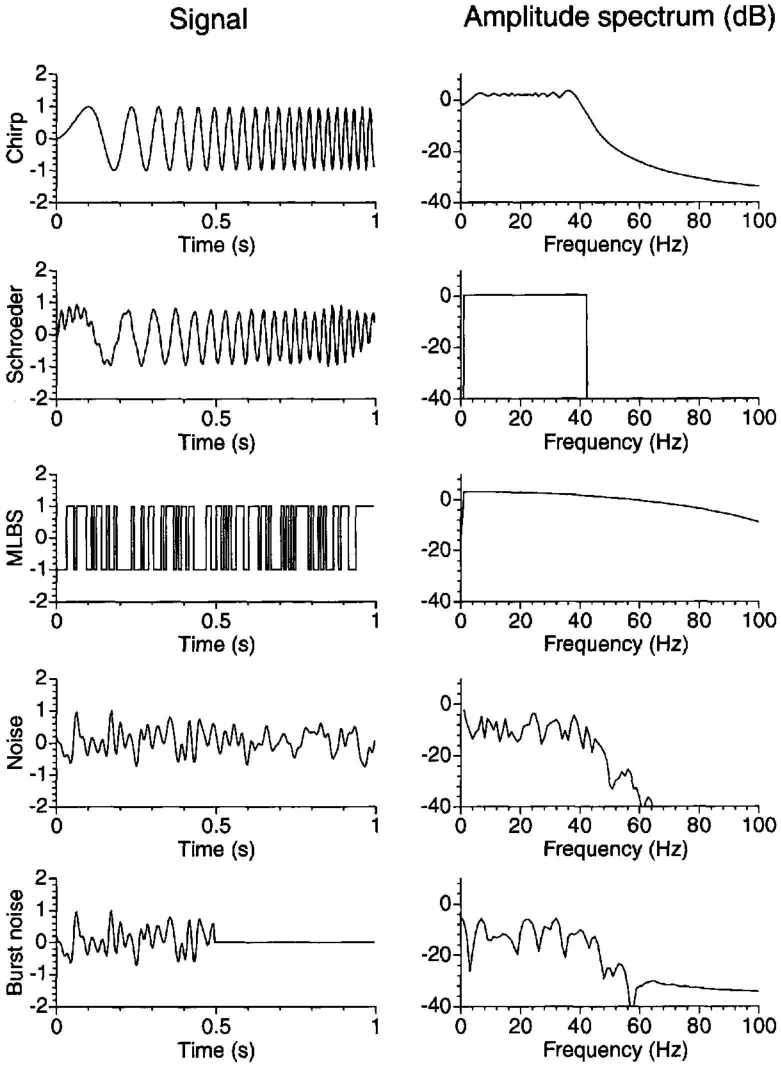


Figure 4-5. Comparison of the general purpose excitation signals in the time (left column) and frequency (right column) domains.

For the MLBS, a clock frequency of 128 Hz was used in order to get better coverage of the frequency band (N=127). The peak value of every signal was scaled to one. The random excitations consist of filtered Gaussian noise (Butterworth filter of order 7 with a cutoff frequency of 42 Hz). The figure is very informative. The multisine is the only signal that exclusively excites the frequency band of interest. All the other signals also excite outside this band. The first three signals inject considerably more power into the system than the noise excitations. After normalization, the power in the frequency band of interest is 1 for the MLBS, 0.81 for the periodic chirp, 0.60 for the multisine, and 0.08 and 0.05 for the random and burst random signals, respectively. The worst measurements will appear at the lines with the smallest amplitude spectrum. This was 0.009 and 0.004 for the random and burst random, 0.55 for the chirp, 1.037 for the multisine, and 1.18 for the MLBS. The amplitude of the chirp droops only at a few border lines of the frequency band of interest; it is slightly above the multisine on most other lines. From this we can conclude that the chirp, multisine, and MLBS have about the same quality and the selection should be based on personal preference, technical possibilities, and second rank arguments that are important for specific situations (e.g., no power outside the band). The random excitations have inferior properties compared with the first three deterministic excitations. They are prone to leakage and inject significantly less power into the system, resulting in a poor SNR.

4.3.2 Optimized Test Signals

Whereas in the previous section we considered signals that could be applied directly, we consider in this section excitation signals where an iterative algorithm is needed to optimize their design. Because of the continuously increasing computer power, this is not a real drawback. The design time runs from a few seconds for simple signals to a few minutes for complex signals with a few hundred frequency components.

Two classes of signals are considered. First the design of multisine excitations with minimized crest factor is discussed, then optimized binary sequences are designed.

4.3.2.1 Optimized Multisines. These are classical multisines where the user chooses the excited frequencies on the equidistant frequency grid kf_0 and also selects the desired amplitude spectrum. This is the signal preferred by the authors because it gives maximal flexibility combined with a minimum measurement time and a maximum quality of the measurements. Moreover, by making a dedicated selection of the components of the excitation signal, it is even possible to detect, qualify, and quantify the presence of nonlinear distortions (see Chapter 3).

Properties

- Periodic signal with period $T_0 = 1/f_0$ → no leakage.
- Frequency resolution $1/T_0$.
- All the power at the user-selected frequencies that can be freely chosen on the discrete grid kf_0.
- The amplitude of the harmonic components can be freely chosen and is exactly realized, no out-of-band power appears.
- Crest factor from 1.4 to 2, depending on the complexity of the amplitude spectrum.

Discussion. Instead of using explicit phase relations for the multisine, a numerical search method is used to select optimal phases that minimize the crest factor. In the literature, many crest factor minimization methods have been presented. In the explicit expressions, the

Schroeder phases are given, allowing a direct calculation of the phases. For multisines with a sparse spectrum, where the frequency lines are few and far apart, or for multisines with an amplitude spectrum that is not flat, the Schroeder phases give no better results than those obtained with a random phase selection, uniformly distributed in $[0,2\pi[$. In these situations, more sophisticated methods are needed and no explicit formulas are available. Two algorithms are proposed. The first one is a clipping procedure that cuts the largest peaks of the signal. The second one is based on the successive minimization of a series of $l_{2p}(\phi)$ norms w.r.t. for increasing p

$$\begin{aligned} l_{2p}(\phi) &= \|u(t,\phi)\|_{2p} = \left(\frac{1}{T_0}\int_0^{T_0} u^{2p}(t,\phi)dt\right)^{\frac{1}{2p}} \\ u(t,\phi) &= \sum\nolimits_{k=1}^{F} A_k\cos(2\pi f_k t + \phi_k) \end{aligned} \qquad (4\text{-}10)$$

with $\phi^T = [\phi_1\phi_2\ldots\phi_F]$ the phases of the multisine $u(t)$, T_0 the period of the multisine, and $p = 2, 4, 8, 16, \ldots$. Compared with the first one, it gives smaller crest factors but needs a larger memory, especially for multisines with a large number of components. With increasing computing power, the last method becomes more and more attractive. Both algorithms are discussed in Appendix 4.A.

Note: Both algorithms can be generalized easily to generate a signal with a power spectrum $S_{uu}(j\omega) + S_{aa}(j\omega)$, with $S_{uu}(j\omega)$ the desired power spectrum and $S_{aa}(j\omega)$ additional power that is added by the algorithm at other frequencies such that the crest factor of the signal decreases further (e.g., by adding additional harmonics to a sine wave, a block-like signal results, pushing the crest factor well below $\sqrt{2}$) (Guillaume et al., 1991). This is called snowing. During the calculation of the crest factor, the additional power is not considered when calculating u_{RMSe}.

Example 4.9 (Flat (Snow) Multisine): The signal of the previous section is also optimized with the l_{2p} algorithm, resulting in a crest factor of 1.42 (compared with 1.67 for the Schroeder multisine). It is shown in Figure 4-6(a). Next, snowing was allowed on the lines 43–255, pushing down the crest factor to 1.19. This made it possible to get 40% more power

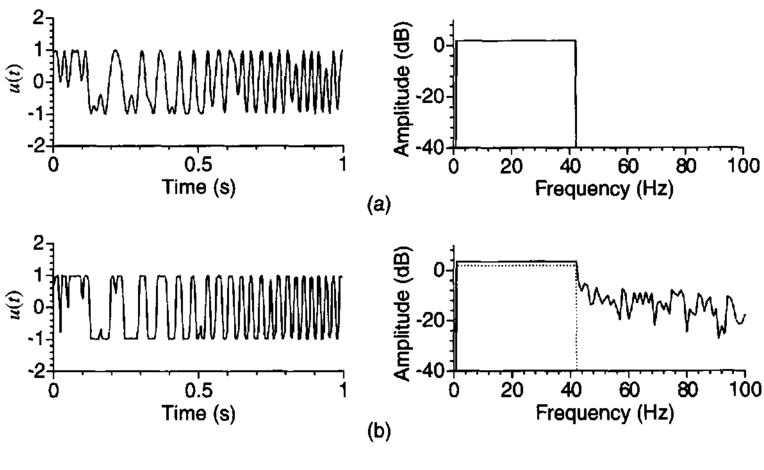


Figure 4-6. Example of the general purpose multisine after optimizing phases. (a) Without snow, (b) with snow: ... (reference signal without snow as in a), ___ with snow.

in the frequency band of interest, compared with the original signal, which had no snow. Compared with the PRBS, 19% more power is injected in the frequency band of interest. About 5% of the totally available power is "wasted" at the snow lines. □

Example 4.10 (Quasi-Logarithmic Excitation): The advantage of the iterative algorithms becomes most obvious when dealing with more complex power spectra. In this example, a quasi-logarithmic multisine is generated, depositing the power at an almost logarithmic frequency grid ($N_{\log} = 4096$, $f_{\max} = 0.4N_{\log}$, $f_{k+1}/f_k \approx 1.05$). Each time the frequencies are shifted to the nearest harmonic line. After optimization, the crest factor is 2.0 (Schroeder phases: 3.3) so that almost three times more power can be injected for the same peak value of the excitation. In this example the crest factor is reduced using the successive minimization algorithm (4-10). The alternative is to use the clipping algorithm (Van der Ouderaa and Renneboog, 1988), but the first algorithm gives better results in a shorter time, at a cost of needing a larger memory. The signal is shown in Figure 4-7. □

4.3.2.2 Discrete Interval Binary Sequence (DIBS). The second class of optimized excitation signals are the discrete interval binary sequences. These are periodic binary sequences, where the sign can change only at an equidistant discrete set of points in time (Van den Bos, 1974; Paehlike and Rake, 1979; Van den Bos and Krol, 1979). The amplitude spectrum of the sequence can be optimized by choosing a good switching sequence so that the energy is concentrated within the frequency band of interest.

Properties

- Periodic signal with arbitrary period length $T_0 = 1/f_0$ → no leakage.
- Frequency resolution $1/T_0$.
- The power is concentrated at the user- selected frequencies that can be freely chosen on the discrete grid kf_0, but the other frequencies are also excited.
- The amplitude of the harmonic components can be freely chosen but is only approximately realized.
- The crest factor depends on the complexity of the signal but is usually rather small.

DISCUSSION. The generation of a DIBS is based on an iterative algorithm proposed by Van den Bos and Krol (1979). The procedure is begun a number of times from different starting values, and the best signal is retained. With a DIBS, it is possible to concentrate the

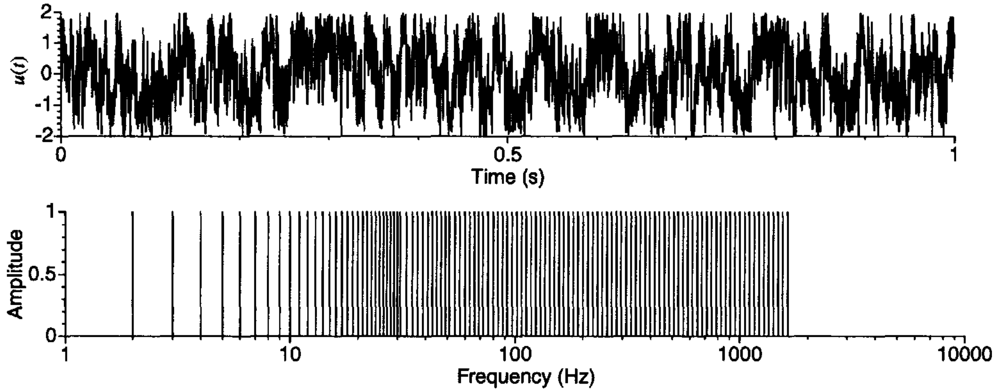


Figure 4-7. Example of a quasi-logarithmic multisine on an equidistant frequency grid.

energy in a discrete set of spectral lines. The crest factor is greater than one because not all of the power is concentrated at the frequencies of interest; but even then, most of the energy can be confined to the frequency band required, which is not possible with the MLBS. Paehlike and Rake (1979) have presented an iterative scheme for putting more of the energy into the weakest spectral lines, thus improving the SNR and decreasing the time factor. Compared with the PRBS, the DIBS can be generated for any sequence length with an arbitrary power spectrum.

Example 4.11 (Low-Pass Spectrum): The general purpose signal of the previous section was also recalculated using this method and is compared with the results of the MLBS in Figure 4-8. The crest factor of the DIBS signal is 1.36, compared with 1.5 for the MLBS. □

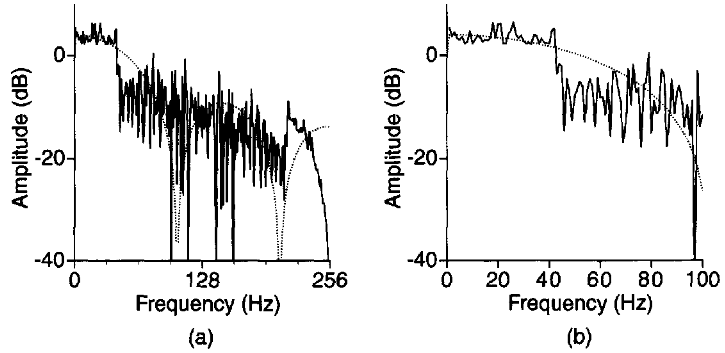


Figure 4-8. Comparison of the spectrum of a DIBS (f_c = 256 Hz, N = 256) and an PRBS (N = 103, f_c = 103 Hz) to generate a flat spectrum in a band 1–42 Hz. (a) Global view, (b) zoom on the frequency band of interest, —— DIBS, ---- PRBS.

Example 4.12 (Bandpass Spectrum): Figure 4-9 illustrates the possibility of creating a bandpass spectrum using a DIBS (crest factor 1.29). Note that this is not possible at all with an MLBS. □

4.3.3 Advanced Test Signals

In this section we discuss some excitation signals with very specialized properties, for example, signals where the crest factor of the first or second derivative is also minimized. These should be used only in critical conditions, where the special shape of the excitation gives a significant advantage. Even for these signals, the additional design time is quite restricted (from a few seconds to a few minutes), but their proper design and application requires a good user's insight into the properties of these signals and their application.

4.3.3.1 Crest Factor Minimization of Linearly Related Multiple Multisines. In some problems, it is not sufficient to keep the crest factor of the excitation low; the system

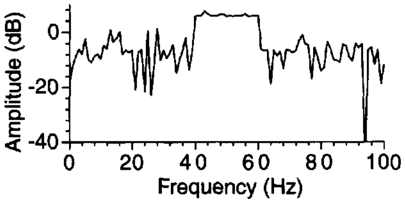


Figure 4-9. Spectrum of a DIBS (f_c = 256 Hz, N = 256) designed to generate a flat spectrum in the band 40–60 Hz.

output should also have a small crest factor. In other applications the signal and its first or second derivative should be small. For example, in mechanical systems the acceleration should be restricted in order to avoid excessive forces, while excessive displacements are avoided to keep the stroke of the shaker small and to maintain a linear behavior of the system. Again, it would be useful if the crest factor of both signals is minimized at the same time.

The l_{2p} crest factor minimization algorithm of the previous section makes it possible to optimize multiple multisines linked by linear systems, e.g., $Y(j\omega) = G(j\omega)U(j\omega)$. Criterion (4-10) is generalized to

$$\begin{aligned}
\left\|\frac{u(t,\phi)}{u_{\mathrm{RMS}}}, \frac{y(t,\phi)}{y_{\mathrm{RMS}}}\right\|_{2p} &= \left(\frac{1}{T_0}\int_0^{T_0}\left(\frac{u^{2p}(t,\phi)}{u_{\mathrm{RMS}}^{2p}} + \frac{y^{2p}(t,\phi)}{y_{\mathrm{RMS}}^{2p}}\right)dt\right)^{\frac{1}{2p}} \\
u(t,\phi) &= \sum\nolimits_{k=1}^{F} A_k\cos(2\pi f_k t + \phi_k) \\
y(t,\phi) &= \sum\nolimits_{k=1}^{F} A_k|G_0(\Omega_k)|\cos(2\pi f_k t + \phi_k + \angle G_0(\Omega_k))
\end{aligned} \tag{4-11}$$

with $\phi^T = [\phi_1\phi_2\ldots\phi_F]$, $\Omega_k = j\omega_k$ for continuous-time systems and $\Omega_k = e^{-j\omega_k T_s}$ for discrete-time systems. In Guillaume et al. (1991), it is shown that the minimum of (4-11), with respect to ϕ, for p growing to infinity results in two multisines with equal and minimum crest factors. Sometimes, it is more advantageous to minimize the scaled peak values of both multisines, allowing optimal use of the full scale of the measurement equipment. This is done by minimizing

$$\left\|u(t,\phi), \frac{y(t,\phi)}{S}\right\|_{2p} \tag{4-12}$$

with S a scaling factor. When S is chosen as the ratio of the rms values, signals with equal crest factors are obtained. Clearly, when S is chosen too large (or too small), the problem reduces to the minimization of $\|u(t,\phi)\|_{2p}$ or $\|y(t,\phi)\|_{2p}$.

Example 4.13 (Crest Factor Minimization of Linearly Related Multisines): A multisine $u(t)$ with $F = 512$ consecutive components is designed to have minimum crest, together with its second derivative $d^2u(t)/dt^2$. Table 4-1 gives the crest factors that are obtained using l_{2p} and the resulting output signals $d^2u(t)/dt^2$ are shown in Figure 4-10. As can be seen, the crest factor is reduced to 60% of its original value. In the case of a mechanical system this allows a significant reduction of the forces and, hence, the dimensions and the cost of the shaker used to generate the signals. □

TABLE 4-1 Crest Factor Minimization of $u(t)$ and $d^2u(t)/dt^2$

	Crest Factor Input	Crest Factor Output
Input min. (4-10)	1.39	2.85
Input/output min. (4-11)	1.61	1.63

4.3.3.2 Multilevel Excitation Signals. The DIBS (see Section 4.3.2) is a binary sequence that excites two levels only. In some applications, ternary signals can be used (e.g., levels −1, 0, 1), allowing greater flexibility during the design. In general, this leads to the following results:

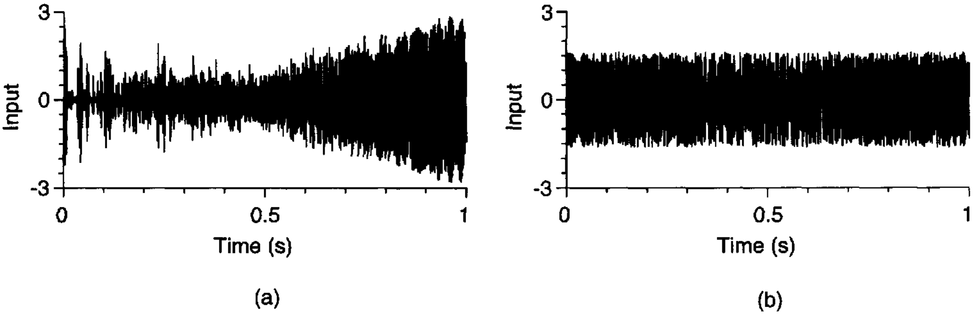


Figure 4-10. The second order derivative of a multisine with minimum crest factor (a) and input-output optimized crest factor (b).

The total power of the signal decreases: since the signal is set equal to zero at some points (instead of −1 or 1), it is clear that less power is available in the design.

The out-of-band power is reduced: the greater flexibility due to the additional level gives better control over the power spectrum. This makes it possible to reduce the out-of-band power.

The lowest in-band level is about the same: although less power is generated, the lowest amplitude at a frequency line of interest remains almost the same. This guarantees that the minimum uncertainty of the measurement will be the same for binary and three-level signals. However, by using the ternary signal, less power is wasted.

The design of multilevel signals is extensively discussed by McCormack et al. (1995) and Barker and Zhuang (1997).

4.3.3.3 Harmonic Suppression. In Section 3.5.1 it was shown that periodic signals with an odd (spectral lines $2k+1$ present) or odd-odd (spectral lines $4k+1$ present) spectrum make it possible to eliminate the even nonlinearities and detect the presence of odd nonlinearities. Such signals can easily be obtained from multisines where the amplitudes of the corresponding lines are put to zero. It is also possible to create such signals from binary sequences. The inversely repeated sequence $[u(t), -u(t)]$ has no even components in its spectrum. Using multilevel designs (Barker and Zhuang, 1997), it is also possible to suppress the second and third harmonics of a set of specified primes. Finally, it is also possible to design sparse harmonic multisines which facilitate a direct probing of the second and third degree Volterra kernels (Evans, 1998; Boyd et al., 1983) with a minimum interference.

4.4 OPTIMIZATION OF EXCITATION SIGNALS FOR PARAMETRIC MEASUREMENTS

4.4.1 Introduction

Here, the parametric measurement problem is studied. We will concentrate on the parameters θ of the mathematical model $G(\Omega_k, \theta)$, with $\Omega_k = j\omega_k$ for continuous-time systems and $\Omega_k = e^{-j\omega_k T_s}$ for discrete-time systems, which describes the measured transfer function $G_0(\Omega_k)$. To fit the model $G(\Omega_k, \theta)$ on the measurements $G(\Omega_k)$, a cost function

$V(\theta, Z)$, with Z a vector containing the measured input-output DFT spectra, which is an indication of the quality of the fit, is minimized. As explained in Chapter 1, a simple and very popular choice for $V(\theta, Z)$ is the least squares cost function in which the squared differences between the model and the measurements are summed together. Another possibility is to embed the choice of the cost function in a statistical framework, as done for the maximum likelihood (ML) estimator, resulting in a weighted least squares estimator if the disturbing noise is normally distributed (Chapter 1). The quality of the estimates strongly depends on the excitation signals applied during the experiment. As in the nonparametric case, the excitation signal will be optimized in two steps, the first being the selection of an optimized power spectrum followed by a crest factor minimization of the involved signals in the second step.

To optimize the input spectrum, we need a scalar criterion that is sensitive to the accuracy of all the parameters of the system. The determinant of the covariance matrix, which is equal to the volume of the uncertainty ellipsoid, is such a criterion.

A range of criteria, other than the determinant, can be found in the literature, optimization of the trace being the most popular. For the sake of brevity, we limit ourselves in this text to examining the minimization of the determinant of the covariance matrix. For more information on other criteria, the reader is referred to other publications (Federov, 1972; Goodwin and Payne, 1977; Zarrop, 1979; Walter and Pronzato, 1997).

For computational simplicity, the covariance matrix is approximated by the Cramér-Rao lower bound (inverse information matrix) because the latter is easier to calculate (Chapter 1). General expressions of the information matrix can be obtained (without specifying an estimator) and the problem of minimizing the determinant of the covariance matrix is replaced by maximizing the determinant of the information matrix. This approximation is valid if the covariance matrix of the actual estimator approximates the Cramér-Rao lower bound sufficiently close for the considered experiments.

4.4.2 Optimization of the Power Spectrum of a Signal

4.4.2.1 Preliminary Aspects. The information matrix is the kernel of optimizing algorithms. It is a real symmetric and semipositive definite $n_\theta \times n_\theta$ matrix, where n_θ is the number of unknown model parameters. Each optimal design in the frequency domain can be reduced to a design consisting of a discrete set of $n_\theta(n_\theta + 1)/2 + 1$ frequencies (Federov, 1972; Goodwin and Payne, 1977), which corresponds to the number of free parameters in a symmetric $n_\theta \times n_\theta$ matrix + 1. The minimum number of frequencies required to avoid a nonsingular information matrix is $\text{int}(n_\theta/2)$ (with $\text{int}(x)$ the integer part of x). When using classical optimizing algorithms, the computer time needed to search for an extreme value depends strongly on the number of frequencies. From a modeling point of view, however, the minimum number is undesirable, because if an estimate of n_θ parameters is made using $\text{int}(n_\theta/2)$ frequencies, there is no possibility of detecting model errors. A second drawback of working with the minimum number of frequencies is that it is more difficult to compress the signals in the time domain.

Most algorithms presented in the literature searched for optimal designs with the minimum number of frequencies. We present a method for designing optimal power spectra based on a discrete frequency grid: this is not in itself a restriction because we look for periodic signals that have discrete spectra. This will lead to a significant reduction of the computation time. The method can be applied in the Laplace domain (continuous-time systems) as well as in the z-domain (discrete-time systems). In order to stress this equivalence, we use Ω as the frequency variable in the following interchangeable manner: $\Omega = j\omega$ (Laplace), or $\Omega = e^{-j\omega T_s}$ (z-domain). The following function is used in the optimization algorithm.

Definition 4.14 (Dispersion Function): The dispersion function $v(\chi, \Omega_k)$ for a given input power spectrum

$$\chi(\Omega) = (|U(1)|^2 ... |U(F)|^2), \text{ with } \textstyle\sum_{k=1}^{F} |U(k)|^2 = \wp \qquad (4\text{-}13)$$

is

$$v(\chi, \Omega_k) = \text{trace}([Fi(\chi)]^{-1} fi(\Omega_k)) \qquad (4\text{-}14)$$

with $Fi(\chi)$ the information matrix resulting from the design $\chi(\Omega)$, $fi(\Omega_k)$ the information matrix corresponding to a single frequency input with a normalized power spectrum $|U(k)|^2 = \wp$, and Ω_k the frequency.

The dispersion function has the following properties:

- The dispersion function can be related to the input and output noise on the measurements (Schoukens and Pintelon, 1991) as

$$v(\chi, \Omega_k) = \frac{2\sigma_G^2(\Omega_k, \theta)\wp}{\sigma_U^2(k)|G(\Omega_k)|^2 + \sigma_Y^2(k) - 2\text{Re}(\sigma_{YU}^2(k)\bar{G}(\Omega_k))} \qquad (4\text{-}15)$$

  with $\sigma_G^2(\Omega_k, \theta)$ the uncertainty on the transfer function using the Cramér-Rao lower bound as covariance matrix for the model parameters. The dispersion can be interpreted as the ratio of the variance of the system frequency response, calculated with the estimated parameters, to the noise power of the measurements referred to the output of the system at the frequency Ω_k.

- The dispersion function is a normalized quantity:

$$\textstyle\sum_{k=1}^{F} v(\chi, \Omega_k)\dfrac{|U(k)|^2}{\wp} = n_\theta \qquad (4\text{-}16)$$

  (Goodwin and Payne, 1977).

- The maximum of the dispersion function $v(\chi, \Omega_k)$ over the frequency grid is larger than or equal to the number of parameters n_θ (Goodwin and Payne, 1977).

These three properties will be used in the algorithm for designing an optimized excitation signal.

4.4.2.2 An Efficient Algorithm for Maximizing the Information Matrix. Although the optimal input may be found analytically for simple situations, in general, no closed form solution can be found. Therefore, an iterative design is required. Most algorithms carry out a search in the continuous frequency space to find the frequency with the maximum dispersion and then add extra energy at this frequency. The resulting spectrum is normalized, and the procedure is repeated until the variations are negligible. More sophisticated algorithms combine this procedure with a mechanism that removes components from the spectrum (Federov, 1972; Zarrop, 1979). The search for a maximum is very time consuming, and the final spectrum is difficult to generate because the optimal frequencies are not harmonically related. For these reasons, it is better to reduce the frequency space to a discrete set of frequencies in the analysis; the implications of this restriction for the attainable accuracy are studied in more detail by Van den Eijnde and Schoukens (1991), and it turns out that there is no significant loss in attainable accuracy if the discrete set of frequencies is sufficiently dense.

In general, any discrete set of frequencies can be used, but if only periodic signals are retained, it is obvious that the selected frequencies should be harmonically related. For the initial design, the simplest first choice is that of equally spaced spectral lines within the frequency band of interest, with the total fixed input power uniformly distributed over the F frequencies in this set. The resulting spectrum constitutes the initial design χ_0. The response dispersion function $v(\chi_0, \Omega_k)$ is computed for every spectral line Ω_k in the set, and the available power is redistributed over all spectral lines proportionally to the corresponding values of the dispersion function. The optimal input is found by repeating this procedure; the iteration can be stopped when the variation of the determinant of the information matrix is small. This method was described in the late 1970s (see Walter and Pronzato, 1997, pp. 305–306, and the references therein). If we express this approach in mathematical terms, we end up with an algorithm with the following consecutive steps:

Algorithm 4.15 (Optimization Power Spectrum)

1. Initiation:
   Select a set $\mathbb{F}$ of F frequencies $\Omega_1, ..., \Omega_F$ within the frequency band of interest: $\mathbb{F} = \{\Omega_1, ..., \Omega_F\}$. Distribute the input power equally over these F frequencies. This constitutes the initial design χ_0.
2. Iteration:
   2a. Set $i = i + 1$ and compute the response dispersion function $v(\chi_i, \Omega_k)$ for $k = 1, ..., F$.
   2b. Compose a new design in the following way:

$$\chi_{i+1}(\Omega_k) = \chi_i(\Omega_k)v(\chi_i, \Omega_k)/n_\theta \text{ for } k = 1,...,F \tag{4-17}$$

   2c. If $\max(v(\chi_i, \Omega_k) - n_\theta) < \varepsilon$ with ε sufficiently small and $\Omega_k \in \mathbb{F}$, then the optimum design is found; otherwise go to step 2a.

Proof. See Van den Eijnde and Schoukens (1991) and Delbaen (1990). □

It has been shown (Walter and Pronzato, 1997; Delbaen, 1990) that each run of this algorithm yields a superior input design and that consecutive designs converge monotonously to a design with the optimum dispersion function and, hence, the minimum determinant of the Cramér-Rao bound.

4.4.2.3 Importance of Crest Factor Minimization. In a second step, after the selection of the power spectrum, the crest factor of the corresponding multisine(s) should be minimized. To compare different excitations, it is necessary to scale the determinant of the Cramér-Rao lower bound and the dispersion function with the optimized crest factor so that all signals are compared for the same peak value.

$$\det(CR_{\text{scaled}}(\theta)) = \det(CR(\theta))Cr^{2n_\theta}(u) \tag{4-18}$$

4.4.2.4 Practical Implementation. It is obvious that the calculation of the optimum amplitude spectrum is possible only if enough knowledge of the system is available. In most situations, a two-step procedure is required, restricting the applicability of these methods significantly. In the first step, the unknown parameters are estimated using a multisine with a flat amplitude spectrum; in the second step, these estimated values are used to optimize the amplitude spectrum. The covariance matrix of the estimated, unknown model parameters should be close enough to the Cramér-Rao lower bound.

4.4.2.5 Example: An Experimental Verification. The power spectrum optimization for a parametric measurement is illustrated in the following example:

$$G_0(s) = \frac{b_2s^2 + b_3s^3 + b_4s^4}{a_0 + a_1s + \ldots + a_6s^6} \tag{4-19}$$

The coefficients are given in Table 4-2 and the corresponding amplitude characteristic is given in Figure 4-11. The system is excited with a multisine at the frequencies $f_k = kf_0$, with $k = 25, 26, \ldots, 100$ and $f_0 = 50/2048$ MHz. The rms value of the multisine is set equal to $1/\sqrt{2}$. Two multisines are considered, the first one having a flat amplitude spectrum and the second one being optimized on the basis of the procedure described before. The evolution of the power spectrum optimization process is given in Figure 4-12. The optimization is stopped before the final convergence is reached (after three iterations) to avoid signals with a sparse spectrum. These are very difficult to compress and have a large crest factor. From (4-18) it is seen that this would jeopardize the accuracy gain that is obtained with the design of an optimal spectrum. In this example the determinant of the corresponding Cramér-Rao lower bound was reduced with a factor 43 after three iterations.

TABLE 4-2 Coefficients of the Transfer Function of the Sixth-Order Continuous-Time Bandpass Filter

		b_2	b_3	b_4		
		8.973e-10	5.5155e-12	3.2010e-17		
a_0	a_1	a_2	a_3	a_4	a_5	a_6
1	2.5017e-4	3.5869e-7	5.5550e-11	3.36031e-14	2.5351e-18	1.0131e-21

The crest factors or peak values of the multisine at the input and output are minimized using the l_{2p} algorithm (4-12) and the results are given in Table 4-3. Three situations are considered:

- Minimization of the crest factor of the input signal
- Simultaneous minimization of the crest factors of the input and output
- Simultaneous minimization of the peak values of the input and output

For our purpose, the last possibility is the most interesting because it will determine the settings of the full scale of the measurement instruments. In Table 4-3, it is seen that the peak values of the multisine, with the optimized power spectrum, are equal to those of the multi-

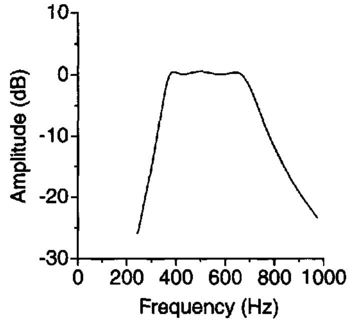


Figure 4-11. Amplitude transfer characteristic of the studied system.

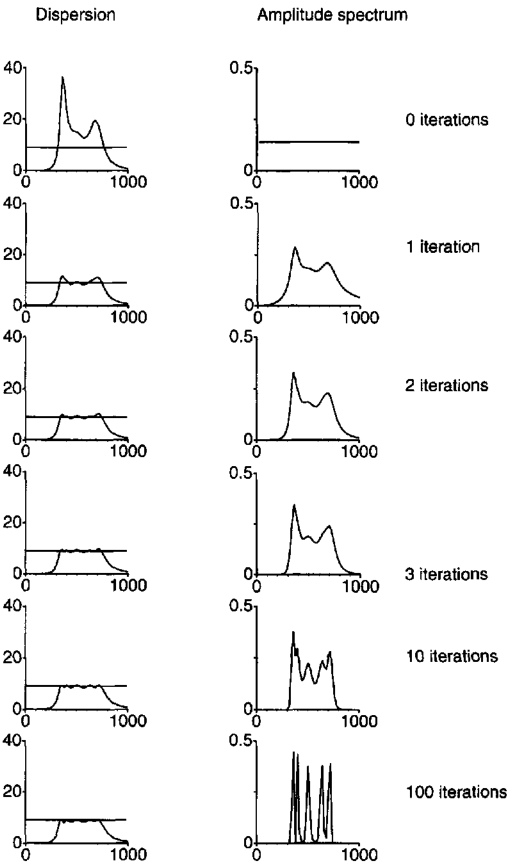


Figure 4-12. Evolution of the power spectrum optimization process.

sine with flat power spectrum (optimization c). So the settings of the measurement instruments can remain the same for both excitations, and, consequently, the noise on both measurements will be equal. However, the uncertainty on the estimated parameters will be smaller in the second case because the determinant $\det(Fi(\theta))$ is much smaller than in the first case, resulting in a smaller uncertainty on the calculated transfer characteristics.

TABLE 4-3 Minimization of the Crest Factor(s) or Peak Values of Two Multisines, Related by the Linear System (4-19)

	Input		Output	
	Crest Factor	Peak Value	Crest Factor	Peak Value
Flat input power spectrum				
a	1.459	1.031	2.749	1.418
b	1.667	1.170	1.667	0.862
c	1.509	1.067	2.065	1.067
Optimized input power spectrum				
c	1.459	1.031	1.860	1.200
b	1.582	1.118	1.582	1.026
c	1.508	1.066	1.643	1.066

a: minimization of the crest factor of the input

b: simultaneous minimization of the crest factors of the input and output

c: simultaneous minimization of the peak values of the input and output

From experimental tests, it turned out that these signals can be generated in practice; small disturbances at the amplitudes or the phases in the generator (and reconstruction filter) do not result in an excessive growth of the crest factor. In Figure 4-13(b) measurements of the calculated multisines are given. They were generated with a 12-bit arbitrary waveform generator with 2048 points in one period (sampling frequency 20 kHz). The generator was followed by a reconstruction filter (a Cauer filter with a cutoff frequency of 2 kHz). No phase or amplitude compensation was made for the distortion introduced by the reconstruction filter. If this amplitude/phase distortion becomes disturbing, it is always possible to give a precompensation to the amplitudes/phases of the multisine. The measurements were made with an 8-bit digitizer (full scale ± 1 V) at 512 points with a sampling frequency of 5 kHz.

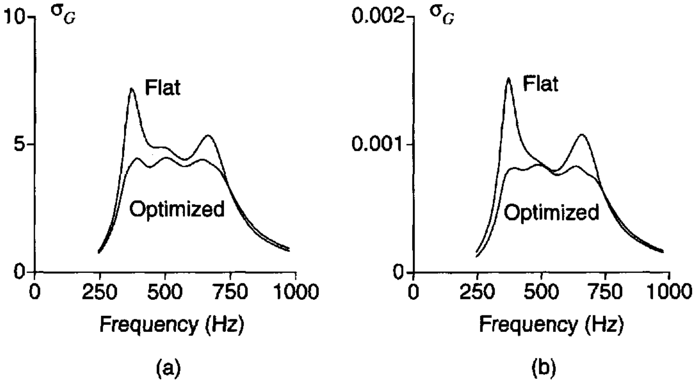


Figure 4-13. Comparison of the model uncertainty with the flat and the optimized power spectrum, (a) theoretical (scaled) results, (b) experimental results.

Figure 4-13(a) compares the uncertainty $\sigma_G(\Omega_k, \theta)$ on the estimated transfer function model in case a multisine with a flat and an optimized amplitude spectrum is used. These results were experimentally verified using the setup described before. Sixty measurements were made to measure the standard deviation of the FRF measurement. The results are shown in Figure 4-13(b). It is obvious that this result is relevant only if the model errors of the parametric model in the identification step are smaller than the identification uncertainty due to the noise.

4.5 APPENDIX

Appendix 4.A Minimizing the Crest Factor of a Multisine

4.A.1 Clipping Algorithm. In Van der Ouderaa et al. (1988a, 1988b) an iterative method has been developed to optimize the phases. The method is very close to an algorithm presented by Van den Bos (1987). The basic idea behind this method is a clipping procedure, which is illustrated in Figure 4-14. For a given amplitude spectrum, a time signal with a minimum peak value has to be found. The iteration procedure is started from the specified amplitude spectrum, and arbitrary phases are taken as starting values. Using the inverse Fourier transform, the signal is calculated at a set of discrete equidistant times. A new time signal is then generated by clipping off all the values larger than a given maximum, and the new modified spectrum and phases are calculated using the FFT. These new phases are retained as a first approximation to the solution, but the modified amplitude spectrum is rejected in favor of the original one. This procedure is repeated until no further significant reduction of the

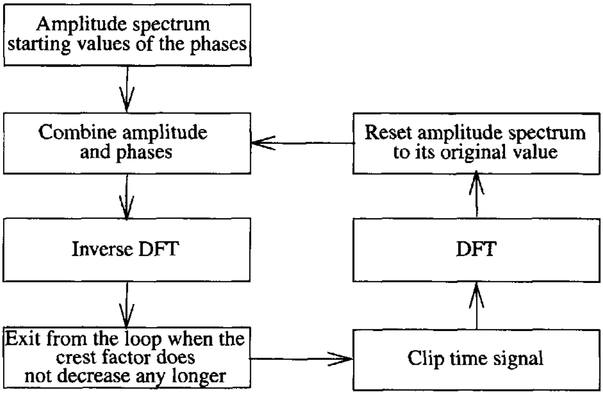

Amplitude spectrum
starting values of the phases
Combine amplitude and phases
Reset amplitude spectrum to its original value
Inverse DFT
DFT
Exit from the loop when the crest factor does not decrease any longer
Clip time signal

Figure 4-14. Minimization of the crest factor of a multisine: clipping algorithm.

crest factor is obtained. During the iteration process, the clipping level is changed from a low value in the beginning (e.g., 0.7 $u_{\max}$) to almost no clipping (e.g., 0.999 $u_{\max}$) at the end of the process, for strongly compressed signals. In general, the algorithm needs a few hundred iterations to obtain useful signals (for example, a flat multisine with a crest factor of 1.5), but in order to obtain near-optimal crest factors (of 1.4) a few hundred thousand iteration steps are more likely to be required. This algorithm is called the clipping algorithm.

4.A.2 Infinity Norm Algorithm. In Guillaume et al. (1991) an algorithm has been developed based on the minimization of the l_{2p} norm

$$\begin{aligned} l_{2p}(\phi) &= \|u(t, \phi)\|_{2p} = \left(\frac{1}{T_0}\int_0^{T_0} u^{2p}(t, \phi)dt\right)^{\frac{1}{2p}} \\ u(t, \phi) &= \sum\nolimits_{k=1}^{F} A_k \cos(2\pi f_k t + \phi_k) \end{aligned} \tag{4-20}$$

with T_0 the period of the multisine and $p = 2, 4, 8, 16, \ldots$. It is shown that the $l_{2p}(\phi)$ norm is equal to

$$l_{2p}(\phi) = \left(\frac{1}{N}\sum\nolimits_{t=0}^{N-1} u^{2p}(tT_s, \phi)\right)^{\frac{1}{2p}} \text{ if } N \geq 2pf_{\max}T_0 + 1 \tag{4-21}$$

with $f_{\max}$ the maximum frequency occurring in the multisine and N the number of samples in one period. Condition $N \geq 2pf_{\max}T_0 + 1$ in (4-21) expresses that no alias contribution may appear on the DC component.

The l_{2p} norm is minimized with respect to the phases using a Marquardt algorithm for values of p that are gradually increased during the iteration process. This defines a descent algorithm that converges to a local minimum. From our experiences, it turned out that the results of this algorithm were better than those obtained with the previous method. In practice, conditions (4-21) may be violated as long as a sufficiently large number of points is considered (e.g., $N \geq 16f_{\max}T_0 + 1$), leading to a significant reduction of the calculation time.

5

Models of Linear Time-Invariant Systems

Abstract: This chapter presents the nonparametric and parametric system (signal) and noise models used throughout this book. The models are described in the frequency domain and cover linear time-invariant discrete-time systems (z-domain), continuous-time systems (s-domain), diffusion phenomena ($\sqrt{s}$-domain), commensurate microwave systems ($\tanh(\tau_R s)$), and damped (complex) exponentials. The classical transfer function models describing the relationship between the DFT spectra of the input and output signals are valid for periodic and time-limited signals only. These models are extended to arbitrary excitations for discrete-time and continuous-time systems. Extended transfer function models are also derived in case samples are missing at the input and/or output signals. The identifiability issues of the different models are discussed and generalizations to the multivariable case are given. The basic concepts of linear system theory are assumed to be known. Textbooks on the topic are by Oppenheim et al. (1997), Kailath (1980), and Kwakernaak and Sivan (1991).

5.1 INTRODUCTION

Although most real-life processes are nonlinear and time variant, they can often be approximated very well by linear time-invariant systems. Linear time-invariant continuous-time systems are described by differential equations (finite dimensional or lumped systems) or partial differential equations (infinite dimensional or distributed systems) with constant coefficients. The transfer function between the input $u(t)$ and the output $y(t)$ of the process is calculated assuming that the initial conditions are zero.

Example 5.1 (Lumped Continuous-Time System): Consider the LC resonator of Figure 5-1.

The input of the system is the voltage source $u(t)$ and the output is the voltage $y(t)$ across the capacitor. Both are related by a second-order differential equation,

$$LC\frac{d^2y(t)}{dt^2} + y(t) = u(t) \tag{5-1}$$

System Identification: A Frequency Domain Approach. Rik Pintelon and Johan Schoukens
Copyright © 2001 Institute of Electrical and Electronics Engineers. ISBN: 0-780-36000-1

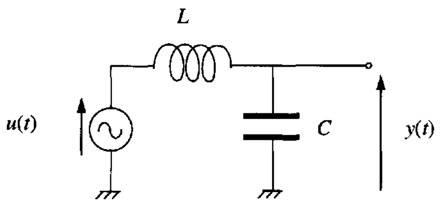


Figure 5-1. *LC* series resonator.

Taking the Laplace transform of (5-1) assuming that the initial conditions are zero ($y(0) = 0$ and $y'(0) = 0$) gives the transfer function

$$G(s) = \frac{Y(s)}{U(s)} = \frac{1}{1 + LCs^2} \tag{5-2}$$

Note that $G(s)$ has one complex conjugate pole pair $s = \pm j/\sqrt{LC}$ on the imaginary axis. □

Example 5.2 (Distributed Continuous-Time System): Consider the clamped beam of Figure 5-2.

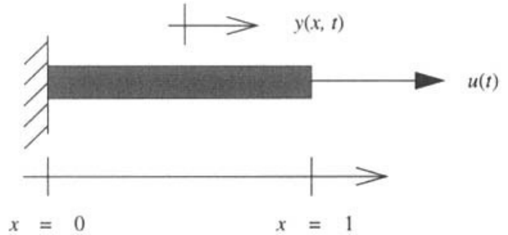


Figure 5-2. Longitudinal vibrations of a clamped beam.

The input of the system is the force per unit area $u(t)$ and the output is the longitudinal displacement $y(x, t)$. Both are related by a second-order partial differential equation,

$$\frac{\partial^2 y(x, t)}{\partial t^2} = \frac{E}{\rho}\frac{\partial^2 y(x, t)}{\partial x^2} \tag{5-3}$$

with boundary conditions $y(0, t) = 0$ and $\partial y(x, t)/\partial x|_{x=l} = u(t)/E$. E, ρ are, respectively, the elasticity modulus and the density of the beam. The transfer function between the force per unit area $u(t)$ and the longitudinal displacement at the end of the beam $y(l, t)$ is calculated, assuming zero initial conditions $y(x, 0) = 0$, $\partial y(x, t)/\partial t|_{t=0} = 0$. We find

$$G(s) = \frac{Y(l, s)}{U(s)} = \frac{l}{E}\frac{\tanh(\tau s)}{\tau s} \tag{5-4}$$

with $\tau = \sqrt{\rho l^2/E}$. Note that $G(s)$ has an infinite number of complex conjugate pole pairs $s = \pm(2k+1)\frac{\pi}{2\tau}j$, $k \in \mathbb{N}$ on the imaginary axis (see Exercise 5.1). According to the Mittag-Leffler theorem (Henrici, 1974), (5-4) can be expanded in an infinite series of partial fractions (see Exercise 5.2)

$$G(s) = \frac{l}{E}\sum_{k=0}^{\infty}\frac{2}{(\tau s)^2 + (\pi(2k+1)/2)^2} \tag{5-5}$$

Because the active frequency range of $|2/((\tau s)^2 + (\pi(2k+1)/2)^2)|_{s = j\omega}$ is limited, it follows from (5-5) that, within a given frequency band, (5-4) can be approximated very well by a rational transfer function of finite order in s. □

The conclusions of Example 5.2 are valid for most physical infinite-dimensional processes: their irrational transfer functions have an infinite (countable) number of poles (those at infinity included) and can be approximated well in a limited frequency band by a rational form of finite order in s. The advantage of using a rational approximation is that the form of the model is robust w.r.t. (small) changes in the geometry and/or the boundary conditions. This is not the case for the irrational transfer function models, because they must be recalculated for each particular geometry and boundary condition. The disadvantage of the rational approximation is that the model contains too many parameters; for example, the irrational transfer function (5-4) has two independent parameters while a rational approximation of order two uses five independent parameters.

The irrational transfer functions of systems where diffusion phenomena such as mass or heat transfer are important are very often a function of $\sqrt{s}$. For such systems it might be a good idea to use a rational approximation in $\sqrt{s}$ instead of s. Examples of such systems are electrochemical processes where the charge transport, controlled by diffusion, is modeled by an impedance (Warburg impedance) that is proportional to $\sqrt{s}$ (Wang, 1987).

The irrational transfer functions of lossless commensurate microwave devices are a rational function of the Richards variable $S = \tanh(\tau_R s)$ (Rizzi, 1988). For real (lossy) microwave devices it might be a good idea to use rational approximations in $\tanh(\tau_R s)$ instead of s.

When a lumped continuous-time system is excited by a piecewise constant signal, then there exists a discrete-time model that, exactly, describes the input-output behavior of system at the sampling instances (see Example 5.3). This result is used in control applications where the input of the system (plant) is the piecewise constant output of a digital controller.

Example 5.3 (Discrete-Time System): Consider a lumped continuous-time system (see Figure 5-3) excited by a piecewise constant excitation signal

$$u_{\text{zoh}}(t) = \sum\nolimits_{r=0}^{\infty} u(r)\text{zoh}(t - rT_s) \tag{5-6}$$

with $\text{zoh}(t) = 1$ for $t \in [0, T_s)$ and $\text{zoh}(t) = 0$ elsewhere. The Laplace transform of the output $y(t)$ equals

$$Y(s) = \frac{G(s)}{s}(1 - z^{-1})U(z)\Big|_{z = e^{sT_s}} \tag{5-7}$$

with $U(z)$ the Z-transform of the samples $u(k)$. Applying the residue formula

$$Z\{Y(s)\} = \sum_{\text{poles } Y(s)} \text{Res}\left(\frac{z}{z - e^{sT_s}}Y(s)\right) \tag{5-8}$$

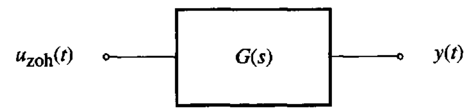


Figure 5-3. Lumped continuous-time system excited by a piecewise constant signal.

(Selby, 1973) to (5-7), we find the Z-transform of the sampled output $y(kT_s)$

$$Y(z) = (1-z^{-1})U(z)Z\{G(s)/s\} \tag{5-9}$$

It follows that there exists a discrete-time model with transfer function

$$G_{\mathrm{ZOH}}(z^{-1}) = Y(z)/U(z) = (1-z^{-1})Z\{G(s)/s\} \tag{5-10}$$

that exactly describes the input-output behavior of the continuous-time model at the sampling times $t = kT_s$. Result (5-10) can be generalized to the cascade of two systems (see Figure 5-4). However, in this case the discrete-time model relating the sampled input $u(t)$ to the sampled output $y(t)$ of the plant $G(s)$

$$G_d(z^{-1}) = Y(z)/U(z) = \frac{Y(z)/R(Z)}{U(z)/R(Z)} = \frac{Z\{L(s)G(s)/s\}}{Z\{L(s)/s\}} \tag{5-11}$$

depends on the characteristics of the preceding system $L(s)$ (see Exercise 5.4). □

The results of Example 5.3 can be generalized to a certain class of nonlinear continuous-time systems. If a continuous-time Volterra system is excited by a piecewise constant signal, then there exists a discrete-time Volterra model that, exactly, describes the input-output behavior of the system at the sampling instances (see Example 5.4).

Example 5.4 (Nonlinear Discrete-Time System): The output $y(t)$ of a time-invariant continuous-time Volterra system can be written as

$$\begin{aligned} y(t) &= \textstyle\sum_{\alpha=1}^{\infty} y_\alpha(t) \\ y_\alpha(t) &= \int_0^\infty \int_0^\infty \cdots \int_0^\infty g_\alpha(\tau_1, \tau_2, \ldots, \tau_\alpha)u(t-\tau_1)u(t-\tau_2)\ldots u(t-\tau_\alpha)d\tau_1 d\tau_2 \ldots d\tau_\alpha \end{aligned} \tag{5-12}$$

with $u(t)$ the input, $y_\alpha(t)$ the nonlinear contribution of degree α, and $g_\alpha(\tau_1, \ldots, \tau_\alpha)$ the multidimensional impulse response of degree α (Schetzen, 1980). Note that $y_\alpha(t)$ is written as a multidimensional convolution of $g_\alpha(\tau_1, \ldots, \tau_\alpha)$ with the input. The contribution of degree α in (5-12) can always be written as

$$y_\alpha(t) = \sum_{n_1, \ldots, n_\alpha = 1}^{\infty} \int_{(n_1-1)T_s}^{n_1 T_s} \cdots \int_{(n_\alpha-1)T_s}^{n_\alpha T_s} g_\alpha(\tau_1, \ldots, \tau_\alpha)u(t-\tau_1)\ldots u(t-\tau_\alpha)d\tau_1 \ldots d\tau_\alpha \tag{5-13}$$

Evaluating (5-13) at $t = kT_s$ for piecewise constant inputs $u_{\mathrm{zoh}}(t)$ (5-6), taking into account that $u_{\mathrm{zoh}}(kT_s - \tau) = u(k-n)$ for $\tau \in ((n-1)T_s, nT_s]$, gives

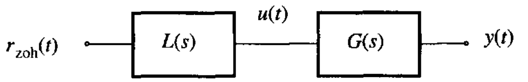


Figure 5-4. Cascade of continuous-time systems excited by a piecewise constant signal.

$$y(kT_s) = \sum_{\alpha=1}^{\infty} y_\alpha(kT_s)$$
$$y_\alpha(kT_s) = \sum_{n_1, n_2, \ldots, n_\alpha = 1}^{\infty} g_{\alpha\text{zoh}}(n_1, n_2, \ldots, n_\alpha)u(k-n_1)u(k-n_2)\ldots u(k-n_\alpha) \tag{5-14}$$

where $g_{\alpha\text{zoh}}(n_1, n_2, \ldots, n_\alpha)$ is defined as

$$g_{\alpha\text{zoh}}(n_1, \ldots, n_\alpha) = \int_{(n_1-1)T_s}^{n_1 T_s} \ldots \int_{(n_\alpha-1)T_s}^{n_\alpha T_s} g_\alpha(\tau_1, \ldots, \tau_\alpha)d\tau_1\ldots d\tau_\alpha \tag{5-15}$$

Equation (5-14) is a shift-invariant discrete-time Volterra model (Brillinger, 1981) that exactly describes the input-output behavior of the time-invariant continuous-time Volterra system (5-12) at the sampling times $t = kT_s$.

Note that the Z-transform of the linear contribution in (5-14),

$$y_1(kT_s) = \sum_{n_1=1}^{\infty} g_{1\text{zoh}}(n_1)u(k-n_1) \text{ with } g_{1\text{zoh}}(n_1) = \int_{(n_1-1)T_s}^{n_1 T_s} g_1(\tau_1)d\tau_1$$

is exactly (5-9) and (5-10). □

We conclude from Examples 5.1 to 5.3 that rational transfer function models of some generalized frequency variable are appropriate for describing a broad class of (in)finite-dimensional linear time-invariant systems. The stable and minimum phase regions of the poles and zeros in the s-, z- and $\sqrt{s}$-domains are shown in Figure 5-5 (proof: see Appendix 5.A). In what follows, we discuss several possible parameterizations of transfer function models and establish the relationship with the discrete Fourier transforms (DFTs) of the input and output signals.

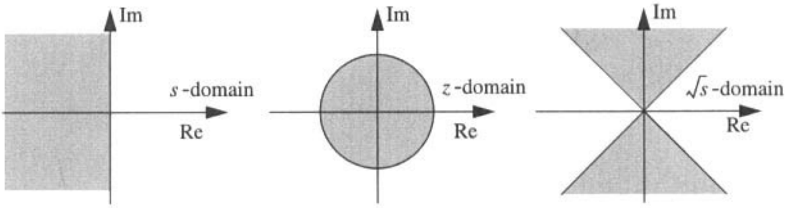


Figure 5-5. Gray area: stable and minimum phase regions of, respectively, the poles and zeros. s-domain: $\text{Re}(s) < 0$, z-domain: $|z| < 1$, and $\sqrt{s}$-domain: $|\text{Re}(\sqrt{s})| < |\text{Im}(\sqrt{s})|$.

For lumped continuous-time and discrete-time systems the transfer function models, and their relationship with the input-output DFT spectra, are obtained by taking, respectively, the Laplace transform of the following differential equation:

$$\sum_{n=0}^{n_a} a_n y^{(n)}(t) = \sum_{m=0}^{n_b} b_m u^{(m)}(t) \tag{5-16}$$

and the Z-transform of the following difference equation:

$$\sum_{n=0}^{n_a} a_n y(t-n) = \sum_{m=0}^{n_b} b_m u(t-m) \tag{5-17}$$

If the system is proper, $n_a \geq n_b$, then (5-16) and (5-17) can be written under their state space representation form as, respectively,

$$\begin{aligned} \frac{dx(t)}{dt} &= Ax(t) + Bu(t) \\ y(t) &= Cx(t) + Du(t) \end{aligned} \tag{5-18}$$

and

$$\begin{aligned} x(t+1) &= Ax(t) + Bu(t) \\ y(t) &= Cx(t) + Du(t) \end{aligned} \tag{5-19}$$

where $x(t) \in \mathbb{R}^{n_a}$ is the state vector (Kailath, 1980). The parameters $A \in \mathbb{R}^{n_a \times n_a}$, $B \in \mathbb{R}^{n_a \times 1}$, $C \in \mathbb{R}^{1 \times n_a}$, and $D \in \mathbb{R}$ of the state space equations (5-18) and (5-19) can easily be related to the a_n and b_m coefficients of Eqs. (5-16) and (5-17) (see Exercise 5.6).

5.2 PLANT MODELS

The parametric model that will be used mostly throughout this book is a *rational form*

$$G(\Omega, \theta) = \frac{B(\Omega, \theta)}{A(\Omega, \theta)} = \frac{\sum_{r=0}^{n_b} b_r \Omega^r}{\sum_{r=0}^{n_a} a_r \Omega^r} \tag{5-20}$$

where $\Omega = s$ for lumped continuous-time systems, $\Omega = z^{-1}$ for discrete-time systems, $\Omega = \sqrt{s}$ for diffusion phenomena, $\Omega = \tanh(\tau_R s)$ for commensurate microwave devices, and with $\theta \in \mathbb{R}^{n_\theta}$ the vector of the plant model parameters

$$\theta^T = [a_0 a_1 \ldots a_{n_a} b_0 b_1 \ldots b_{n_b}] \tag{5-21}$$

The reason for this is that it is very easy to get good starting values for (5-20) (see Chapter 7). For lumped continuous-time and discrete-time systems, (5-20) is obtained by taking the Laplace and Z-transform of (5-17) and (5-16) respectively, assuming that the initial conditions are zero. For large order systems (typically $n_a, n_b > 30$) parameterization (5-20) becomes numerically unstable (leads to ill-conditioned normal equations, see Chapter 7), thus requiring other representations to be used.

In modal analysis (Ewins, 1991) and nuclear magnetic resonance modeling (see Section 5.4) a *partial fraction expansion* of (5-20) is often used. Assuming that $G(\Omega, \theta)$ has simple poles, it has the form (Henrici, 1974)

$$G(\Omega, \theta) = \sum_{\substack{r = -p \\ r \neq 0}}^{p} \frac{L_r}{\Omega - \lambda_r} + \sum_{r=1}^{q} \frac{S_r}{\Omega - \sigma_r} \tag{5-22}$$

for strictly proper ($n_b < n_a$) continuous-time models ($\Omega = s$, $\sqrt{s}$ or $\tanh(\tau_R s)$) and

$$G(z^{-1}, \theta) = \sum_{\substack{r = -p \\ r \neq 0}}^{p} \frac{L_r z^{-1}}{1 - \lambda_r z^{-1}} + \sum_{r = 1}^{q} \frac{S_r z^{-1}}{1 - \sigma_r z^{-1}} \tag{5-23}$$

for proper ($n_b \leq n_a$) discrete-time models with $b_0 = 0$ (see Exercise 5.5). In both cases we have $L_{-r} = \bar{L}_r$, $\lambda_{-r} = \bar{\lambda}_r$ and $S_r, \sigma_r \in \mathbb{R}$ with $2p + q = n_a$ so that

$$\theta^T = [\sigma_1 \ldots \sigma_q \mathrm{Re}(\lambda_1)\mathrm{Im}(\lambda_1) \ldots \mathrm{Re}(\lambda_p)\mathrm{Im}(\lambda_p) S_1 \ldots S_q \mathrm{Re}(L_1)\mathrm{Im}(L_1) \ldots \mathrm{Re}(L_p)\mathrm{Im}(L_p)] \tag{5-24}$$

Because parameterizations (5-22) and (5-23) are numerically more stable than (5-20) (except in the case of poles of multiplicity larger than one), one could think of using these models to identify high-order systems (typically $n_a, n_b > 30$). In practice, these representations are not really helpful because the starting values, generated by using parameterization (5-20), are of insufficient quality for higher order systems resulting in poor transfer function estimates (5-22) and (5-23) (one gets stuck in a local minimum). The disadvantage of parameterizations (5-22) and (5-23) is that they do not allow the choice of the order n_b of the numerator polynomial of $G(\Omega, \theta)$. The advantage is that they can deal very easily with constraints on the residues and the poles (see Section 5.4).

An alternative solution for high-order systems is to *factorize* transfer function (5-20) in its poles and zeros. Assuming that $G(\Omega, \theta)$ has simple poles and zeros, we get

$$G(\Omega, \theta) = K \frac{\prod_{r=1}^{n_b} (\Omega - \zeta_r)}{\prod_{r=1}^{n_a} (\Omega - \lambda_r)}$$

However, this representation suffers from the same problems as (5-22) and (5-23): (i) starting values should be generated via (5-20), and (ii) it leads to ill-conditioned normal equations if the true plant model contains multiple poles and/or zeros. Note that the latter is not the case for parameterization (5-20).

To handle high-order systems (typical $n_a, n_b > 30$) the numerator and denominator polynomials of the transfer function (5-20) are expanded in *scalar or vector orthogonal polynomials* (see Section 13.11 and Exercise 1.13)

$$G(\Omega, \theta) = \frac{B(\Omega, \theta)}{A(\Omega, \theta)} = \frac{\sum_{r=0}^{n_q} b_r q_r(\Omega)}{\sum_{r=0}^{n_p} a_r p_r(\Omega)} \tag{5-25}$$

For *scalar orthogonal polynomials* we have $n_p = n_a$, $n_q = n_b$ and $p_r(\Omega)$, $q_r(\Omega)$ are polynomials of order r; for *vector orthogonal polynomials* $b_r = a_r$, $n_q = n_p = n_a + n_b + 1$ and $p_r(\Omega)$, $q_r(\Omega)$ are polynomials of increasing order with $p_{n_p}(\Omega)$, $q_{n_q}(\Omega)$ polynomials of order n_a, n_b, respectively. These are chosen such that they maximize the numerical stability of the model (minimize the condition number of the normal equations, see Chapter 7).

The *state space representation* form of a proper ($n_b \leq n_a$) transfer function (5-20) is

$$G(s, \theta) = C(sI_{n_a} - A)^{-1} B + D \tag{5-26}$$

for lumped continuous-time systems and

$$G(z^{-1}, \theta) = z^{-1}C(I_{n_a} - z^{-1}A)^{-1}B + D \tag{5-27}$$

for discrete-time systems. Equations (5-26) and (5-27) are obtained by taking the Laplace and Z-transform of (5-19) and (5-18) respectively, assuming that the initial conditions are zero. In both cases we have $A \in \mathbb{R}^{n_a \times n_a}$, $B \in \mathbb{R}^{n_a \times 1}$, $C \in \mathbb{R}^{1 \times n_a}$, and $D \in \mathbb{R}$, so that

$$\theta^T = [\text{vec}^T(A)\ B^T\ C\ D] \tag{5-28}$$

The disadvantages of the state space representation are that it does not exist for improper systems ($n_b > n_a$) and that it does not allow one to choose the order n_b of the numerator polynomial of $G(s, \theta)$. The advantage is that it allows straightforward extension to multivariable systems (see Section 5.6).

A *time delay* can be added to transfer function models (5-20), (5-22), (5-23), (5-25), (5-26), and (5-27). For example, for continuous-time models ($\Omega = s$, $\sqrt{s}$ or $\tanh(\tau_R s)$) (5-20) becomes

$$G(\Omega, \theta) = e^{-\tau s}\frac{B(\Omega, \theta)}{A(\Omega, \theta)} = e^{-\tau s}\frac{\sum_{r=0}^{n_b} b_r \Omega^r}{\sum_{r=0}^{n_a} a_r \Omega^r} \tag{5-29}$$

and for discrete-time models

$$G(z^{-1}, \theta) = z^{-\tau/T_s}\frac{B(z^{-1}, \theta)}{A(z^{-1}, \theta)} = z^{-\tau/T_s}\frac{\sum_{r=0}^{n_b} b_r z^{-r}}{\sum_{r=0}^{n_a} a_r z^{-r}} \tag{5-30}$$

where $\tau \in \mathbb{R}$ is an arbitrary time delay (not necessarily an integer multiple of the sampling period T_s). Then the vector of the model parameters θ also contains the delay τ.

5.3 RELATION BETWEEN THE INPUT-OUTPUT DFT SPECTRA

In this section we establish the relationship between the DFTs of the sampled input and output signals of a linear dynamic system

$$U(k) = \frac{1}{\sqrt{N}}\sum_{t=0}^{N-1} u(tT_s)z_k^{-t},\ Y(k) = \frac{1}{\sqrt{N}}\sum_{t=0}^{N-1} y(tT_s)z_k^{-t} \text{ with } z_k = e^{j2\pi k/N} \tag{5-31}$$

and the transfer function models $G(\Omega, \theta)$ of Section 5.2. We start with periodic excitation signals, proceed with arbitrary signals, and finally handle the case where data samples are missing at the input and/or output signals. For the continuous-time systems ($\Omega = s$, $\sqrt{s}$ or $\tanh(\tau_R s)$) we will assume that the excitation is band limited.

5.3.1 Models for Periodic Signals

Assume that we apply a periodic signal $u(t)$ with harmonically related frequencies hf_0, $h \in \mathbb{H} \subset \mathbb{N}$, and period $T_0 = 1/f_0$ to the system and that we observe the steady-state response during an integer number of periods $NT_s = MT_0$ with $M \in \mathbb{N}$. If the excitation is band limited (continuous-time systems) or piecewise constant (discrete-time systems), then the ratio of the output to the input DFT spectra at the excited frequency lines $k = Mh$, $h \in \mathbb{H}$, gives the true transfer function

$$Y(k) = G(\Omega_k, \theta)U(k) \tag{5-32}$$

where $\Omega_k = s_k$, z_k^{-1}, $\sqrt{s_k}$ or $\tanh(\tau_R s_k)$ with $s_k = j\omega_k$ and $z_k = e^{j\omega_k T_s}$, and where $G(\Omega, \theta)$ can take any parameterization of Section 5.2 (Brigham, 1974; Oppenheim et al., 1997). For single sine excitations (5-32) is valid at arbitrary (not related to a DFT grid) frequencies.

5.3.2 Models for Arbitrary Signals

5.3.2.1 Introduction. Spectral leakage occurs in the calculation of the DFT of nonperiodic signals and of periodic signals observed at a noninteger number of periods (see Section 2.2.3 and Brigham, 1974). For these signals, relationship (5-32) is no longer valid and should, therefore, be generalized. We will show that the DFT spectra $Y(k)$, $U(k)$ satisfy an extended transfer function model that includes the beginning and end effects of the data record (see Figure 2-25 on page 59). The relationship is exact, without any approximation for discrete-time systems, and is approximately valid within some spectral alias errors for lumped continuous-time systems.

5.3.2.2 The Extended Transfer Function Model. The DFT spectra $Y(k)$, $U(k)$ of the observed samples $y(t)$, $u(t)$, $t = 0, T_s, \ldots, (N-1)T_s$ satisfy

$$A(s_k, \theta)Y(k) = B(s_k, \theta)U(k) + I(s_k, \theta) + \Delta(s_k) \tag{5-33}$$

$$A(z_k^{-1}, \theta)Y(k) = B(z_k^{-1}, \theta)U(k) + I(z_k^{-1}, \theta) \tag{5-34}$$

where the polynomial $I(\Omega, \theta) = \sum_{r=0}^{n_i} i_r \Omega^r$ ($\Omega = z^{-1}, s$) with $n_i = \max(n_a, n_b) - 1$ is independently parameterized of the plant model parameters (5-21) (proof: see Appendix 5.B). The coefficients i_r are a linear function of the difference between the initial and final conditions of the system and, therefore, will be called the equivalent initial conditions. The term $\Delta(s_k)$ in (5-33) represents the residual alias error and is present even if the signals have been low-pass filtered before sampling. Dividing (5-33), (5-34) by $A(\Omega_k, \theta)$ gives the extended transfer function models

$$Y(k) = G(s_k, \theta)U(k) + T(s_k, \theta) + \delta(s_k) \tag{5-35}$$

$$Y(k) = G(z_k^{-1}, \theta)U(k) + T(z_k^{-1}, \theta) \tag{5-36}$$

where $G(\Omega, \theta)$ and $T(\Omega, \theta)$, with $\Omega = s$ or z^{-1}, can take any parameterization of Section 5.2. $T(\Omega, \theta)$ is called the plant transient term.

For the *rational form* representation $G(\Omega, \theta)$ is as in (5-20) and

$$T(\Omega, \theta) = \frac{I(\Omega, \theta)}{A(\Omega, \theta)} = \frac{\sum_{r=0}^{n_i} i_r \Omega^r}{\sum_{r=0}^{n_a} a_r \Omega^r} \tag{5-37}$$

where $i_0 i_1 \ldots i_{n_i}$ is added to θ (5-21), for the *partial fraction expansion* $G(\Omega, \theta)$ is as in (5-22), (5-23) and

$$T(s, \theta) = \sum_{\substack{r=-p \\ r \neq 0}}^{p} \frac{l_r}{s - \lambda_r} + \sum_{r=1}^{q} \frac{s_r}{s - \sigma_r} \tag{5-38}$$

$$T(z^{-1}, \theta) = \sum_{\substack{r=-p \\ r \neq 0}}^{p} \frac{l_r}{1 - \lambda_r z^{-1}} + \sum_{r=1}^{q} \frac{s_r}{1 - \sigma_r z^{-1}} \tag{5-39}$$

where $s_1 \ldots s_q \mathrm{Re}(l_1)\mathrm{Im}(l_1) \ldots \mathrm{Re}(l_p)\mathrm{Im}(l_p)$ is added to θ (5-24), for the *orthogonal polynomials* $G(\Omega, \theta)$ is as in (5-25) and

$$T(\Omega, \theta) = \frac{\sum_{r=0}^{n_r} i_r r_r(\Omega)}{\sum_{r=0}^{n_p} a_r p_r(\Omega)} \tag{5-40}$$

For *scalar orthogonal polynomials* $n_p = n_a$, $n_q = n_b$, $n_r = n_i$, and $p_r(\Omega)$, $q_r(\Omega)$, $r_r(\Omega)$ are polynomials of order r; for *vector orthogonal polynomials* $a_r = b_r = i_r$, $n_p = n_q = n_r = n_a + n_b + n_i + 2$ and $p_r(\Omega)$, $q_r(\Omega)$, $r_r(\Omega)$ are polynomials of increasing order with $p_{n_p}(\Omega)$, $q_{n_q}(\Omega)$, $r_{n_r}(\Omega)$ polynomials of order n_a, n_b, n_i respectively. These are chosen such that they maximize the numerical stability of the model (minimize the condition number of the normal equations, see Chapter 7). Finally, for the *state space representation* $G(\Omega, \theta)$ is as in (5-26), (5-27) and

$$T(s, \theta) = C(sI_{n_a} - A)^{-1} x_I \tag{5-41}$$

$$T(z^{-1}, \theta) = C(I_{n_a} - z^{-1}A)^{-1} x_I \tag{5-42}$$

where $x_I \in \mathbb{R}^{n_a}$ is added to θ (5-28) (proof: see Appendix 5.C).

The convergence rate to zero of the transient term $T(\Omega_k, \theta)$ and the alias term $\delta(s_k)$ in the extended transfer function models (5-35) and (5-36) is established in the following two lemmas.

Lemma 5.5 (Convergence Rate $T(\Omega_k, \theta)$): For bounded excitations $u(t)$ (bounded excitations $u(t)$ with finite left $(n_b - 1)$th order derivative) applied to stable plants or unstable plants captured within a stabilizing feedback loop, the transient term $T(z_k^{-1}, \theta)$

$(T(s_k, \theta))$ tends to zero as $O(N^{-1/2})$. For bounded random excitations $T(z_k^{-1}, \theta)$ is an $O_{\text{a.s.}}(N^{-1/2})$.

Proof. See Appendix 5.D. □

Lemma 5.6 (Convergence Rate $\delta(s_k)$): Consider band-limited periodic signals, $U(j\omega) = 0$ for $|\omega| > \omega_B$, and band-limited random signals, $S_{uu}(j\omega) = 0$ for $|\omega| > \omega_B$, with $\omega_B < \omega_s/2$. Assume furthermore that these signals have finite nonzero power

$$\frac{1}{NT_s} \int_{-NT_s/2}^{+NT_s/2} \mathcal{E}\{x^2(t)\} dt = O(N^0) > 0 \tag{5-43}$$

for any N, ∞ included. The residual alias error $\delta(s_k)$ tends to zero as $O(N^{-1/2})$ for band-limited periodic excitations, and $O_{\text{m.s.}}(N^{-1/2})$ for band-limited random excitations with differentiable power spectrum $S_{uu}(j\omega)$ $(dS_{uu}(j\omega)/d\omega < \infty$ for $|\omega| \le \omega_B)$.

Proof. See Appendix 5.F. □

Using Lemmas 5.5 and 5.6, we can calculate how fast the extended transfer function models (5-35) and (5-36) tend to the transfer function model (5-32) as $N \to \infty$.

Lemma 5.7 (Convergence Rate Extended Transfer Function Models): Under the assumptions of Lemma 5.5, the convergence rates of discrete-time model (5-36) to (5-32) are $O(N^{-1/2})$ for normalized periodic signals (see Definition 3.4, $F = O(N)$), $O(N^{-1})$ at the excited DFT frequencies for periodic signals with a fixed number of frequencies $(F = O(N^0))$, and $O_{\text{a.s.}}(N^{-1/2})$ for random excitations with differentiable power spectrum. Under the assumptions of Lemmas 5.5 and 5.6, the convergence rates of continuous-time model (5-35) to (5-32) are $O(N^{-1/2})$ for normalized periodic signals (see Definition 3.4, $F = O(N)$), $O(N^{-1})$ at the excited DFT frequencies for periodic signals with a fixed number of frequencies $(F = O(N^0))$, and $O_{\text{p}}(N^{-1/2})$ for random excitations with differentiable power spectrum and $O(N^{-1/2})$.

Proof. It follows directly from Lemmas 5.5 and 5.6 and the fact that the DFT spectrum of band-limited signals with finite nonzero power is $O(N^0)$ for random signals, $O(N^0)$ for normalized periodic signals, and $O(N^{1/2})$ at the excited DFT frequencies for periodic signals with a fixed number of frequencies. □

5.3.2.3 Discussion. The extended transfer models (5-33) and (5-34) show that the leakage errors on the input and output DFT spectra can be modeled by a polynomial and are, in fact, an initial condition (transient) problem. This is illustrated in Figure 2-25 on page 59. The difference from time domain identification is that the equivalent initial conditions take into account the beginning as well as the end effects of the finite data record. In the time domain the initial conditions remain the same as the number of data N increases, whereas in the frequency domain they vary with N (not only due to the scaling factor $N^{-1/2}$ but also due to the varying final conditions of the experiment). Asymptotically $(N \to \infty)$, the extended transfer function models (5-33) and (5-34) reduce to (5-32) (Lemma 5.7).

Lemma 5.7 shows that the classical transfer function model $Y(k) = G(s_k, \theta)U(k)$ contains no asymptotic $(N \to \infty)$ approximation errors in the complete frequency band from DC to Nyquist for band-limited input signals with finite nonzero power.

The transient term $T(\Omega, \theta)$ is zero if the initial and final conditions of the experiment are the same (see Appendix 5.B, Eqs. (5-84) and (5-91)). This is the case for periodic signals observed during an integer number of periods and for time-limited signals. For the band-limited versions of these signals the alias term $\delta(s_k)$ is also zero.

From Lemmas 5.5 and 5.6 it follows that the transient term $T(s_k, \theta)$ and the alias error $\delta(s_k)$ tend to zero at the same rate. Hence, $\delta(s_k)$ cannot be neglected w.r.t. $T(s_k, \theta)$, even for "large" values of N. However, practice has shown that the alias error $\delta(s_k)$ can be approximated well by a polynomial (Pintelon and Schoukens, 1997b). Therefore, to reduce $\delta(s_k)$ in (5-35), the order of the polynomial $I(s, \theta)$ is increased: $n_i \geq \max(n_a, n_b) - 1$.

5.3.3 Models for Records with Missing Data

5.3.3.1 Introduction. Because of temporary sensor failure and/or data transmission errors, it may happen that data samples are missing in the measured signals. The best thing to do then is to throw away the data set and to repeat the experiment. This is not always possible because, for example, the experiment is too expensive, or some of the data are collected in an irregular way using laboratory analysis. Sometimes the output is sampled at a lower rate than the input, which is a periodic missing output data problem (Goodwin and Adams, 1994; Albertos et al., 1999). Treating the missing data as unknown parameters, a generalized version of the extended transfer models (5-35) and (5-36) is constructed. It can handle missing input and/or output data and does not assume any particular missing data pattern.

5.3.3.2 The Extended Transfer Function Model. For simplicity of notation we will assume that M_u consecutive input samples starting at $t = K_u T_s$ and M_y consecutive output samples starting at $t = K_y T_s$ are missing. The sets $\mathbb{M}_u$ and $\mathbb{M}_y$ describing the time instances of the missing input and output samples are then

$$\mathbb{M}_x = \{K_x, K_x + 1, \ldots, K_x + M_x - 1\} \tag{5-44}$$

where $x = u, y$. Define $x^{\mathrm{m}}(tT_s)$, $t = 0, 1, \ldots, N-1$, as the data set where the missing samples are replaced by zeros

$$x^{\mathrm{m}}(tT_s) = \begin{cases} 0 & t \in \mathbb{M}_x \\ x(tT_s) & \text{elsewhere} \end{cases} \tag{5-45}$$

and $X^{\mathrm{m}}(k)$ as the corresponding DFT spectrum ($X = U, Y$ and $x = u, y$). The DFT spectra $Y^{\mathrm{m}}(k)$, $U^{\mathrm{m}}(k)$ of the observed samples (missing data sets) $y^{\mathrm{m}}(t)$, $u^{\mathrm{m}}(t)$, $t = 0, T_s, \ldots, (N-1)T_s$ satisfy

$$\begin{aligned} A(s_k, \theta)Y^{\mathrm{m}}(k) = {} & B(s_k, \theta)U^{\mathrm{m}}(k) + I(s_k, \theta) + \\ & z_k^{-K_u}B(s_k, \theta)I_u(z_k^{-1}, \psi) - z_k^{-K_y}A(s_k, \theta)I_y(z_k^{-1}, \psi) + \Delta(s_k) \end{aligned} \tag{5-46}$$

$$\begin{aligned} A(z_k^{-1}, \theta)Y^{\mathrm{m}}(k) = {} & B(z_k^{-1}, \theta)U^{\mathrm{m}}(k) + I(z_k^{-1}, \theta) + \\ & z_k^{-K_u}B(z_k^{-1}, \theta)I_u(z_k^{-1}, \psi) - z_k^{-K_y}A(z_k^{-1}, \theta)I_y(z_k^{-1}, \psi) \end{aligned} \tag{5-47}$$

where the polynomials $I_x(z^{-1}, \psi) = N^{-1/2}\sum_{t=0}^{M_x-1} x(K_x + t)z^{-t}$, $x = u, y$, contain the missing data and ψ is the parameter vector of the missing samples

$$\psi^T = [u(K_uT_s)\dots u((K_u + M_u - 1)T_s)y(K_yT_s)\dots y((K_y + M_y - 1)T_s)] \tag{5-48}$$

(proof: see Appendix 5.G). Note that models (5-46), (5-47) are bilinear in the parameters θ, ψ. Dividing (5-46), (5-47) by $A(\Omega_k, \theta)$ gives the extended transfer function models

$$Y^{\mathrm{m}}(k) = G(s_k, \theta)U^{\mathrm{m}}(k) + T(s_k, \theta) + z_k^{-K_u}G(s_k, \theta)I_u(z_k^{-1}, \psi) - z_k^{-K_y}I_y(z_k^{-1}, \psi) + \delta(s_k) \tag{5-49}$$

$$Y^{\mathrm{m}}(k) = G(z_k^{-1}, \theta)U^{\mathrm{m}}(k) + T(z_k^{-1}, \theta) + z_k^{-K_u}G(z_k^{-1}, \theta)I_u(z_k^{-1}, \psi) - z_k^{-K_y}I_y(z_k^{-1}, \psi) \tag{5-50}$$

where $G(\Omega, \theta)$ and $T(\Omega, \theta)$ can take any parameterization of Sections 5.2 and 5.3.2 and where the alias error $\delta(s_k)$ has the same properties as in Section 5.3.2. The generalization of (5-49) and (5-50) to the case where data are missing at more than one place is straightforward (see Exercise 5.7).

5.4 MODELS FOR DAMPED (COMPLEX) EXPONENTIALS

In some applications an impulse excitation is applied to the system and only the free decay response is observed, which consists of the sum of (complex) exponentially damped cosines. For real strictly proper lumped continuous-time systems ($\theta \in \mathbb{R}^{n_\theta}$ in (5-37)) with simple complex conjugate pole pairs, it has the form

$$y(t) = 2\sum_{r=1}^{n_a} a_r e^{-d_r(t+\tau)}\cos(\omega_r(t+\tau) + \phi_r) \tag{5-51}$$

with $a_r \in \mathbb{R}^+$ the amplitude, $d_r \in \mathbb{R}^+$ the decay, $\omega_r \in \mathbb{R}^+$ the angular frequency, and $\phi_r \in \mathbb{R}$ the phase of the rth exponentially damped cosine. τ is the (known) delay between the beginning of the free decay experiment and the start of the observations. In modal analysis (5-51) is parameterized in the resonant angular frequency $\omega_0 = \sqrt{d^2 + \omega^2}$ and the damping coefficient $\zeta = d/\omega_0$, while in circuit theory the resonant angular frequency ω_0 and the quality factor $Q = 1/(2\zeta)$ are used. For complex strictly proper lumped continuous-time systems ($\theta \in \mathbb{C}^{n_\theta}$ in (5-37)) with simple complex poles the response is

$$y(t) = \sum_{r=1}^{n_a} a_r e^{j\phi_r} e^{(-d_r + j\omega_r)(t+\tau)} \tag{5-52}$$

Examples of (5-51) and (5-52) are, respectively, impact testing in modal analysis (Ewins, 1991) and nuclear magnetic resonance (NMR) measurements (Kumaresan et al., 1990). In the first application the mechanical structure is excited with an impulse, and the free decay response of the structure, for example, the displacement or the acceleration, is measured at a given location. In the second application the free decay responses of a magnetic field in two orthogonal directions are combined into one complex signal.

The DFT spectrum $Y(k)$ of the free decay response $y(t)$ of a strictly proper lumped continuous-time system or a proper discrete-time system with $b_0 = 0$ satisfies

$$Y(k) = T(z_k^{-1}, \theta) \tag{5-53}$$

where $T(z^{-1}, \theta)$ is the rational function (5-37) with $n_i = n_a - 1$ (proof: see Appendix 5.H). $T(z^{-1}, \theta)$ can also be parametrized as in (5-39), (5-40), and (5-42). The parameters of the free decay responses (5-51) and (5-52) can easily be related to the parameters of the partial fraction expansion (5-39) with $q = 0$ ($\lambda_{-r} \neq \bar{\lambda}_r$ and $l_{-r} \neq \bar{l}_r$ for complex systems). In both cases we have

$$a_r e^{j\phi_r} = \frac{l_r\sqrt{N}}{\lambda_r^{\tau/T_s}(1-\lambda_r^N)}, \quad -d_r + j\omega_r = \frac{1}{T_s}\ln(\lambda_r) \tag{5-54}$$

(proof: see Appendix 5.I).

In NMR measurements the response is typically of the form (5-52) where each term corresponds to the response of a particular chemical substance in a (human) tissue. The amplitude a_k is a measure of the concentration of the substance. Often it is known that a particular substance with known frequency f_r is present in the tissue. Sometimes the chemical structure of the substance imposes the ratio of some amplitudes. All this prior information results in parameter constraints that can easily be taken into account in the partial fraction expansion (5-39). This is not the case for the other parameterizations, which explains why representation (5-39) is popular in NMR modeling. Parameterization (5-37) is appropriate for obtaining starting values for (5-39).

5.5 IDENTIFIABILITY

Loosely speaking, a parametric model $M(\theta, Z)$ is identifiable when the parameters θ can be estimated uniquely using the data Z. It requires that the data are informative enough to distinguish between different models (= condition on the experiment) and that different parameter values give different models (= condition on the model structure). More formally, the identifiability concept can be defined as follows.

Definition 5.8 (Identifiability): A model $M(\theta, Z)$, with θ the model parameters and Z the data, is identifiable at θ_1 if for any θ in a (possibly small) neighborhood of θ_1, $M(\theta, Z) = M(\theta_1, Z)$ implies that $\theta = \theta_1$.

Note that Definition 5.8 gives a definition of *local identifiability*. If the implication in Definition 5.8 is valid for almost all θ and θ_1 values, then one has *global identifiability* (see Ljung, 1999 for a detailed discussion of this issue). In this section we give necessary conditions for the identifiability of the transfer function models of Section 5.3. These conditions can be split into constraints on the parameters θ (identifiable parametrization) and constraints on the input signal (persistent excitation).

5.5.1 Models for Periodic Signals

The identifiability of transfer function model (5-32) depends on the particular parameterization of $G(\Omega, \theta)$. The *rational forms* (5-20) and (5-25) are not identifiable because replacing θ by $\lambda\theta$, with $\lambda \in \mathbb{R}_0$, results in the same input-output description: $G(\Omega, \lambda\theta) = G(\Omega, \theta)$. This parameter ambiguity is removed by constraining the model parameters, for example, $\theta_{[1]} = 1$ or $\|\theta\|_2 = 1$. For transfer functions with a time delay (5-29) and (5-30), the parameter ambiguity is removed by constraining the numerator and denominator coefficients, but not the delay. The *partial fraction expansions* (5-22) and (5-23) contain no parameter ambiguities and, hence, are identifiable. Replacing (A, B, C, D) by

$(TAT^{-1}, TB, CT^{-1}, D)$ in the *state space representations* (5-26) and (5-27) with $T \in \mathbb{R}^{n_a \times n_a}$, a regular matrix $(\det(T) \neq 0)$, leaves $G(\Omega, \theta)$ unchanged. This parameter ambiguity is removed by imposing n_a^2 constraints on θ, which leads to the so-called identifiable state space representations (Van Overbeek and Ljung, 1982). Besides possible constraints on θ, the identifiability of transfer function model (5-32) also puts conditions on the DFT spectrum $U(k)$ of the input signal.

Theorem 5.9 (Identifiability Transfer Function Model (5-32)): Transfer function model (5-32), parameterized as in (5-20) and (5-25) with, for example, constraint $a_{n_a} = 1$, is identifiable if and only if

1. The polynomials $A(\Omega, \theta)$ and $B(\Omega, \theta)$ have no common roots.
2. The input DFT spectrum $U(k)$ is different from zero for at least $(n_a + n_b + 1)/2$ different DFT frequencies, where DC $(k = 0)$ and Nyquist $(N/2)$, each, counts for $1/2$.

Proof. See Appendix 5.K. □

With appropriate additional assumptions on $G(\Omega, \theta)$, Theorem 5.9 also applies for the other parameterizations. For example, the partial fraction expansions (5-22) and (5-23) assume that $G(\Omega, \theta)$ has simple poles. The condition on $U(k)$ is fulfilled, for example, if $u(t)$ consists of the sum of at least $(n_a + n_b + 1)/2$ sine waves. Note that for complex systems, $\theta \in \mathbb{C}^{n_\theta}$, $(n_a + n_b + 1)$ frequencies are required.

5.5.2 Models for Arbitrary Signals

The identifiability of transfer function models (5-35) and (5-36) depends on the particular parameterization of $G(\Omega, \theta)$ and $T(\Omega, \theta)$. The *partial fraction expansions* (5-22), (5-23) and (5-38), (5-39) are identifiable, while the same parameter ambiguities occur as in the periodic case (see Section 5.5.1) for the *rational forms* (5-20), (5-25) and (5-37), (5-40) $(G(\Omega, \lambda\theta) = G(\Omega, \theta),\ T(\Omega, \lambda\theta) = T(\Omega, \theta))$ and the *state space representations* (5-26), (5-27) and (5-41), (5-42) (replacing (A, B, C, D) by $(TAT^{-1}, TB, CT^{-1}, D)$ leaves $G(\Omega, \theta)$ and $T(\Omega, \theta)$ unchanged). Compared with the periodic case, the identifiability of transfer function models (5-35) and (5-36) requires additional conditions on the DFT spectrum, $U(k)$, of the input signal. Necessary conditions for the identifiability of transfer function models (5-35) and (5-36) are

1. The polynomials $A(\Omega, \theta)$, $B(\Omega, \theta)$, and $I(\Omega, \theta)$ have no common roots.
2. The input DFT spectrum $U(k)$ is different from zero for at least $(n_b + n_i + 2)/2$ different DFT frequencies, where DC $(k = 0)$ and Nyquist $(N/2)$, each, counts for $1/2$.
3. $U(k)$ cannot be written as a rational form in Ω_k of order n_i over n_b or less.

(Proof: See Appendix 5.L). □

Note that condition 1 does not exclude $A(\Omega, \theta)$ and $B(\Omega, \theta)$ for having common roots and/or $B(\Omega, \theta)$ and $I(\Omega, \theta)$ for having common roots (see Exercise 5.9). If condition 3 is not fulfilled, then the terms $G(\Omega_k, \theta)U(k)$ and $T(\Omega_k, \theta)$ are indistinguishable. This is, for example, the case when the DFT spectrum $U(k)$ is a constant ($u(t)$ is an impulse (Dirac)).

5.5.3 Models for Records with Missing Data

The identifiability of transfer function models (5-46) and (5-47) depends on the particular parameterization of $G(\Omega, \theta)$ and $T(\Omega, \theta)$, the missing data pattern, and the input DFT spectrum $U^{\mathrm{m}}(k)$. The same parameter constraints should be applied on θ as in Section 5.5.2. A similar analysis, as in Section 5.5.2, gives the following necessary conditions on $U^{\mathrm{m}}(k)$ and the missing data pattern:

1. The polynomials $A(\Omega, \theta)$, $B(\Omega, \theta)$, and $I(\Omega, \theta)$ have no common roots.
2. The input DFT spectrum $U^{\mathrm{m}}(k)$ is different from zero for at least $(n_b + n_i + 2)/2$ different DFT frequencies, where DC ($k = 0$) and Nyquist ($N/2$), each, counts for $1/2$.
3. It is not possible to write $U^{\mathrm{m}}(k)$ as a rational form in Ω_k of order n_i over n_b or less.
4. For discrete-time systems, it is not possible to write $U^{\mathrm{m}}(k) + z_k^{-K_u} I_u(z_k^{-1}, \psi)$ as a rational form in z_k^{-1} of order n_i over n_b or less.
5. For discrete-time systems it is not possible to write $z_k^{-(K_y - K_u)} I_y(z_k^{-1}, \psi)/I_u(z_k^{-1}, \psi)$ as a rational form in z_k^{-1} of order n_b over n_a or larger.

Condition 5 constrains the missing data pattern. For example, discrete-time systems are not identifiable (the condition number of (5-47) is infinitely large) if the input and output samples are missing at the same place, $K_u = K_y$ and the number of consecutive missing samples is larger than or equal to the system order, $M_u, M_y \geq \max(n_a, n_b)$. The missing input and output samples ψ cannot be estimated and the plant model parameters θ should be estimated from the two sets of complete input-output data. Everything happens as if two experiments with full data are available. Section 11.3.4.5 discusses this issue in more detail. Continuous-time systems are still identifiable if $K_u = K_y$ and $M_u = M_y \geq \max(n_a, n_b)$; however, the condition number of model (5-46) increases quickly with the number of consecutive missing samples. For too large an $M_u = M_y$, (5-46) is no longer identifiable within a given finite arithmetic precision (Pintelon and Schoukens, 1999b). The identifiability conditions can easily be extended to the case where data are missing at more than one place.

5.6 MULTIVARIABLE SYSTEMS

The n_y outputs and the n_u inputs of a multivariable system are related by an $n_y \times n_u$ transfer function matrix $G(\Omega, \theta)$ where each entry $G_{[i,j]}(\Omega, \theta)$, $i = 1, 2, \ldots, n_y$ and $j = 1, 2, \ldots, n_u$, is a rational function of Ω ($\Omega = s$, $\sqrt{s}$, $\tanh(\tau_{\mathrm{R}} s)$ or z^{-1}, see Section 5.2). If no relationships exist between the coefficients of the different transfer functions $G_{[i,j]}(\Omega, \theta)$, then the multivariable system is the parallel connection of separate multiple input, single output (MISO) systems. Often, the transfer functions $G_{[i,j]}(\Omega, \theta)$ have the same denominator, for example, in modal analysis (Ewins, 1991) and the two port description of *LC*, *LR*, and *RC* circuits (Balabanian and Bickart, 1969). This leads to the *common denominator* model

$$G(\Omega, \theta) = \frac{B(\Omega, \theta)}{A(\Omega, \theta)} \tag{5-55}$$

where $A(\Omega, \theta) = \sum_{r=0}^{n_a} a_r \Omega^r$ is the common denominator polynomial and $B(\Omega, \theta) = \sum_{r=0}^{n_b} B_r \Omega^r$, with $B_r \in \mathbb{R}^{n_y \times n_u}$, a polynomial matrix.

A natural generalization of the scalar transfer function (5-20) is the so-called matrix-fraction descriptions (Kailath, 1980). Writing the transfer function matrix as a *left matrix fraction* gives

$$G(\Omega, \theta) = A^{-1}(\Omega, \theta)B(\Omega, \theta) \tag{5-56}$$

where $A(\Omega, \theta) = \sum_{r=0}^{n_a} A_r\Omega^r$, with $A_r \in \mathbb{R}^{n_y \times n_y}$, and $B(\Omega, \theta) = \sum_{r=0}^{n_b} B_r\Omega^r$, with $B_r \in \mathbb{R}^{n_y \times n_u}$, are polynomial matrices. Writing the transfer function matrix as a *right matrix fraction* gives

$$G(\Omega, \theta) = B(\Omega, \theta)A^{-1}(\Omega, \theta) \tag{5-57}$$

where $A(\Omega, \theta)$ and $B(\Omega, \theta)$ are, respectively, $n_u \times n_u$ and $n_y \times n_u$ polynomial matrices.

The *partial fraction expansion* of $G(\Omega, \theta)$ has the same form (5-22), (5-23) where each residue matrix L_r, $S_r \in \mathbb{R}^{n_y \times n_u}$ may have a different rank. Sometimes the rank is known beforehand and this should be taken into account in the parameterization. For example, in modal analysis the residue matrices have rank one (Heylen et al., 1997) and are written as $L_r = v_r w_r^T$ with $v_r \in \mathbb{R}^{n_y}$ and $w_r \in \mathbb{R}^{n_u}$ the modal vectors.

The *state space representation* has the same form (5-26), (5-27) with $A \in \mathbb{R}^{n_a \times n_a}$, $B \in \mathbb{R}^{n_a \times n_u}$, $C \in \mathbb{R}^{n_y \times n_a}$, and $D \in \mathbb{R}^{n_y \times n_u}$.

The relation to the input and output DFT spectra and the identifiability issues of the multivariable parametric models are similar to the single input, single output case. For example, the left matrix fraction description (5-56) is made identifiable with the parameter constraint $A_{n_a} = I_{n_a}$. Note that the common denominator (5-55) and the left matrix fraction (5-56) descriptions allow straightforward generalization of the scalar relationships (5-33), (5-34) between the numerator and denominator polynomials of the transfer function model and the input and output DFT spectra. This is important for generating starting values (see Chapter 7). Formula (5-33), (5-34) are then valid with $A(\Omega, \theta)$, $B(\Omega, \theta)$ as defined in (5-55) and (5-56) and $I(\Omega, \theta) = \sum_{r=0}^{n_b} I_r\Omega^r$, $I_r \in \mathbb{R}^{n_y}$, a polynomial vector. This is not the case for the right matrix fraction description (5-57), which can, however, be used if the identification starts from the measured frequency response matrix $G(\Omega_k)$

$$G(\Omega_k) = B(\Omega_k, \theta)A^{-1}(\Omega_k, \theta) \Rightarrow G(\Omega_k)A(\Omega_k, \theta) = B(\Omega_k, \theta)$$

5.7 NOISE MODELS

5.7.1 Introduction

In practice, disturbing noise sources occur everywhere in the measurement setup (see Figure 2-16). The DFT spectra $U(k)$ and $Y(k)$ of the observed input $u(t)$ and output $y(t)$ signals are noisy replicas of the true (unknown) DFT spectra $U_0(k)$ and $Y_0(k)$

$$\begin{aligned} Y(k) &= Y_0(k) + N_Y(k) \\ U(k) &= U_0(k) + N_U(k) \end{aligned} \tag{5-58}$$

where $N_U(k) = \text{DFT}(n_u(t))$ and $N_Y(k) = \text{DFT}(n_y(t))$ are functions of the measurement noise, the process noise, and possibly the generator noise (see Section 2.4). In order to put a quality label (uncertainty bounds) on the measured frequency response function (see

Chapter 2) and the estimated transfer function model (see Chapter 7), we need a model for the disturbing errors $N_U(k)$ and $N_Y(k)$.

5.7.2 Nonparametric Noise Model

As a nonparametric noise model, we will take the (co-)variances of the discrete Fourier transform of the input and output errors

$$\sigma_U^2(k) = \text{var}(N_U(k)),\ \sigma_Y^2(k) = \text{var}(N_Y(k)),\ \sigma_{YU}^2(k) = \text{covar}(N_Y(k), N_U(k)) \tag{5-59}$$

at the DFT frequencies k of interest. It can be obtained via a noise analysis without excitation signal ($r(t) = 0$ in Figure 2-16 on page 44) or via independent, repeated experiments with the same excitation signal $r(t)$. The last approach is strongly recommended because it reduces the total measurement time (the frequency response function and the noise model are measured at the same time) and because the noise model is measured at nominal operating conditions. In practice, the independent, repeated experiments are obtained using periodic signals (see Chapter 8).

5.7.3 Parametric Noise Model

In control applications the input is assumed to be known, $n_u(t) = 0$, and the disturbance $n_y(t)$ is modeled at the sampling instances as filtered white noise $e(t)$

$$n_y(t) = H(q, \theta)e(t) \tag{5-60}$$

with $q = z^{-1}$ the backward shift operator, $e(t)$ a stationary white noise sequence with zero mean and variance σ^2, and

$$H(z^{-1}, \theta) = \frac{C(z^{-1}, \theta)}{D(z^{-1}, \theta)} = \frac{\sum_{r=0}^{n_c} c_r z^{-r}}{\sum_{r=0}^{n_d} d_r z^{-r}} \tag{5-61}$$

The unknown parameters are $c_0, c_1, \ldots, c_{n_c}$, $d_0, d_1, \ldots, d_{n_d}$, and σ. Model (5-60) contains two parameter ambiguities: replacing c_r, d_r and σ by $\lambda_1^{-1}\lambda_2 c_r$, $\lambda_2 d_r$ and $\lambda_1\sigma$, with $\lambda_1, \lambda_2 \neq 0$, leaves (5-60) unchanged ($e(t)$ is multiplied with the same factor as σ). These parameter ambiguities are removed by adding two constraints on the numerator and denominator coefficients of (5-61). In most cases, the choice $d_0 = c_0 = 1$ is made (monic transfer function).

Noise model (5-60) implicitly assumes that the white noise sequences $e(t)$, $t = 0, 1, \ldots, N-1$ are the samples of a piecewise constant continuous-time random variable. This is rarely the case in practice. Think, for example, of the thermal noise generated by resistors (Pyati, 1992) or the flicker and generation-recombination noise generated by semiconductor devices (Lowen and Teich, 1990). However, by increasing the orders n_c and n_d of the polynomials in (5-61) the approximation errors in (5-60) can be made sufficiently small. In general, (5-60) will be only approximately true and any physical interpretation of the results should be done with care.

Taking the DFT of (5-60) gives

$$N_Y(k) = H(z_k^{-1}, \theta)E(k) + T_H(z_k^{-1}, \theta) \tag{5-62}$$

where $T_H(z^{-1}, \theta)$ is the noise transient term,

$$T_H(z^{-1}, \theta) = \frac{J(z^{-1}, \theta)}{D(z^{-1}, \theta)} = \frac{\sum_{r=0}^{n_j} j_r z^{-r}}{\sum_{r=0}^{n_d} d_r z^{-r}} \tag{5-63}$$

$n_j = \max(n_c, n_d) - 1$, and where the coefficients j_r are a function of the initial and final conditions of the noise process (proof: apply (5-36) to (5-60)). Because$T_H(z^{-1}, \theta)$ decreases to zero as $O_{\mathrm{p}}(N^{-1/2})$ and $H(z^{-1}, \theta)E(k)$ is an $O_{\mathrm{a.s.}}(N^0)$ (see Section 14.16), (5-62) is usually approximated by $N_Y(k) = H(z^{-1}, \theta)E(k)$.

The parametric noise model (5-62) can be combined with any plant model of Section 5.3. For example, combining (5-35), (5-36), (5-58), and (5-62) with $N_U(k) = 0$ gives

$$Y(k) = G(s_k, \theta)U(k) + T(s_k, \theta) + H(z_k^{-1}, \theta)E(k) + T_H(z_k^{-1}, \theta) + \delta(s_k) \tag{5-64}$$

$$Y(k) = G(z_k^{-1}, \theta)U(k) + T(z_k^{-1}, \theta) + H(z_k^{-1}, \theta)E(k) + T_H(z_k^{-1}, \theta) \tag{5-65}$$

where $c_1, \ldots, c_{n_c}$, $d_1, \ldots, d_{n_d}$ and possibly $j_0, \ldots, j_{n_j}$ are added to the parameter vector θ. Model (5-65) represents the classical time domain model structures, for example, ARX (AutoRegressive with eXogenous input) for $C(z^{-1}, \theta) = 1$ and $D(z^{-1}, \theta) = A(z^{-1}, \theta)$,

$$\text{ARX:} \qquad Y(k) = \frac{B(z_k^{-1}, \theta)}{A(z_k^{-1}, \theta)}U(k) + \frac{1}{A(z_k^{-1}, \theta)}E(k) + \frac{K(z_k^{-1}, \theta)}{A(z_k^{-1}, \theta)} \tag{5-66}$$

with $K(z^{-1}, \theta) = I(z_k^{-1}, \theta) + J(z_k^{-1}, \theta)$ and $n_k = \max(n_a, n_b) - 1$; ARMAX (AutoregRessive Moving Average with eXogenous input) for $D(z^{-1}, \theta) = A(z^{-1}, \theta)$,

$$\text{ARMAX:} \qquad Y(k) = \frac{B(z_k^{-1}, \theta)}{A(z_k^{-1}, \theta)}U(k) + \frac{C(z_k^{-1}, \theta)}{A(z_k^{-1}, \theta)}E(k) + \frac{K(z_k^{-1}, \theta)}{A(z_k^{-1}, \theta)} \tag{5-67}$$

with $K(z^{-1}, \theta) = I(z_k^{-1}, \theta) + J(z_k^{-1}, \theta)$ and $n_k = \max(n_a, n_b, n_c) - 1$; ARMA (AutoRegressive Moving Average) for $G(z^{-1}, \theta) = 0$ and $T(z^{-1}, \theta) = 0$,

$$\text{ARMA:} \qquad Y(k) = \frac{C(z_k^{-1}, \theta)}{D(z_k^{-1}, \theta)}E(k) + \frac{J(z_k^{-1}, \theta)}{D(z_k^{-1}, \theta)} \tag{5-68}$$

OE (Output Error) for $H(z^{-1}, \theta) = 1$ and $T_H(z^{-1}, \theta) = 0$,

$$\text{OE:} \qquad Y(k) = \frac{B(z_k^{-1}, \theta)}{A(z_k^{-1}, \theta)}U(k) + \frac{I(z_k^{-1}, \theta)}{A(z_k^{-1}, \theta)} + E(k) \tag{5-69}$$

and BJ (Box-Jenkins)

$$\text{BJ:} \qquad Y(k) = \frac{B(z_k^{-1}, \theta)}{A(z_k^{-1}, \theta)}U(k) + \frac{I(z_k^{-1}, \theta)}{A(z_k^{-1}, \theta)} + \frac{C(z_k^{-1}, \theta)}{D(z_k^{-1}, \theta)}E(k) + \frac{J(z_k^{-1}, \theta)}{D(z_k^{-1}, \theta)} \tag{5-70}$$

when the plant $G(z^{-1}, \theta)$ and the noise $H(z^{-1}, \theta)$ models have no common parameters (Ljung, 1999). Note that model (5-64) can be seen as a hybrid Box-Jenkins model: it combines a continuous-time plant model with a discrete-time noise model.

The plant transient $T(z_k^{-1}, \theta)$ and the noise transient $T_H(z_k^{-1}, \theta)$ terms in (5-65) are not always distinguishable (separately identifiable), for example,

$$T(z^{-1}, \theta) + T_H(z^{-1}, \theta) = \frac{I(z^{-1}, \theta) + J(z^{-1}, \theta)}{A(z^{-1}, \theta)} \tag{5-71}$$

for ARX and ARMAX models and only the sum $i_r + j_r$ of the coefficients can be identified. Therefore, we replaced $I(z^{-1}, \theta) + J(z^{-1}, \theta)$ by $K(z^{-1}, \theta)$ in (5-66) and (5-67). For Box-Jenkins models we have

$$T(z^{-1}, \theta) + T_H(z^{-1}, \theta) = \frac{I(z^{-1}, \theta)D(z^{-1}, \theta) + A(z^{-1}, \theta)J(z^{-1}, \theta)}{A(z^{-1}, \theta)D(z^{-1}, \theta)} \tag{5-72}$$

where $T(z_k^{-1}, \theta)$ and $T_H(z_k^{-1}, \theta)$ are distinguishable (i_r and j_r are identifiable) if $A(z^{-1}, \theta)$ and $D(z^{-1}, \theta)$ have no common roots and if $n_b \le n_a$ and $n_c \le n_d$ (see Exercise 5.11). If $A(z^{-1}, \theta)$ and $D(z^{-1}, \theta)$ have common roots then the parameterization should be adapted accordingly (see Exercise 5.12). Although the transient terms $T_H(z^{-1}, \theta)$ and $T(z^{-1}, \theta)$ are often neglected, they can be important, for example, in model validation tests (see Section 10.8.1).

5.8 NONLINEAR SYSTEMS

In this section the response of the nonlinear system $y(t) = G[u(t)]$ (see Figure 5-6) is studied for random phase multisine (see Definition 3.2) and periodic noise (see Definition 3.3) excitations $u(t)$. These are periodic signals with a deterministic (random multisine) or random (periodic noise) amplitude spectrum and a random phase spectrum. The class of nonlinear distortions considered is restricted to the nonlinear systems that can be approximated arbitrarily well in least squares sense by a Volterra series on a given input domain (see Definition 3.5). It makes it possible to describe strongly nonlinear phenomena such as saturation (e.g., amplifiers) and discontinuities (e.g., relays, quantizers).

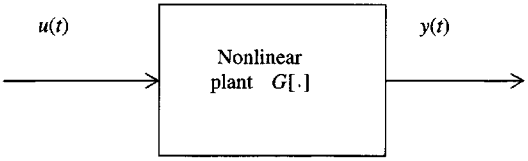


Figure 5-6. Nonlinear system $y(t) = G[u(t)]$.

From Theorem 3.8 it follows that the input-output DFT spectra are related to the best linear approximation $G_R(s)$ by

$$Y(k) = G_R(s_k)U(k) + Y_S(k) \tag{5-73}$$

where $Y_S(k) = G_S(k)U(k)$ with $G_S(k)$ the stochastic contributions to the transfer function. The related linear dynamic system (best linear approximation) $G_R(s_k) = G_0(s_k) + G_B(s_k)$ consists of two parts: the true underlying linear system $G_0(s_k)$ and the bias term $G_B(s_k)$, which depends on the nonlinear distortions and the power spectrum of the input signal

$$\begin{aligned} G_B(s_k) &= \sum\nolimits_{n=2}^{\infty} G_B{}^{2n-1}(s_k) \\ G_B^{2n-1}(s_k) &= \frac{c_n}{N^{n-1}} \sum_{k_1, \ldots, k_{n-1}=1}^{N} G^{2n-1}_{k, -k_1, k_1, \ldots, -k_{n-1}, k_{n-1}} \mathcal{E}\{|U(k_1)|^2 \ldots |U(k_{n-1})|^2\} \\ &\quad + O_n(N^{-1}) \end{aligned} \tag{5-74}$$

with $c_n = 2^{n-1}(2n-1)!!$, $\sum_{n=2}^{\infty} O_n(N^{-1}) = O(N^{-1})$ and where $\mathcal{E}\{.\}$ denotes the expected value with respect to the random amplitudes of the periodic noise (see Theorem 3.7). Multiplying (5-73) by $e^{-j\angle U(k)}$ gives

$$Y(k)e^{-j\angle U(k)} = G_R(s_k)|U(k)| + Y_S(k)e^{-j\angle U(k)} \tag{5-75}$$

Written in this form, it is obvious that the noise term $Y_S(k)e^{-j\angle U(k)}$ is independent of the signal term $G_R(s_k)|U(k)|$. Because $\angle Y_S(k) - \angle U(k) = \angle G_S(k)$, it follows that $\sqrt{N} Y_S(k)e^{-j\angle U(k)}$ has exactly the same stochastic properties as $G_S(k)$ in Theorems 3.9, 3.10 and 3.11: it is mixing of order infinity and is asymptotically ($N \to \infty$) circular complex normally distributed. Both observations motivate the block diagram of Figure 5-7.

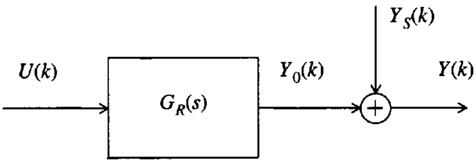


Figure 5-7. Input-output behavior of a nonlinear system excited by a random phase multisine.

5.9 EXERCISES

5.1. Show that the transfer function between the force per unit area and the longitudinal displacement of the clamped beam is given by (5-4). Calculate the poles of (5-4) (hint: values of s such that $\cosh(\tau s) = 0$).

5.2. Calculate the partial fraction expansion (5-5) of transfer function (5-4) (hint: note that $G(\infty) = 0$ and calculate $\sum_{k=-\infty}^{\infty} \frac{R_{2k+1}}{s - s_{2k+1}}$ with R_k the residue of the pole s_k).

5.3. Consider the charging of a capacitor (see Figure 5-8). Show that the transfer function between

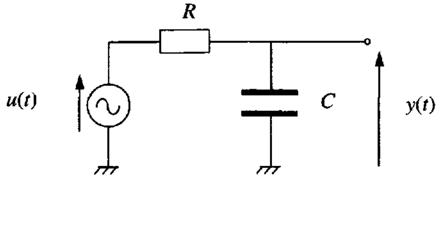


Figure 5-8. Charging of a capacitor with a voltage $u(t)$.

the piecewise constant voltage source $u(t)$ and the sampled voltage $y(t)$, $t = kT_s$, across the capacitor is given by

$$G_{\text{ZOH}}(z^{-1}) = \frac{1 - e^{-T_s/(RC)}}{z - e^{-T_s/(RC)}} \tag{5-76}$$

(hint: first show that $G(s) = 1/(1 + RCs)$).

5.4. Consider the cascade of two continuous-time systems shown in Figure 5-4 and show that discrete-time model (5-11) describes the input-output behavior of the continuous-time model exactly at the sampling times $t = kT_s$ (hint: apply (5-10) on the transfer functions from $r_{\text{zoh}}(t)$ to $u(t)$ and from $r_{\text{zoh}}(t)$ to $y(t)$).

5.5. Show that the partial fraction expansion of a proper ($n_b \le n_a$) discrete-time system $G(z^{-1}, \theta)$ with $b_0 = 0$ is given by (5-23) (hint: multiply the numerator and denominator polynomial of $G(z^{-1}, \theta)$ with z^{n_a} and calculate the partial fraction expansion in z).

5.6. Show that a state space representation of difference equation (5-17) is given by (5-19) with

$$A = \begin{bmatrix} -a_1/a_0 & -a_2/a_0 & \dots & -a_{n_a-1}/a_0 & -a_{n_a}/a_0 \\ 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & \dots & \dots \\ \dots & \dots & \dots & 0 & \dots \\ 0 & \dots & 0 & 1 & 0 \end{bmatrix}, B = \begin{bmatrix} 1/a_0 \\ 0 \\ \dots \\ 0 \end{bmatrix},$$

$$C = \begin{bmatrix} b_1 - a_1 b_0/a_0 \dots b_{n_b} - a_{n_b} b_0/a_0, & -a_{n_b+1} b_0/a_0 \dots -a_{n_a} b_0/a_0 \end{bmatrix} \text{ and } D = b_0/a_0$$

(hint: eliminate the state vector in (5-19)).

5.7. Assume that $M_{u[i]}$ input samples are missing at time instants $t = K_{u[i]}T_s$, $i = 1, 2, \dots, M_{u[i]}$, and $M_{y[j]}$ output samples are missing at time instants $t = K_{y[j]}T_s$, $j = 1, 2, \dots, M_{y[j]}$. Show that the extended transfer function model for discrete-time systems is given by

$$Y^{\text{m}}(k) = G(z_k^{-1}, \theta)U^{\text{m}}(k) + T(z_k^{-1}, \theta) +$$
$$G(z_k^{-1}, \theta)\sum\nolimits_{i=1}^{M_{u[i]}} z_k^{-K_{u[i]}} I_{u[i]}(z_k^{-1}, \psi) - \sum\nolimits_{j=1}^{M_{y[j]}} z_k^{-K_{y[j]}} I_{y[j]}(z_k^{-1}, \psi)$$

where $I_{x[i]}(z^{-1}, \psi) = N^{-1/2}\sum\nolimits_{t=0}^{M_{x[i]}-1} x(K_{x[i]} + t)z^{-t}$, $x = u, y$ (hint: follow the lines of Appendix 5.G).

5.8. Show relation (5-54) for the real case (5-51) (hint: use $\cos(x) = (e^{jx} + e^{-jx})/2$ and follow the lines of Appendix 5.I).

5.9. Show that a_0 in $a_0 y(t) + \frac{dy(t)}{dt} = a_0 u(t) + \frac{du(t)}{dt}$ is identifiable if and only if $y(0\text{-}) \ne u(0\text{-})$. Note that $y(t)$ can never be made different from $u(t)$ if $y(0\text{-}) = u(0\text{-})$. Only internal action in the system can make $y(0\text{-}) \ne u(0\text{-})$.

5.10. Show that the functions $f_r(\Omega_k) = \Omega_k^r U(k)$, $r = 0, 1, \dots, n_b$, and $k = 0, 1, \dots, N/2$ are linearly independent if and only if $U(k) \ne 0$ for at least $(n_b + 1)/2$ DFT frequencies where DC and Nyquist each count for $1/2$ (hint: study $\sum\nolimits_{r=0}^{n_b} \beta_r f_r(\Omega_k) = 0$ at the DFT frequencies where $U(k) \ne 0$).

5.11. Consider model (5-65) where $n_b \le n_a$, $n_c \le n_d$, $A(z^{-1}, \theta)$ and $D(z^{-1}, \theta)$ have no common roots, and $G(z^{-1}, \theta)$, $H(z^{-1}, \theta)$ have respective minimal orders n_b over n_a and n_c over n_d. Show that $T(z^{-1}, \theta)$ and $T_H(z^{-1}, \theta)$ are identifiable (hint: suppose that $A(z^{-1}, \theta)$ and $D(z^{-1}, \theta)$ are given and write $T(z_k^{-1}, \theta) + T_H(z_k^{-1}, \theta)$ as $\sum_{r=0}^{n_i} i_r f_r(z_k^{-1}) + \sum_{r=0}^{n_j} j_r g_r(z_k^{-1})$ with $f_r(z_k^{-1}) = z_k^{-r}/A(z_k^{-1}, \theta)$, $g_r(z_k^{-1}) = z_k^{-r}/D(z_k^{-1}, \theta)$; next show, following the lines of Appendix 5.J, that $f_r(z_k^{-1})$, $g_r(z_k^{-1})$ are independent functions).

5.12. Consider model (5-65) where $n_b \le n_a$, $n_c \le n_d$, $A = \tilde{A}F$, $D = \tilde{D}F$, $\tilde{A}$ and $\tilde{D}$ have no common roots, and G, H have respective minimal orders n_b over n_a and n_c over n_d. Show that

$$G = \frac{B}{\tilde{A}F},\ H = \frac{C}{\tilde{D}F} \text{ and } T + T_H = \frac{1}{F}\left(I_2 + \frac{\tilde{I}}{\tilde{A}} + \frac{\tilde{C}}{\tilde{D}}\right),$$

where I_2, $\tilde{I}$, and $\tilde{C}$ are polynomials in z^{-1} of respective orders $n_f - 1$, $n_a - n_f - 1$, and $n_d - n_f - 1$, is an identifiable parameterization.

5.10 APPENDIXES

Appendix 5.A: Stability and Minimum Phase Regions

To determine the stability regions of the poles we expand the rational form $G(\Omega, \theta)$ in partial fractions. We find

$$G(s, \theta) = \sum_r \frac{L_r}{s - \lambda_r},\ G(z^{-1}, \theta) = \sum_r \frac{L_r}{z - \lambda_r} \text{ and } G(\sqrt{s}, \theta) = \sum_r \frac{L_r}{\sqrt{s} - \lambda_r} \qquad (5\text{-}77)$$

(see Eqs. 5-22 and 5-23). Calculating the impulse responses of $G(\Omega, \theta)$ in (5-77) gives

$$\begin{aligned} g(t) &= L^{-1}\{G(s, \theta)\} = \sum_r L_r e^{\lambda_r t} && \text{(a)} \\ g(n) &= Z^{-1}\{G(z^{-1}, \theta)\} = \sum_r L_r \lambda_r^{(n-1)},\ \text{for } n > 0 && \text{(b)} \qquad (5\text{-}78) \\ g(t) &= L^{-1}\{G(\sqrt{s}, \theta)\} = \sum_r L_r \frac{1}{\sqrt{\pi t}} + \lambda_r e^{\lambda_r^2 t}\operatorname{erfc}(-\lambda_r \sqrt{t}) && \text{(c)} \end{aligned}$$

with $L^{-1}\{\ \}$ the inverse Laplace transform, $Z^{-1}\{\ \}$ the inverse Z-transform, and erfc( ) the complementary error function (Selby, 1973; Spiegel, 1965). It follows that the impulse responses are asymptotically zero (poles are stable) if and only if $\operatorname{Re}(\lambda_r) < 0$ in the s-domain (5-78a), $|\lambda_r| < 1$ in the z-domain (5-78b), and $\operatorname{Re}(\lambda_r^2) < 0$ or $|\operatorname{Re}(\lambda_r)| < |\operatorname{Im}(\lambda_r)|$ in the $\sqrt{s}$-domain (5-78c). By definition, the minimum phase region of the zeros equals the stable region of the poles. □

Appendix 5.B: Relation between DFT Spectra and Transfer Function for Arbitrary Signals

First, the result is proved for discrete-time systems, and next, for lumped continuous-time systems. In both cases we assume that the input and output samples are known (no measurement and no process noise) exactly for $t = 0, 1, \ldots, N-1$ and are unknown elsewhere.

5.B.1 Discrete-Time Systems. The discrete input and output samples satisfy difference equation (5-17) for any t. Taking the one-sided Z-transform of both sides of (5-17) using

$$\sum_{t=0}^{\infty} x(t-n)z^{-t} = z^{-n}(X(z) + X_1(z))$$

where $X(z) = \sum_{t=0}^{\infty} x(t)z^{-t}$ ($X = U, Y$ and $x = u, y$) is the one-sided Z-transform of $x(t)$ and $X_1(z) = \sum_{t=1}^{n} x(-t)z^{t}$, gives

$$A(z^{-1})Y(z) = B(z^{-1})U(z) + I_1(z^{-1}) \qquad (5\text{-}79)$$

$A(z^{-1})$ and $B(z^{-1})$ are, respectively, the denominator and numerator polynomials of the plant transfer function (5-20) and $I_1(z^{-1})$ stands for the influence of the initial conditions of the experiment (past samples of $u(t)$ and $y(t)$)

$$I_1(z^{-1}) = \sum_{m=1}^{n_b}\sum_{t=1}^{m} b_m u(-t)z^{t-m} - \sum_{n=1}^{n_a}\sum_{t=1}^{n} a_n y(-t)z^{t-n} \qquad (5\text{-}80)$$

Model (5-79) cannot be evaluated on the unit circle because the input and output samples for $t \geq N$ are unknown. These samples must, hence, be eliminated. We solve, thereto, difference equation (5-17) for $t = N, N+1, \ldots, \infty$. Multiplying both sides of (5-17) with z^{-t} and making the summation over $t = N, N+1, \ldots, \infty$ using

$$\sum_{t=N}^{\infty} x(t-n)z^{-t} = z^{-n}(\tilde{X}(z) + z^{-N}X_2(z))$$

where $\tilde{X}(z) = \sum_{t=N}^{\infty} x(t)z^{-t}$ and $X_2(z) = \sum_{t=1}^{n} x(N-t)z^{t}$ ($X = U, Y$ and $x = u, y$), gives

$$A(z^{-1})\tilde{Y}(z) = B(z^{-1})\tilde{U}(z) + z^{-N}I_2(z^{-1}) \qquad (5\text{-}81)$$

$I_2(z^{-1})$ stands for the influence of the final conditions (samples of $u(t)$ and $y(t)$ at the end of the experiment)

$$I_2(z^{-1}) = \sum_{m=1}^{n_b}\sum_{t=1}^{m} b_m u(N-t)z^{t-m} - \sum_{n=1}^{n_a}\sum_{t=1}^{n} a_n y(N-t)z^{t-n} \qquad (5\text{-}82)$$

Subtracting (5-81) from (5-79) gives

$$A(z^{-1})Y_N(z) = B(z^{-1})U_N(z) + I_1(z^{-1}) - z^{-N}I_2(z^{-1}) \qquad (5\text{-}83)$$

where $X_N(z) = X(z) - \tilde{X}(z) = \sum_{t=0}^{N-1} x(t)z^{-t}$ $(X = U, Y$ and $x = u, y)$. Evaluation of (5-83) on the unit circle at the DFT frequencies $z_k = \exp(j2\pi k/N)$, taking into account that $z_k^N = 1$, $Y_N(z_k) = N^{1/2}Y(k)$ and $U_N(z_k) = N^{1/2}U(k)$, finally gives

$$A(z_k^{-1})Y(k) = B(z_k^{-1})U(k) + I(z_k^{-1}) \qquad (5\text{-}84)$$

where $I(z^{-1}) = N^{-1/2}(I_1(z^{-1}) - I_2(z^{-1}))$ is a polynomial of order $n_i = \max(n_a, n_b) - 1$. The polynomial $I(z^{-1})$ can be parameterized independent of the numerator and denominator coefficients of $G(z^{-1})$ (5-20) because its coefficients i_r depend, for a given plant model, linearly on $\max(n_a, n_b)$ independent, initial conditions.

5.B.2 Lumped Continuous-Time Systems. The proof follows the same lines as in the previous section. We assume that the excitation $u(t)$ is band limited. The input and output continuous-time signals satisfy differential equation (5-16). Taking the one-sided Laplace transform of (5-16) gives

$$A(s)Y(s) = B(s)U(s) + I_1(s) \qquad (5\text{-}85)$$

where $U(s)$ and $Y(s)$ are the one-sided Laplace transforms of $u(t)$ and $y(t)$, respectively. $A(s)$, $B(s)$ are, respectively, the numerator and denominator polynomials of the plant transfer function (5-20) and $I_1(s)$ represents the influence of the initial conditions (value and derivatives of $u(t)$ and $y(t)$ at $t = 0\text{-}$)

$$I_1(s) = \sum_{n=1}^{n_a} \sum_{r=0}^{n-1} a_n s^{n-r-1} y^{(r)}(0\text{-}) - \sum_{m=1}^{n_b} \sum_{r=0}^{m-1} b_m s^{m-r-1} u^{(r)}(0\text{-}) \qquad (5\text{-}86)$$

The integrals appearing in the model (5-85) cannot be evaluated because the input and output signals are unknown for $t \geq NT_s$. The differential equation (5-16) is, therefore, solved for $t \geq NT_s$ using the one-sided Laplace transform. Multiplying both sides of (5-16) by e^{-st} and integrating from $t = NT_s$ to $t = \infty$ gives

$$A(s)\tilde{Y}(s) = B\tilde{U}(s) + e^{-NT_s s} I_2(s) \qquad (5\text{-}87)$$

where $\tilde{X}(s) = \int_{NT_s}^{\infty} e^{-st}x(t)dt$ $(X = U, Y$ and $x = u, y)$. $I_2(s)$ stands for the influence of the final conditions (value and derivatives of $u(t)$ and $y(t)$ at $t = NT_s\text{-}$)

$$I_2(s) = \sum_{n=1}^{n_a} \sum_{r=0}^{n-1} a_n s^{n-r-1} y^{(r)}(NT_s\text{-}) - \sum_{m=1}^{n_b} \sum_{r=0}^{m-1} b_m s^{m-r-1} u^{(r)}(NT_s\text{-}) \qquad (5\text{-}88)$$

Subtracting (5-87) from (5-85) gives

$$A(s)Y_N(s) = B(s)U_N(s) + I_1(s) - e^{-NT_s s} I_2(s) \qquad (5\text{-}89)$$

where $X_N(s) = X(z) - \tilde{X}(z) = \int_0^{NT_s} e^{-st}x(t)dt$ $(X = U, Y$ and $x = u, y)$. Evaluating (5-89) along the $j\omega$-axis at the DFT frequencies $s_k = j2\pi f_s k/N$ using the relationship between the DFT and the Fourier integral (Brigham, 1974)

$$X(k) = \frac{1}{\sqrt{N}}\sum\nolimits_{t=0}^{N-1} x(tT_s)z_k^{-t} = \frac{1}{T_s\sqrt{N}}\sum\nolimits_{n=-\infty}^{n=+\infty} X_N(s_k - nj\omega_s) \tag{5-90}$$

and taking into account that $e^{-NT_s s_k} = 1$, results in

$$A(s_k)Y(k) = B(s_k)U(k) + I(s_k) + \Delta(s_k) \tag{5-91}$$

$I(s) = N^{-1/2}(I_1(s) - I_2(s))/T_s$ is a polynomial of order $\max(n_a, n_b) - 1$, and $\Delta(s_k)$ is the residual spectral alias error

$$\Delta(s_k) = \frac{1}{T_s\sqrt{N}}\sum_{\substack{n = -\infty \\ n \neq 0}}^{n = +\infty} [B(s_k)U_N(s_k - nj\omega_s) - A(s_k)Y_N(s_k - nj\omega_s)] \tag{5-92}$$

Note that the spectral alias error is due to the piecewise constant approximation of the Fourier integrals $U_N(j\omega)$ and $Y_N(j\omega)$ by the discrete Fourier transforms $U(k)$ and $Y(k)$ (see (5-90)): it is present even if the signals $u(t)$ and $y(t)$ passed through a low-pass filter before sampling. For the same reason as in the previous section, the polynomial $I(s)$ is parameterized independent of the numerator and denominator coefficients of $G(s)$ (5-20).

Appendix 5.C: Parameterizations of the Extended Transfer Function Model

The partial fraction expansions (5-38) and (5-39) follow directly from the fact that $T(\Omega, \theta)$ has the same poles as $G(\Omega, \theta)$. The particular form (5-39) is obtained by rewriting $T(z^{-1}, \theta)$ as $z(z^{-1}T(z^{-1}, \theta))$, where the partial fraction expansion of $z^{-1}T(z^{-1}, \theta)$ has the form (5-23), because the orders of the numerator and the denominator of $z^{-1}T(z^{-1}, \theta)$ are equal (see Exercise 5.5).

The state space equations of a proper $(n_b \leq n_a)$ discrete-time system are given by (5-19). Following the lines of Appendix 5.B, we solve (5-19) for $t = 0, 1, \ldots, \infty$ and for $t = N, N+1, \ldots, \infty$ using the one-sided Z-transform. Using the same notations as in Appendix 5.B, we find

$$Y(z) = G(z^{-1})U(z) + C(I_{n_a} - z^{-1}A)^{-1}x(0) \tag{5-93}$$

$$\tilde{Y}(z) = G(z^{-1})\tilde{U}(z) + z^{-N}C(I_{n_a} - z^{-1}A)^{-1}x(N) \tag{5-94}$$

where $G(z^{-1})$ is given by (5-27). Subtracting (5-94) from (5-93) and evaluating the result at the DFT frequencies $z = z_k$ gives (5-42) with $x_I = N^{-1/2}(x(0) - x(N))$.

The state space equations of a proper $(n_b \leq n_a)$ lumped continuous-time system are given by (5-18). Following the lines of Appendix 5.B, we solve (5-18) for $t \in [0, \infty]$ and for $t \in [NT_s, \infty]$ using the one-sided Laplace transform. Using the same notations as in Appendix 5.B, we find

$$Y(s) = G(s)U(s) + C(sI_{n_a} - A)^{-1}x(0\text{-}) \tag{5-95}$$

$$\tilde{Y}(s) = G(s)\tilde{U}(s) + e^{-NT_s s}C(sI_{n_a} - A)^{-1}x(NT_s\text{-}) \tag{5-96}$$

where $G(s)$ is given by (5-26). Subtracting (5-96) from (5-95) and evaluating the result at the DFT frequencies $s = s_k$ gives (5-41) with $x_I = N^{-1/2}(x(0\text{-}) - x(NT_s\text{-}))/T_s$. □

Appendix 5.D: Convergence Rate of the Equivalent Initial Conditions

We prove the result for a discrete-time system ($\Omega = z^{-1}$). The proof for continuous-time systems ($\Omega = s$) follows the same lines. Using Eqs. (5-80), (5-82), and (5-84) of Appendix 5.B, we find the following relationship between the coefficients of the polynomial $I(z^{-1}, \theta) = \sum_{r=0}^{n_i} i_r z^{-r}$ in the plant model (5-34) and the initial and final conditions of the experiment:

$$I(z^{-1}, \theta) = N^{-1/2}\left(\sum_{m=1}^{n_b}\sum_{t=1}^{m} b_m \Delta_N u(t) z^{t-m} - \sum_{n=1}^{n_a}\sum_{t=1}^{n} a_n \Delta_N y(t) z^{t-n}\right) \tag{5-97}$$

where $\Delta_N x(t) = x(-t) - x(N-t)$ with $x = u, y$. It shows that the coefficients i_r, $r = 0, 1, \ldots, n_i$, of $I(z^{-1}, \theta)$ depend linearly on $2n_a$ output and $2n_b$ input samples (finite number independent of N). A bounded input applied to a stable linear system results in a bounded output; see Kailath (1980). The same is true for unstable plants captured in a stabilizing feedback loop. Therefore, it follows from (5-97) that i_r in (5-37) is an $O(N^{-1/2})$. For bounded random inputs we still have $|i_r| \le C/\sqrt{N}$ with probability one, so that $i_r = O_{\text{a.s.}}(N^{-1/2})$. Because the residues of the poles of a rational function are proportional to its numerator coefficients, the same conclusions hold for the residues l_r and s_r in (5-39). Similar reasoning proves the results for x_I in (5-42) (see Appendix 5.C for explicit expressions of x_I). □

Appendix 5.E: Some Integral Expressions

5.E.1 Definite Integrals Involving sin(x)/x Functions. Using

$$\cos(ax)\sin(bx) = 0.5[\sin((a+b)x) - \sin((a-b)x)] \tag{5-98}$$

$$\int_{-\infty}^{+\infty} \frac{\sin(cx)}{x} dx = \begin{cases} \pi & c > 0 \\ -\pi & c < 0 \end{cases} \tag{5-99}$$

and $0 < a < b$ we find

$$\int_{-\infty}^{+\infty} \frac{\cos(ax)\sin(bx)}{\pi x} dx = \int_{-\infty}^{+\infty} \frac{\sin((a+b)x) - \sin((a-b)x)}{2\pi x} dx = \frac{1}{2} - \left(-\frac{1}{2}\right) = 1 \tag{5-100}$$

Note that (5-100) is zero if $a > b$. Because $\sin(ax)\sin(bx)/(\pi x)$ is a uniformly bounded odd function of x in $[-\infty, +\infty]$, we have

$$\int_{-\infty}^{+\infty} \frac{\sin(ax)\sin(bx)}{\pi x} dx = 0 \tag{5-101}$$

for any value of a and b.

5.E.2 Convergence Rate of Integrals Involving sin(x)/x Functions. In this section we study the convergence rate to zero of

$$\int_{N}^{+\infty} f_1(x)dx,\ \int_{N}^{+\infty} f_2(x)dx,\ \int_{-\infty}^{-N} f_1(x)dx \text{ and } \int_{-\infty}^{-N} f_2(x)dx \tag{5-102}$$

as $N \to \infty$, where

$$\begin{aligned} f_1(x) &= \frac{\cos(ax)\sin(bx)}{\pi x} = \frac{\sin((a+b)x) - \sin((a-b)x)}{2\pi x} \\ f_2(x) &= \frac{\sin(ax)\sin(bx)}{\pi x} = \frac{\cos((a-b)x) - \cos((a+b)x)}{2\pi x} \end{aligned} \tag{5-103}$$

are uniformly bounded functions of x. Because $f_1(x)$ and $f_2(x)$ are, respectively, even and odd functions of x, it follows from (5-102) and (5-103) that it is sufficient to analyze the convergence rate of

$$\int_{N}^{+\infty} \frac{\sin(cx)}{x} dx \tag{5-104}$$

with $c > 0$. The basic idea is to write the integral (5-104) as an infinite sum of integrals over one period $2\pi/c$ of the $\sin(cx)$ function

$$\int_{N}^{+\infty} \frac{\sin(cx)}{x} dx = \int_{N}^{2k_1\pi/c} \frac{\sin(cx)}{x} dx + \sum_{k=k_1}^{+\infty} \int_{2k\pi/c}^{2(k+1)\pi/c} \frac{\sin(cx)}{x} dx \tag{5-105}$$

with $k_1 = \text{int}(Nc/(2\pi)) + 1$ and int( ) the integer part of a number. Each integral in the infinite sum can be bounded above by

$$\begin{aligned} \int_{2k\pi/c}^{2(k+1)\pi/c} \frac{\sin(cx)}{x} dx &= \int_{2k\pi/c}^{(2k+1)\pi/c} \frac{\sin(cx)}{x} dx + \int_{(2k+1)\pi/c}^{2(k+1)\pi/c} \frac{\sin(cx)}{x} dx \\ &\le \int_{2k\pi/c}^{(2k+1)\pi/c} \frac{\sin(cx)}{2k\pi/c} dx + \int_{(2k+1)\pi/c}^{2(k+1)\pi/c} \frac{\sin(cx)}{2(k+1)\pi/c} dx \end{aligned} \tag{5-106}$$

because $\sin(cx) \ge 0$ for any x in $[2k\pi/c, (2k+1)\pi/c]$ and $\sin(cx) \le 0$ for any x in $[(2k+1)\pi/c, 2(k+1)\pi/c]$. Working out the integrals in (5-106) gives

$$\int_{2k\pi/c}^{2(k+1)\pi/c} \frac{\sin(cx)}{x}dx \le \frac{1}{\pi k} - \frac{1}{\pi(k+1)} = \frac{1}{\pi k(k+1)} \tag{5-107}$$

Hence, the second term in the right-hand side of (5-105) can be bounded above by

$$\sum_{k=k_1}^{+\infty} \int_{2k\pi/c}^{2(k+1)\pi/c} \frac{\sin(cx)}{x}dx \le \sum_{k=k_1}^{+\infty} \frac{1}{\pi k(k+1)} \le C_1 \int_{k_1}^{+\infty} \frac{dx}{x^2} = \frac{C_1}{k_1} = O(N^{-1}) \tag{5-108}$$

with C_1 a constant independent of N. The second inequality is due to the Cauchy integral test (Gradshteyn and Ryzhik, 1980). The first term in the right-hand side of (5-105) can be bounded above by

$$\left| \int_{N}^{2k_1\pi/c} \frac{\sin(cx)}{x}dx \right| \le \frac{2k_1\pi/c - N}{N} \le \frac{2\pi}{cN} = O(N^{-1}) \tag{5-109}$$

Collecting (5-108) and (5-109) proves that the integral (5-105) is an $O(N^{-1})$.

Appendix 5.F: Convergence Rate of the Residual Alias Errors

Because the output Fourier spectrum $Y(j\omega)$ is related to the input Fourier spectrum $U(j\omega)$ by $Y(j\omega) = G(j\omega)U(j\omega)$ with $G(s)$ stable, the output signal has exactly the same spectral properties as the input signal, for example, band-limited, discrete Fourier spectrum. Similarly, because $S_{yy}(j\omega) = |G(j\omega)|^2 S_{uu}(j\omega)$ with $G(s)$ stable, the output power spectrum $S_{yy}(j\omega)$ has the same spectral properties as the input power spectrum $S_{uu}(j\omega)$, for example, band-limited, differentiable power spectrum. Therefore, to study $\delta(s_k) = \Delta(s_k)/A(s_k)$ (see (5-92)) it is sufficient to study the spectral content of a band-limited signal $x(t)$, observed during a time NT_s. The errors in the DFT spectra giving the term $\delta(s_k)$ are in fact, leakage errors that can be interpreted as alias errors. Indeed, due to the multiplication of $x(t)$ with a rectangular window $w_N(t)$, sharp transitions occur at the edges of $x_N(t) = x(t)w_N(t)$. These sharp transitions have a high frequency content. For ease of notation, we will take the time origin in the middle of the observation window $w_N(t)$: $w_N(t) = 1$ for $t \in [-N/2, N/2)T_s$ and zero elsewhere. First, we prove the result for normalized periodic signals (see Definition 3.4), and next, for random signals.

5.F.1 Periodic Signals. A normalized periodic signal has the form

$$x(t) = \sum\nolimits_{k=1}^{F} \frac{A_k}{\sqrt{N}} \sin(\omega_k t + \phi_k) \tag{5-110}$$

where $A_k > 0$ and where F increases with N, $F = O(N)$ (see Definition 3.4). By assumption, the signal $x(t)$ is band limited so that $\max_k f_k \le f_B < f_s/2$. The outline of the proof is as follows. First, we calculate the high frequency content $x_a(t)$ of the observed signal $x_N(t) = w_N(t)x(t)$. Next, the energy of $x_a(t)$ is compared with that of $x_N(t)$. Finally, via

Parceval's theorem, we draw conclusions concerning the Fourier spectra $X_N(j\omega)$ and $X_a(j\omega)$ of $x_N(t)$ and $x_a(t)$ respectively.

The high frequency content $x_a(t)$ is found by multiplying $X_N(j\omega)$ with a window $P(f)$ that excludes all frequencies in the band $[-f_s/2, f_s/2]$, and by taking afterward the inverse Fourier transform. We find

$$X_a(j2\pi f) = X_N(j2\pi f)P(f) \Rightarrow x_a(t) = x_N(t)*p(t) \tag{5-111}$$

with $p(t)$ the inverse Fourier transform of $P(f)$ and * the convolution product. The window $P(f)$ can be written as $P(f) = 1 - B(f)$, where $B(f) = 1$ for $|f| \le f_s/2$ and zero elsewhere, so that

$$p(t) = \delta(t) - \sin(\omega_s t/2)/(\pi t) \tag{5-112}$$

with $\sin(\omega_s t/2)/(\pi t)$ the inverse Fourier transform of $B(f)$. Using (5-112), we get the following expression for $x_a(t)$:

$$\begin{aligned} x_a(t) &= x_N(t) - x_N(t)*(\sin(\omega_s t/2)/(\pi t)) \\ &= x_N(t) - \int_{t-NT_s/2}^{t+NT_s/2} x(t-\tau)\frac{\sin(\omega_s \tau/2)}{\pi\tau}d\tau \end{aligned} \tag{5-113}$$

Putting (5-110) in (5-113) gives, using $\sin(a-b) = \sin(a)\cos(b) - \cos(a)\sin(b)$,

$$x_a(t) = x_N(t) - x_1(t) + x_2(t) \tag{5-114}$$

with

$$\begin{aligned} x_1(t) &= \sum_{k=1}^{F}\frac{A_k}{\sqrt{N}}\sin(\omega_k t + \phi_k)\int_{t-NT_s/2}^{t+NT_s/2} f_1(\tau)d\tau, \qquad f_1(\tau) = \frac{\cos(\omega_k\tau)\sin(\omega_s\tau/2)}{\pi\tau} \\ x_2(t) &= \sum_{k=1}^{F}\frac{A_k}{\sqrt{N}}\cos(\omega_k t + \phi_k)\int_{t-NT_s/2}^{t+NT_s/2} f_2(\tau)d\tau, \qquad f_2(\tau) = \frac{\sin(\omega_k\tau)\sin(\omega_s\tau/2)}{\pi\tau} \end{aligned} \tag{5-115}$$

We now study (5-114) for the four following cases: (i) $t \in (-NT_s/2, NT_s/2)$, (ii) $t = -NT_s/2$ and $t = NT_s/2$, (iii) $t > NT_s/2$, and (iv) $t < -NT_s/2$.

If $t \in (-NT_s/2, NT_s/2)$, then we can split up the integrals in (5-115) as

$$\int_{t-NT_s/2}^{t+NT_s/2} f_i(\tau)d\tau = \int_{-\infty}^{+\infty} f_i(\tau)d\tau - \int_{-\infty}^{t-NT_s/2} f_i(\tau)d\tau - \int_{t+NT_s/2}^{+\infty} f_i(\tau)d\tau \tag{5-116}$$

for $i = 1, 2$. From Appendix 5.E it follows that

$$\int_{-\infty}^{+\infty} f_1(\tau)d\tau = 1 \qquad \int_{-\infty}^{+\infty} f_2(\tau)d\tau = 0 \tag{5-117}$$

and that for $N \to \infty$ and t fixed, or $t = \alpha NT_s/2$ with $\alpha \in (-1, 1)$,

$$\int_{-\infty}^{t-NT_s/2} f_i(\tau)d\tau = O(\frac{1}{t-NT_s/2}) \qquad \int_{t+NT_s/2}^{+\infty} f_i(\tau)d\tau = O(\frac{1}{t+NT_s/2}) \tag{5-118}$$

The first integral in (5-117) is valid only if $x(t)$ (5-110) is band limited, $\omega_k < \omega_s/2$ for $k = 1, 2, \ldots, F$, while the second integral follows from the fact that $f_2(\tau)$ is an odd function of τ. Collecting (5-114) to (5-118) gives, using $x(t) = O(N^0)$,

$$x_a(t) = O(\frac{1}{t+NT_s/2}) + O(\frac{1}{t-NT_s/2}) \text{ for } t = \alpha NT_s/2 \text{ with } \alpha \in (-1, 1) \tag{5-119}$$

For $t = -NT_s/2$ and $NT_s/2$ the integrals in (5-115) are finite for any N, ∞ included. The same is true for $t = -NT_s/2 \pm \Delta t$ and $NT_s/2 \pm \Delta t$, with Δt independent of N. Together with $x(t) = O(N^0)$, it follows that

$$x_a(t) = O(N^0) \text{ for } t = -NT_s/2 \pm \Delta t \text{ and } NT_s/2 \pm \Delta t \tag{5-120}$$

with Δt independent of N.

If $t > \alpha NT_s/2$ with $\alpha > 1$, then we can split up the integrals in (5-115) as

$$\int_{t-NT_s/2}^{t+NT_s/2} f_i(\tau)d\tau = \int_{t-NT_s/2}^{+\infty} f_i(\tau)d\tau - \int_{t+NT_s/2}^{+\infty} f_i(\tau)d\tau \tag{5-121}$$

where, according to Appendix 5.E,

$$\int_{t-NT_s/2}^{+\infty} f_i(\tau)d\tau = O(\frac{1}{t-NT_s/2}) \qquad \int_{t+NT_s/2}^{+\infty} f_i(\tau)d\tau = O(\frac{1}{t+NT_s/2}) \tag{5-122}$$

Collecting (5-114), (5-115), (5-121), and (5-122) gives, using $x(t) = O(N^0)$ and $x_N(t) = 0$ for $t > NT_s/2$,

$$x_a(t) = O(\frac{1}{t-NT_s/2}) + O(\frac{1}{t+NT_s/2}) \text{ for } t > \alpha NT_s/2 \text{ with } \alpha > 1 \tag{5-123}$$

Following the same lines, we find for $t < -NT_s/2$,

$$x_a(t) = O(\frac{1}{t-NT_s/2}) + O(\frac{1}{t+NT_s/2}) \text{ for } t < -\alpha NT_s/2 \text{ with } \alpha > 1 \tag{5-124}$$

From (5-119), (5-120), (5-123), and (5-124) we conclude that $x_a(t)$ tends to zero everywhere as $O(N^{-1})$, except in a close, N-independent, neighborhood of $t = -NT_s/2$ and $t = NT_s/2$ where it behaves as an $O(N^0)$. A graphical representation of $x_a(t)$ is shown in Figure 5-9. The ringing at the edges of the observation window are known as the Gibbs phenomenon. Note that the difference $x_N(t) - x_a(t)$ is band limited.

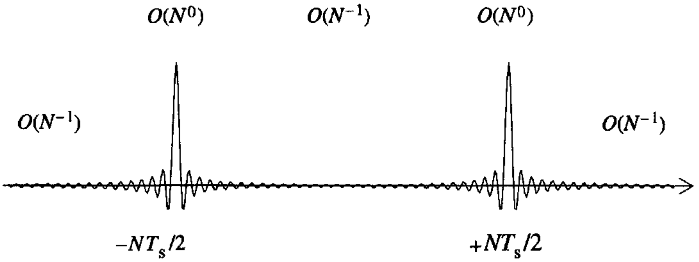


Figure 5-9. Visualization of the high frequency content $x_a(t)$ of the observed signal $x_N(t)$.

Using (5-119), (5-120), (5-123), and (5-124), we can calculate the energy of $x_a(t)$. We find

$$\int_{-\infty}^{+\infty} x_a^2(t)dt = O(N^0) \tag{5-125}$$

From (5-110) it follows that

$$\int_{-\infty}^{+\infty} x_N^2(t)dt = \int_{-NT_s/2}^{+NT_s/2} x^2(t)dt = O(N) \tag{5-126}$$

Applying Parceval's theorem to (5-125) and (5-126), we get

$$\int_{-\infty}^{+\infty} x_a^2(t)dt = \int_{-\infty}^{+\infty} |X_a(j2\pi f)|^2 df = 2\int_{f_s/2}^{+\infty} |X_a(j2\pi f)|^2 df = O(N^0) \tag{5-127}$$

$$\begin{aligned} \int_{-f_s/2}^{+f_s/2} |X_N(j2\pi f)|^2 df &= \int_{-\infty}^{+\infty} |X_N(j2\pi f)|^2 df - \int_{-\infty}^{+\infty} |X_a(j2\pi f)|^2 df \\ &= \int_{-\infty}^{+\infty} x_N^2(t)dt - \int_{-\infty}^{+\infty} x_a^2(t)dt \\ &= O(N) \end{aligned} \tag{5-128}$$

It follows that the ratio of the energy above Nyquist ($|f| > f_s/2$) to the energy below Nyquist ($|f| < f_s/2$) of $x_N(t)$ is an $O(N^{-1})$. By construction, the energy of the normalized periodic signal $x(t)$ is continuously spread over the $F = O(N)$ frequencies f_k (see Definition 3.4), so that the DFT spectrum $X(k)$ of $x_N(t)$ is an $O(N^0)$. As the energy of the pulse-like signal $x_a(t)$ is also continuously spread over the frequency, it follows directly that $\delta(s_k) = O(N^{-1/2})$.

Note that formulas (5-125) to (5-128) are also valid for periodic signals with a fixed number of frequencies $F = O(N^0)$ and fixed amplitudes $A_k/\sqrt{N} = O(N^0)$ in (5-110). The difference with the normalized periodic signals is that the signal energy is concentrated at a fixed number of frequencies f_k. Hence, at the excited frequencies f_k, $X(k)$ is an $O(N^{1/2})$, while $\delta(s_k)$ is still an $O(N^{-1/2})$. It shows that the relative convergence rate of $\delta(s_k)$ is an $O(N^{-1})$ at f_k.

5.F.2 Random Signals. The autocorrelation function $R_N(\tau, t)$ of the observed random signal $x_N(t)$ is related to the autocorrelation $R(\tau)$ of the complete signal $x(t)$ by

$$R_N(\tau, t) = \mathcal{E}\{x_N(t)x_N(t+\tau)\} = w_N(t)w_N(t+\tau)R(\tau) \tag{5-129}$$

Taking the Fourier transform of (5-129) w.r.t. τ gives the power spectrum

$$\begin{aligned} S_N(j\omega, t) &= w_N(t)[S(j\omega)*(W_N(j\omega)e^{j\omega t})] \\ &= w_N(t)\int_{-\infty}^{+\infty} S(j2\pi g)e^{j2\pi(f-g)t}W_N(j2\pi(f-g))dg \\ &= w_N(t)\int_{-f_B}^{+f_B} S(j2\pi g)e^{j2\pi(f-g)t}W_N(j2\pi(f-g))dg \end{aligned} \tag{5-130}$$

with $W_N(j\omega) = 2\omega^{-1}\sin(\omega NT_s/2)$ the spectrum of the window $w_N(t)$, and where the last equality is due to the fact that $x(t)$ is band limited, $S(j\omega) = 0$ for $f > f_B$. Because $S(j2\pi g)/(f-g)$ is finite for any $f > f_s/2$ and $|g| \leq f_B$, we can apply partial integration to (5-130). We find for $f > f_s/2$ □

$$\begin{aligned} S_N(j\omega, t) &= \frac{w_N(t)}{N}\left[\frac{S(j2\pi g)}{f-g}\frac{\cos(\pi NT_s(f-g))}{\pi^2 T_s}\right]_{-f_B}^{+f_B} \\ &\quad -\frac{w_N(t)}{N}\int_{-f_B}^{+f_B}\frac{\cos(\pi(f-g)NT_s)}{\pi^2 T_s}\frac{d}{dg}\left(\frac{S(j2\pi g)}{f-g}\right)dg \end{aligned} \tag{5-131}$$

Clearly, the first term in the right-hand side of (5-131) is an $O(N^{-1})$. Because $S(j2\pi g)$ is differentiable for $|g| \leq f_B$ and $f - g \neq 0$ for any $|f| > f_s/2$ and $|g| \leq f_B$, the integral in (5-131) is finite for any N, ∞ included. Hence, the second term in the right-hand side of (5-131) is also an $O(N^{-1})$, so that $S_N(j\omega, t) = O(N^{-1})$ for $f > f_s/2$. This establishes the mean square convergence (see Chapter 14) of the signal energy above $f_s/2$ to zero (the power spectrum is a second-order moment). As $S_N(j\omega, t) = O(N^0)$ for $|f| \leq f_B$ and the DFT spectrum $X(k)$ is an $O(N^0)$ for stationary random signals, we have $\delta(s_k) = O_{\text{m.s.}}(N^{-1/2})$.

Appendix 5.G: Relation between DFT Spectra and Transfer Function for Arbitrary Signals with Missing Data

The DFT spectrum $X(k) = \text{DFT}(x(t))$ of the complete set (no missing data) can be split into the contributions of the known and the unknown samples

$$X(k) = X^{\text{m}}(k) + z_k^{-K_x} I_x(z_k^{-1}) \tag{5-132}$$

where $X^{\text{m}}(k) = \text{DFT}(x^{\text{m}}(t))$, $I_x(z_k^{-1}) = N^{-1/2}\sum_{t=0}^{M_x-1} x(K_x+t)z^{-t}$, and $x^{\text{m}}(t)$ is defined in (5-45). Applying (5-132), with $X = U, Y$ and $x = u, y$, to (5-33) and (5-34) gives (5-46) and (5-47), respectively. □

Appendix 5.H: Free Decay Response of a Finite-Dimensional System

For discrete-time systems, (5-53) follows directly from (5-34) with $U(k) = 0$. For strictly proper lumped continuous-time systems we use (5-89) with $U_N(s) = 0$

$$Y_N(s) = \frac{I_1(s)}{A(s)} - e^{-NT_s s}\frac{I_2(s)}{A(s)} \tag{5-133}$$

and where the polynomials $I_1(s)$, $I_2(s)$ have order $n_a - 1$. Taking the one-sided Z-transform of (5-133) gives

$$Y_N(z) = T_1(z^{-1}) - z^{-N}T_2(z^{-1}) \tag{5-134}$$

with $Y_N(z) = \sum_{t=0}^{N-1} y(tT_s)z^{-t}$ and $T_1(z^{-1})$, $T_2(z^{-1})$ rational forms in z^{-1} of order $(n_a - 1)$ over n_a. The poles z_p of $T_1(z^{-1})$, $T_2(z^{-1})$ are related to the roots s_p of $A(s)$ by the impulse invariant transformation, $z_p = e^{s_p T_s}$. Evaluating (5-134) at the DFT frequencies $z_k = e^{j2\pi k/N}$ with $z_k^{-N} = 1$ gives (5-53), after division by $\sqrt{N}$. □

Appendix 5.I: Relation between the Free Decay Parameters and the Partial Fraction Expansion

We will prove (5-54) for the complex case (5-52). Using $\sum_{t=0}^{N-1} x^t = (1 - x^N)/(1 - x)$ and $z_k^N = 1$, the DFT transform $Y(k) = N^{-1/2}\sum_{t=0}^{N-1} y(tT_s)z_k^{-t}$ of $y(t)$, (5-52) becomes

$$Y(k) = \frac{1}{\sqrt{N}}\sum_{r=1}^{n_a} a_r e^{j\phi_r}\lambda_r^{\tau/T_s}\frac{1-\lambda_r^N}{1-\lambda_r z_k^{-1}} \tag{5-135}$$

where $\lambda_r = e^{(-d_r + j\omega_r)T_s}$. Comparing (5-135) to (5-39) with $q = 0$, $\lambda_{-r} \neq \bar{\lambda}_r$ and $l_{-r} \neq \bar{l}_r$ gives (5-54). The proof of the real case follows the same lines (see Exercise 5.8). □

Appendix 5.J: Some Properties of Polynomials

Lemma 5.10: Consider the polynomial $P(\Omega)$,

$$P(\Omega) = \left(\sum_{r=0}^{n_b}\beta_r\Omega^r\right)\left(\sum_{r=0}^{n_a}a_r\Omega^r\right) + \left(\sum_{r=0}^{n_a-1}\alpha_r\Omega^r\right)\left(\sum_{r=0}^{n_b}b_r\Omega^r\right) \tag{5-136}$$

with a_r, b_r fixed coefficients and α_r, β_r free parameters, and suppose that $P(\Omega) = 0$ must be fulfilled for any Ω. All the parameters α_r and β_r are zero if and only if the polynomials $A(\Omega, \theta)$ and $B(\Omega, \theta)$ have no common roots.

Proof. If the parameters α_r and β_r are not all zero, then we can rewrite $P(\Omega) = 0$ as

$$\frac{B(\Omega, \theta)}{A(\Omega, \theta)} = -\frac{\sum_{r=0}^{n_b}\beta_r\Omega^r}{\sum_{r=0}^{n_a-1}\alpha_r\Omega^r} \tag{5-137}$$

(if all α_r are zero in $P(\Omega) = 0$ then all β_r are zero and vice versa so that at least one α_r and one β_r should be different from zero).

If the polynomials $A(\Omega, \theta)$ and $B(\Omega, \theta)$ have *no common roots*, then $B(\Omega, \theta)/A(\Omega, \theta)$ has minimal order n_b over n_a. Equation (5-137) implies that $B(\Omega, \theta)/A(\Omega, \theta)$ can be written as a rational form of order n_b over $n_a - 1$, which is impossible. Hence, $P(\Omega) = 0$ can be true for any Ω only if the parameters α_r, β_r are all zero.

If the polynomials $A(\Omega, \theta)$ and $B(\Omega, \theta)$ have *common roots*, then the minimal order of $B(\Omega, \theta)/A(\Omega, \theta)$ is less than n_b over n_a, and (5-137) is fulfilled with α_r and β_r not all zero. □

Lemma 5.11: Consider the polynomial $P(\Omega)$ in (5-136) and suppose that $P(\Omega) = 0$ for at least $(n_a + n_b + 1)/2$ DFT frequencies Ω_k where DC $(k = 0)$ and Nyquist $(k = N/2)$ each count for $1/2$. The free parameters α_r, β_r are zero if and only if the polynomials $A(\Omega, \theta)$ and $B(\Omega, \theta)$ have no common roots.

Proof. The polynomial equation $P(\Omega) = 0$ is fulfilled for any Ω if and only if the coefficients of all the powers of Ω are zero. We will show that this is also true if $P(\Omega) = 0$ for at least $(n_a + n_b + 1)/2$ DFT frequencies. Applying Lemma 5.10 proves the lemma.

Evaluating $P(\Omega) = \sum_{r=0}^{n_p} p_r\Omega^r$, with $n_p = n_a + n_b$, at F DFT frequencies $k_j \in \{1, 2, ..., N/2 - 1\}$, $j = 1, 2, ..., F$, gives

$$V(\Omega_{k_1}, \Omega_{k_2}, ..., \Omega_{k_F})p = 0 \tag{5-138}$$

with $p^T = [p_0 p_1 ... p_{n_p}]$ and where the matrix $V(\Omega_{k_1}, \Omega_{k_2}, ..., \Omega_{k_F}) \in \mathbb{C}^{F\times(n_p+1)}$ has a Vandermonde structure

$$V(\Omega_{k_1}, \Omega_{k_2}, ..., \Omega_{k_F}) = \begin{bmatrix} 1 & \Omega_{k_1} & \Omega_{k_1}^2 & ... & \Omega_{k_1}^{n_p} \\ 1 & \Omega_{k_2} & \Omega_{k_2}^2 & ... & \Omega_{k_2}^{n_p} \\ ... & ... & ... & ... & ... \\ 1 & \Omega_{k_F} & \Omega_{k_F}^2 & ... & \Omega_{k_F}^{n_p} \end{bmatrix} \tag{5-139}$$

The Vandermonde matrix (5-139) is of full rank if and only if the F DFT frequencies Ω_{k_j} are all different (see Golub and Van Loan, 1996 and Exercise 13.6). Adding the F complex conjugate DFT frequencies to (5-138) gives

$$V(\Omega_{k_1}, \ldots, \Omega_{k_F}, \Omega_{-k_1}, \ldots, \Omega_{-k_F})p = 0 \tag{5-140}$$

where $V(\Omega_{k_1}, \ldots, \Omega_{k_F}, \Omega_{-k_1}, \ldots, \Omega_{-k_F}) \in \mathbb{C}^{2F \times (n_p + 1)}$ is of full rank $(\Omega_{k_j} \neq \Omega_{-k_j})$. Hence from (5-140), it follows that $p = 0$ if and only if $F \geq (n_p + 1)/2$. The same reasoning holds if DC $(k = 0)$ and Nyquist $(k = N/2)$ are added to the frequencies. However, since Ω_0, $\Omega_{N/2}$ are real numbers, they increase the rank of $V(\Omega_{k_1}, \ldots, \Omega_{k_F}, \Omega_{-k_1}, \ldots, \Omega_{-k_F})$ by one instead of two as for each complex Ω_k. □

Appendix 5.K: Proof of the Identifiability of Transfer Function Model (5-32) (Theorem 5.9)

Choosing $a_{n_a} = 1$, transfer function model (5-32) can be written as

$$\Omega_k^{n_a} Y(k) = \sum_{r=0}^{n_b} b_r f_r(\Omega_k) - \sum_{r=0}^{n_a - 1} a_r g_r(\Omega_k) \tag{5-141}$$

with $f_r(\Omega_k) = \Omega_k^r U(k)$ and $g_r(\Omega_k) = \Omega_k^r Y(k)$. The coefficients $a_0, a_1, \ldots, a_{n_a - 1}, b_0, \ldots, b_{n_b}$ in (5-141) are identifiable if and only if the functions $f_r(\Omega_k)$, $r = 0, 1, \ldots, n_b$ and $g_r(\Omega_k)$, $r = 0, 1, \ldots, n_a - 1$ are linearly independent. This is the case if and only if

$$\sum_{r=0}^{n_b} \beta_r f_r(\Omega_k) + \sum_{r=0}^{n_a - 1} \alpha_r g_r(\Omega_k) = 0 \tag{5-142}$$

$k = 0, 1, \ldots, N/2$, implies that all parameters α_r, β_r are zero. Multiplying (5-142) by $A(\Omega_k, \theta)$ and using $Y(k) = G(\Omega_k, \theta)U(k)$ gives $P_1(\Omega_k)U(k) = 0$, $k = 0, 1, \ldots, N/2$, with

$$P_1(\Omega) = \left(\sum_{r=0}^{n_b} \beta_r \Omega^r\right)\left(\sum_{r=0}^{n_a} a_r \Omega^r\right) + \left(\sum_{r=0}^{n_a - 1} \alpha_r \Omega^r\right)\left(\sum_{r=0}^{n_b} b_r \Omega^r\right) \tag{5-143}$$

At the DFT frequencies where $U(k) \neq 0$, $P_1(\Omega_k)U(k) = 0$ is equivalent to $P_1(\Omega_k) = 0$. The free parameters α_r and β_r in $P_1(\Omega) = 0$ are zero if and only if $A(\Omega, \theta)$ and $B(\Omega, \theta)$ have no common roots and $U(k) \neq 0$ for at least $(n_a + n_b + 1)/2$ different DFT frequencies (proof: see Appendix 5.J). □

Appendix 5.L: Proof of the Identifiability of Transfer Function Models (5-35) and (5-36)

Choosing $a_{n_a} = 1$, transfer function models (5-33) and (5-34) can be written as

$$\Omega_k^{n_a} Y(k) = \sum_{r=0}^{n_b} b_r f_r(\Omega_k) - \sum_{r=0}^{n_a - 1} a_r g_r(\Omega_k) + \sum_{r=0}^{n_i} i_r \Omega_k^r + \Delta(\Omega_k) \tag{5-144}$$

with $f_r(\Omega_k) = \Omega_k^r U(k)$, $g_r(\Omega_k) = \Omega_k^r Y(k)$ and with $\Delta(z_k^{-1}) = 0$ and $\Delta(s_k)$ given by (5-92). A necessary condition for the identifiability of the coefficients $a_0, \ldots, a_{n_a - 1}, b_0, \ldots,$

b_{n_b}, i_0, ..., i_{n_i} in (5-144) is that $f_r(\Omega_k)$ and Ω_k^r are linearly independent. If $f_r(\Omega_k)$ and Ω_k^r are linearly dependent, then there exist coefficients β_r, γ_r, not all zero, such that

$$\sum_{r=0}^{n_b} \beta_r f_r(\Omega_k) + \sum_{r=0}^{n_i} \gamma_r \Omega_k^r = 0 \tag{5-145}$$

As the functions $f_r(\Omega_k)$ and Ω_k^r are themselves linearly independent (see Exercise 5.10) not all β_r and not all γ_r are zero. From (5-145) it follows, then, that

$$U(k) = -\left(\sum_{r=0}^{n_i} \gamma_r \Omega_k^r\right) / \left(\sum_{r=0}^{n_b} \beta_r \Omega_k^r\right) \tag{5-146}$$

$k = 0, 1, ..., N/2$. We conclude that the functions $f_r(\Omega_k)$ and Ω_k^r are linearly dependent if and only if $U(k)$ can be written as a rational form of order n_i over n_b or less, otherwise they are linearly independent. If $U(k) \neq 0$ for less than $(n_b + n_i + 2)/2$ different DFT frequencies, then $U(k)$ can always be written as in the form (5-146) (a rational function of order n_i over n_b fits exactly $(n_b + n_i + 1)/2$ arbitrary complex numbers).

Transfer function models (5-35) and (5-36) can be written as

$$Y(k) = G(\Omega_k, \theta)U(k) + T(\Omega_k, \theta) + \delta(\Omega_k) \tag{5-147}$$

with $\delta(z_k^{-1}) = 0$ and

$$\delta(s_k) = \frac{1}{T_s\sqrt{N}} \sum_{\substack{n=-\infty \\ n \neq 0}}^{n=+\infty} [G(s_k, \theta)U_N(s_k - nj\omega_s) - Y_N(s_k - nj\omega_s)] \tag{5-148}$$

(see (5-92)). If the polynomials $A(\Omega, \theta)$, $B(\Omega, \theta)$, and $I(\Omega, \theta)$ have common roots, then the rational functions $G(\Omega, \theta)$ and $T(\Omega, \theta)$ can be simplified, leaving (5-147) unchanged. Clearly, the roots that have been removed are not identifiable. □

6

An Intuitive Introduction to Frequency Domain Identification

Abstract: In the next two chapters a detailed study of frequency domain identification schemes will be made. A wide class of methods is discussed and it will be shown how the properties of the estimators are set by the choice of their cost function. Those readers who just want to solve their modeling problem, without passing through all these underlying theories, might still profit from a basic understanding of the methods they will use. For that reason we decided to provide, in this chapter, an intuitive insight into the frequency domain identification problem. First, a straightforward approach will be discussed, starting from the measured FRF of the systems transfer function; next a more general formulation will be made, based on the errors-in-variables concept, leading to a very robust identification method. Finally, it is discussed, briefly, how the general method can be applied to specific situations: no input noise present; the FRF is measured, and so on.

6.1 INTUITIVE APPROACH

The basic aim of this book is to measure and model the transfer function $G_0(\Omega)$ of a plant, starting from noisy input and output measurements (see Figure 6-1). An intuitive approach is to extract, first, a measured FRF $G(\Omega_k)$, $k = 1, ..., F$ of the systems' transfer function at a set of well-chosen frequencies (see Chapter 2 for a detailed discussion). Next, these measurements are approximated by a parametric model $G(\Omega_k, \theta)$ that explains the measurements as much as possible. As explained in Chapter 1, the quality of the match between measurements and model is measured by the cost function. The parameters are then tuned to minimize the cost function so that a best match is obtained. There is no unique choice for the cost function, and because each cost leads to an estimator, it is possible to find different estimators for the same problem. An intuitive choice of the cost function is

$$V_F(\theta, Z) = \frac{1}{F}\sum_{k=1}^{F} \frac{|G(\Omega_k) - G(\Omega_k, \theta)|^2}{\sigma_G^2(k)} \tag{6-1}$$

The weighted least squares distance between the measurement and the model is minimized. Measurements with a small uncertainty ($\sigma_G^2(k)$ is small) are more important than those with

System Identification: A Frequency Domain Approach. Rik Pintelon and Johan Schoukens
Copyright © 2001 Institute of Electrical and Electronics Engineers. ISBN: 0-780-36000-1

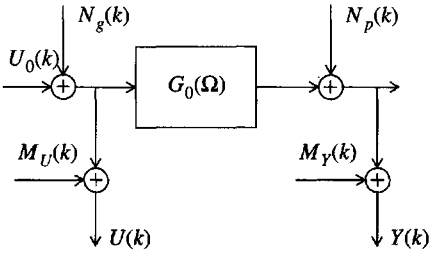


Figure 6-1. Frequency domain representation of the measurement process. Note that the system can be captured in a feedback loop.

a large uncertainty ($\sigma_G^2(k)$ is large). Although this method works amazingly well in many cases, it suffers from a major drawback. It is not always that easy to get a good measurement of G_0 due to the presence of the noise $M_U(k)$ on the input. If the classical correlation methods (H_1 method) are used, a bias appears (see Chapter 2). The measured transfer function converges for an increasing number of averages to:

$$\lim_{M \to \infty} G(\Omega_k) = \frac{S_{YU}(\Omega_k)}{S_{UU}(\Omega_k)} = G_0(\Omega_k)\frac{1}{1 + S_{M_U M_U}(\Omega_k)/S_{UU}(\Omega_k)} \tag{6-2}$$

It is then easy to show that the parametric approximation will also be biased. When periodic excitations are used, alternative methods based on the direct division of output and input spectrum are available: $G(\Omega_k) = Y(k)/U(k)$. Although this method is less sensitive to the bias problem (see Chapter 2) for sufficiently large SNR at the input (better than 6 dB), it can be shown that in general its variance $\sigma_G^2(k)$ does not exist. (Guillaume et al., 1996a; Broersen, 1995). Especially for a low SNR at the input, large spikes frequently appear in the estimate, thus the variance estimate does not converge anymore. For larger SNRs the risk of encountering this problem becomes negligible in practice. However, this puts the user in a situation where he has to decide himself whether the method is applicable or not. For that reason a more robust alternative is formulated in the next section. Although it looks, at a glance, more complicated, it turns out that the computational complexity is not higher than that of the intuitive approach if periodic excitations are used. The major advantage for the less experienced user is that a check is no longer needed to verify whether the operational conditions on the intuitive technique are met or not. The algorithm can be automated fully and included in a general purpose package for public usage by laymen in the identification domain.

6.2 THE ERRORS-IN-VARIABLES FORMULATION

The intuitive methods of the previous section run into problems due to the presence of a division $Y(k)/U(k)$ that is a highly nonlinear operator. The denominator can become almost zero (the noise cancels the input) at some frequencies and this creates outliers. The errors-in-variables approach avoids a direct division of both measured spectra. Instead, the input and output spectra are considered as unknown parameters, connected by the parametric transfer function model:

$$\begin{aligned} Y(k) &= Y_0(k) + N_Y(k) \\ U(k) &= U_0(k) + N_U(k) \end{aligned} \tag{6-3}$$

with $Y_0(k) = G(\Omega_k, \theta)U_0(k)$ and where $N_Y(k)$ and $N_U(k)$ include the generator noise, the process noise, and the measurement noise. Because the exact Fourier coefficients $Y_0(k)$ and $U_0(k)$ are unknown, they are replaced by the parameters $Y_\mathrm{p}(k)$ and $U_\mathrm{p}(k)$, which are estimated by minimizing the distance between the measurements and the parameters $(|U(k) - U_\mathrm{p}(k)|$, $|Y(k) - Y_\mathrm{p}(k)|$), leading to a new constraint optimization problem. If the input and output measurements are uncorrelated with each other, the following least squares cost function can be used:

$$\begin{aligned} V_F(\theta, Z) &= \frac{1}{F}\sum_{k=1}^{F} \frac{|U(k) - U_\mathrm{p}(k)|^2}{\sigma_U^2(k)} + \frac{|Y(k) - Y_\mathrm{p}(k)|^2}{\sigma_Y^2(k)} \\ &= \frac{1}{F}\sum_{k=1}^{F} \begin{pmatrix} Y(k) - Y_\mathrm{p}(k) \\ U(k) - U_\mathrm{p}(k) \end{pmatrix}^H \begin{bmatrix} \sigma_Y^2(k) & 0 \\ 0 & \sigma_U^2(k) \end{bmatrix}^{-1} \begin{pmatrix} Y(k) - Y_\mathrm{p}(k) \\ U(k) - U_\mathrm{p}(k) \end{pmatrix} \end{aligned} \tag{6-4}$$

to be minimized under the constraints

$$Y_\mathrm{p}(k) = G(\Omega_k, \theta)U_\mathrm{p}(k) \qquad k = 1, 2, \ldots, F \tag{6-5}$$

In reality, the noise $N_U(k)$ and $N_Y(k)$ is correlated ($\sigma_{YU}^2(k) \neq 0$), and the full weighted least squares cost function should be considered:

$$V_F(\theta, Z) = \frac{1}{F}\sum_{k=1}^{F} \begin{pmatrix} Y(k) - Y_\mathrm{p}(k) \\ U(k) - U_\mathrm{p}(k) \end{pmatrix}^H \begin{bmatrix} \sigma_Y^2(k) & \sigma_{YU}^2(k) \\ \sigma_{UY}^2(k) & \sigma_U^2(k) \end{bmatrix}^{-1} \begin{pmatrix} Y(k) - Y_\mathrm{p}(k) \\ U(k) - U_\mathrm{p}(k) \end{pmatrix} \tag{6-6}$$

where $\sigma_{UY}(k) = \overline{\sigma_{YU}}(k)$. This cost function should be minimized with respect to the model parameters θ and also to the Fourier coefficients $U_\mathrm{p}(k)$, $Y_\mathrm{p}(k)$, $k = 1, \ldots, F$. As F can be very large, this appears to be a very hard task. However, this cost function can be simplified further. It is possible to eliminate $U_\mathrm{p}(k)$, $Y_\mathrm{p}(k)$ explicitly from the problem, simplifying the cost function to:

$$V_F(\theta, Z) = \frac{1}{F}\sum_{k=1}^{F} \frac{|Y(k) - G(\Omega_k, \theta)U(k)|^2}{\sigma_Y^2(k) + \sigma_U^2(k)|G(\Omega_k, \theta)|^2 - 2\mathrm{Re}(\sigma_{YU}^2(k)\bar{G}(\Omega_k, \theta))} \tag{6-7}$$

Some of the advantages and properties of this formulation are discussed below.

6.2.1.1 Robustness with Respect to Bad Measurements. Compared with (6-1), division of the measured Fourier coefficients is no longer needed. The cost function does not degenerate, even if the measured input equaled zero at some frequencies. The user should not bother anymore with the selection of an appropriate method to measure the FRF.

6.2.1.2 Symmetric Formulation. By replacing in the cost function $G(\Omega_k, \theta) = B(\Omega_k, \theta)/A(\Omega_k, \theta)$ and multiplying the numerator and denominator with $|A(\Omega_k, \theta)|^2$, a complete symmetric formulation is found. The input and output have exactly the same role in the problem:

$$V_F(\theta, Z) = \frac{1}{F}\sum_{k=1}^{F} \frac{|A(\Omega_k, \theta)Y(k) - B(\Omega_k, \theta)U(k)|^2}{\sigma_Y^2(k)|A(\Omega_k, \theta)|^2 + \sigma_U^2(k)|B(\Omega_k, \theta)|^2 - 2\mathrm{Re}(\sigma_{YU}^2(k)A(\Omega_k, \theta)\bar{B}(\Omega_k, \theta))} \tag{6-8}$$

6.2.1.3 Measuring the Noise Model. The cost function depends on the exact values $\sigma_U^2(k)$, $\sigma_Y^2(k)$, and $\sigma_{YU}^2(k)$. In practice, these should be obtained from measured data. Section 2.5.1 shows how the sample covariance matrix can easily be extracted from repeated measurements and, later on, it will be shown that it is sufficient to use only four or six repetitions to guarantee that the properties of the estimator are not lost. This means, again, that a fully identifiability procedure can be set up. If the user can apply a periodic excitation, all other information can be extracted automatically without worrying about determining the noise model. If interested, however, the user can use this information to evaluate the quality of the experiments before starting the actual identification. For example, by examining the measured FRF together with its uncertainty, the complexity of the problem and the quality of the measurements can be revealed.

6.2.1.4 Dealing with Exactly Known Inputs. In some applications (e.g., control problems) a model that links the output of the process directly to the digital controller output is built. In these cases the input signal is exactly known because it is stored in the memory of a computer. The errors-in-variables approach is automatically adapted to this situation by putting $\sigma_U^2(k)$ and $\sigma_{YU}^2(k)$ equal to zero.

6.2.1.5 Starting from Measured FRF. Sometimes the user has only the measured FRF available. In that case it is still possible to use the previous approach by putting $Y(k) = G(\Omega_k)$, and $\sigma_Y^2(k) = \sigma_G^2(k)$, the input is set to $U(k) = 1$, with $\sigma_U^2(k) = 0$. The variance $\sigma_G^2(k)$ can be obtained directly from the coherence as explained in Section 2.5.4.

6.2.1.6 Properties. The properties of the estimator are studied in detail in the next chapter, and it is shown that under weak conditions the estimates converge (for an increasing number of data points) to the parameters θ_* that would be found in the noiseless case. The uncertainty on the estimates approaches the smallest possible level for estimates without systematic errors. The covariance matrix $\text{Cov}(\hat{\theta})$ can be calculated at the end of the identification process. Starting from $\text{Cov}(\hat{\theta})$, it is easy to generate uncertainty bounds on other θ-dependent quantities; for example, for the FRF of the transfer function we get that

$$\text{var}(G(\Omega, \hat{\theta})) \approx \left(\frac{\partial G(\Omega, \theta)}{\partial \theta}\right) \text{Cov}(\hat{\theta}) \left(\frac{\partial G(\Omega, \theta)}{\partial \theta}\right)^H \Bigg|_{\theta = \mathcal{E}\{\hat{\theta}\}} \tag{6-9}$$

(see Section 14.2). In practice the derivatives are evaluated in the estimated value $\hat{\theta}$. Also, the uncertainty bounds on the poles and zeros (see Section 9.2.3) or on the residuals (difference between measured and modeled FRF) (Section 9.2.2) can be obtained.

6.3 GENERATING STARTING VALUES

The cost function (6-8) is highly nonlinear in the parameters θ because they appear in the numerator and denominator. As a result, the minimization of the cost becomes quite difficult. It is possible to solve the problem analytically only in extremely simple cases. In all other situations a numerical search procedure is needed. The convergence of these methods depends strongly on the generation of good starting values for the model parameters θ. In general, it is impossible to get this information from physical principles because the link between the coefficients of the transfer function and the underlying physical systems is very nonlinear, especially for higher order systems. Moreover, often the user does not want to make the effort to collect all the required knowledge because the goal of the experiment is to generate a black box model that describes the input-output behavior. For that reason we need self-starting algorithms that generate the starting values from the measured data and not from unavailable prior knowledge.

A possibility to make the optimization self-starting is to change the cost function in the first step so that its global minimum can be calculated directly. There are a number of possibilities to reach this goal. The simplest solution is just to remove the denominator in (6-8) so that the problem becomes linear-in-the-parameters and the minimum is found by solving a linear set of equations (see Section 7.8.2). The disadvantage of this straightforward approach is that the solution becomes extremely noise sensitive. For that reason attempts were made to make a parameter-independent reconstruction of the denominator of (6-8) using measurement information only (Section 7.12.4). This results in significantly improved starting values. A second possibility to generate starting values is to continue with the nonlinear cost function but to modify it such that the global minimum can easily be found using advanced, but widely available, numerical techniques such as singular value decomposition. This leads to the generalized, total least squares type of solutions (see Section 7.10.3) that minimize a cost function of the form

$$V_F(\theta, Z) = \frac{\sum_{k=1}^{F} |A(\Omega_k, \theta)Y(k) - B(\Omega_k, \theta)U(k)|^2}{\sum_{k=1}^{F} \sigma_Y^2(k)|A(\Omega_k, \theta)|^2 + \sigma_U^2(k)|B(\Omega_k, \theta)|^2 - 2\mathrm{Re}(\sigma_{YU}^2(k)A(\Omega_k, \theta)\bar{B}(\Omega_k, \theta))} \tag{6-10}$$

Although the efficiency of this method is lower than that of the original MLE, it provides good candidate starting values. Again, it is possible to improve the quality by adding a nonparametric frequency weighting as explained in the next chapter. A third possibility to get starting values is to use subspace methods (see Section 7.14) that are based on state space models. Compared with the previous algorithms, this method is less flexible because it is not possible to choose the number of poles different from the numbered zeros; but despite this disadvantage, good quality starting values are generated. A major advantage of subspace methods is that they are very well suited to multiinput, multioutput (MIMO) problems.

As a general procedure, we advise the reader to combine these techniques by calculating two or three candidate starting values and to select, out of these, the solution that results in the smallest MLE cost (6-8).

6.4 DIFFERENCES FROM AND SIMILARITIES TO THE "CLASSICAL" TIME DOMAIN IDENTIFICATION FRAMEWORK

Identification has a long tradition. Over the years, the attention shifted almost completely to the use of discrete time models that were identified starting from arbitrary (no periodicity required) excitations. The major difference with the preceding approach is that a parametric noise model is used (Ljung, 1999; see also Section 8.9). Ljung (1999) gives a frequency domain interpretation of the cost function that is minimized with these techniques. By neglecting the leakage effects, the following equivalent frequency domain representation of the time domain cost is found:

$$V_N(\theta, Z) \approx \frac{1}{N} \sum_{k=0}^{N-1} \frac{|Y(k) - G(z_k^{-1}, \theta)U(k)|^2}{|H(z_k^{-1}, \theta)|^2} \tag{6-11}$$

where $|H(z^{-1}, \theta)|^2$ is a parametric model for the power spectrum of the process noise. These methods work well if the measurement noise $(M_U(k), M_Y(k))$ is negligible, otherwise the results will be prone to systematic errors. The major advantage of this approach is that no periodic signals are needed. Its major disadvantage is the need to estimate an additional model $H(z^{-1}, \theta)$. A more detailed discussion is given in Section 8.9.

6.5 EXTENSIONS OF THE MODEL: DEALING WITH UNKNOWN DELAYS AND TRANSIENTS

In the previous sections, the simplest model was used. The results can be generalized to systems with an unknown delay τ. To do so, the model is extended to $G(\Omega, \theta)e^{-\tau s}$ for continuous time systems or to $z^{-\tau/T_s}G(z^{-1}, \theta)$ for discrete-time systems (see Section 5.2). The reader has to realize that the corresponding optimization problem is much more difficult to solve because it is very sensitive to local minima. Consequently, a good starting value of the delay is needed.

Another generalization is the extension of the model to include transients (before the system reaches its steady-state behavior) or to cover, also, the situation with arbitrary (nonperiodic) excitations. Again, this is simply solved by adding an additional rational term to the model (Section 5.3.2):

$$G(\Omega, \theta) + \frac{I(\Omega, \theta)}{A(\Omega, \theta)} \tag{6-12}$$

Because the additional rational term has the same denominator, the complexity of the numerical optimization process is almost not affected by this generalization. A similar extension can be used to process experiments with missing data (Section 5.3.3).

7

Estimation with Known Noise Model

Abstract: This chapter gives an overview of frequency domain identification methods for single input, single output systems. Estimators such as the (weighted) linear least squares, the weighted nonlinear least squares, the maximum likelihood, the (weighted) total least squares, the instrumental variables, and the subspace algorithms are discussed in detail. The interrelations between the different approaches are highlighted through a study of the (equivalent) cost functions. Special attention is also paid to global minimizers that try to approximate the maximum likelihood estimator. The properties of the different approaches are illustrated by means of an "on-line" simulation example. The chapter ends with an overview of the properties of the estimators and a brief discussion of the particularities of estimating high-order systems, systems with time delay, systems in feedback, systems with missing data, multivariable systems, and transfer function models with complex coefficients.

7.1 INTRODUCTION

In this chapter we handle the identification of the plant model assuming that the noise model is known exactly. We give an overview of frequency domain identification methods for single input, single output systems (Sections 7.8 to 7.15). Afterward, the particularities of high-order systems (Section 7.16), systems with time delay (Section 7.17), systems in feedback (Section 7.18), the missing data problem (Section 7.20), and multivariable systems (Section 7.21) are discussed. A second-order system $G(s, \theta) = 1/(1 + s + s^2)$ is used as an "on-line" illustration through Sections 7.8 to 7.14. Figure 7-1 shows the true transfer function and the simulated noisy frequency response data (see Appendix 7.A for more information concerning the generation of the simulation data). Readers who want only a quick taste of the basics of frequency domain estimation (and accept the claimed properties as they are) may skip the last paragraph of Section 7.4 and Sections 7.5 to 7.7 but should still go through Sections 7.2 and 7.3 before tackling the description of the estimators (Sections 7.8 to 7.15).

Before starting with the overview, we discuss the type of data (experiments) we can handle (Section 7.2), introduce some notations for the parametric plant models (Section 7.3), and present the general form of the identification algorithms (Section 7.4). Section 7.5, quick tools to analyze estimators, is intended for readers who are not interested in the technical

System Identification: A Frequency Domain Approach. Rik Pintelon and Johan Schoukens
Copyright © 2001 Institute of Electrical and Electronics Engineers. ISBN: 0-780-36000-1

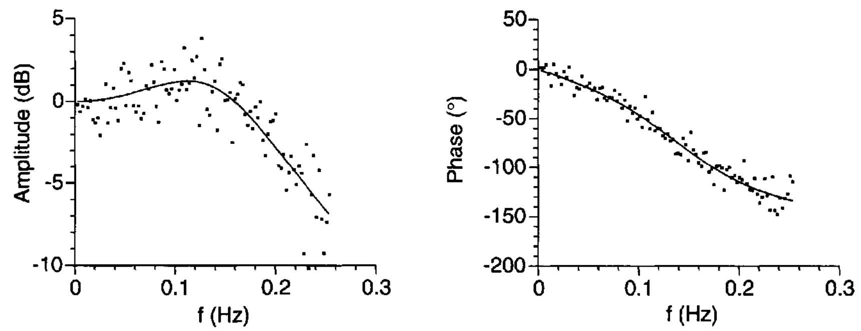


Figure 7-1. Second-order example $G(s, \theta) = 1/(1+s+s^2)$: true transfer function (solid line) and simulated noisy data (dots).

details of the proofs but still want to get some insight into the derivation of some basic properties. Combined with Section 7.7, which discusses the general asymptotic properties of estimators minimizing a cost function that is quadratic-in-the-measurements, it will allow them to easily verify and understand the properties of the different estimators described in Sections 7.8 to 7.14. Those who are interested in the technical details will find a comprehensive list of the basic assumptions needed to prove the asymptotic properties of the estimators (Section 7.6). The proofs of the theorems are given in the Appendix and rely completely on the results of Chapters 15, 16, and 17. The reader is referred to these chapters for more background information concerning the way properties are proved.

7.2 FREQUENCY DOMAIN DATA

The identification starts from measured input-output discrete Fourier transform (DFT) spectra $U(k)$, $Y(k)$,

$$\begin{aligned} Y(k) &= Y_0(k) + N_Y(k) \\ U(k) &= U_0(k) + N_U(k) \end{aligned} \tag{7-1}$$

with $U_0(k)$, $Y_0(k)$ the true unknown values, or from a measured frequency response function $G(\Omega_k)$,

$$G(\Omega_k) = G_0(\Omega_k) + N_G(k) \tag{7-2}$$

with $G_0(\Omega_k)$ the true unknown value, at a set of F frequencies Ω_k, $k = 1, 2, ..., F$, which may be a (sub)set of the DFT frequencies. Note that (7-2) is a special case of (7-1) with $Y(k) = G(\Omega_k)$ and $U(k) = 1$. The $2F$ complex-valued vector Z contains the measured input-output (DFT) spectra

$$Z^T = [Z^T(1)Z^T(2)...Z^T(F)] \text{ with } Z^T(k) = [Y(k)U(k)] \tag{7-3}$$

where $k = 1, 2, ..., F$. It is related to the true values by $Z = Z_0 + N_Z$, where the disturbing noise N_Z has zero mean and is independent of Z_0.

The frequency domain data (7-1), (7-2) can be obtained via time domain or frequency domain experiments. In a *time domain experiment* a broadband random or normalized periodic (see Definition 3.4) excitation is applied to the plant and N samples of the input and output signals are measured (see Figure 7-2). For the periodic signals the steady-state response is observed over an integer number of periods. These N input-output samples are transformed to the frequency domain using the discrete Fourier transform. $F \leq N/2+1$ DFT frequencies of the input and output DFT spectra are used for the identification. For arbitrary signals $u_g(t)$ the generator noise $n_g(t)$ is a part of the excitation, $u_0(t) = u_g(t) + n_g(t)$, so that the frequency domain errors $N_U(k)$ and $N_Y(k)$ in (7-1) are related to the disturbing noise sources in Figure 7-2 as

$$\begin{aligned} N_Y(k) &= \mathrm{DFT}(m_y(t) + n_p(t)) \\ N_U(k) &= \mathrm{DFT}(m_u(t)) \end{aligned} \qquad (7\text{-}4)$$

For periodic signals the generator noise $n_g(t)$ is a disturbing noise source, $u_0(t) = u_g(t)$, which causes a correlation between the input and output errors. Indeed, the frequency domain errors $N_U(k)$ and $N_Y(k)$ in (7-1) are then related to the disturbing noise sources in Figure 7-2 as

$$\begin{aligned} N_Y(k) &= \mathrm{DFT}(n_g(t) * g_0(t) + n_p(t) + m_y(t)) \\ N_U(k) &= \mathrm{DFT}(n_g(t) + m_u(t)) \end{aligned} \qquad (7\text{-}5)$$

with * the convolution operator and $g_0(t)$ the impulse response of the plant. In a *frequency domain experiment*, a single sine excitation is applied to the plant and the input-output spectra of the steady-state response are measured at the excited frequency. This experiment is repeated at F different frequencies. For example, high-frequency network analyzers (microwave measurements) and impedance analyzers follow this measurement procedure. Also, most dynamic signal analyzers have such a measurement mode. The frequency domain errors $N_U(k)$ and $N_Y(k)$ in (7-1) are related to the noise sources in Figure 7-3 as

$$\begin{aligned} N_Y(k) &= N_g(k) G_0(\Omega_k) + M_Y(k) + N_P(k) \\ N_U(k) &= N_g(k) + M_U(k) \end{aligned} \qquad (7\text{-}6)$$

with $G_0(\Omega_k)$ the plant transfer function.

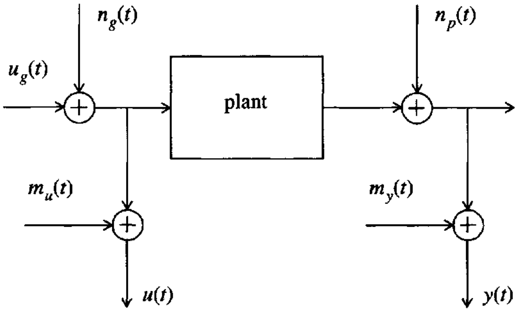


Figure 7-2. Time domain experiment: a broadband excitation $u_g(t)$ is applied to the plant. The DFT spectra of N observed input-output samples are calculated. $F = O(N)$ DFT frequencies of the input-output DFT spectra are retained. $n_g(t)$ is the generator noise, $m_u(t)$ and $m_y(t)$ are the input and output measurement errors, and $n_p(t)$ is the process noise.

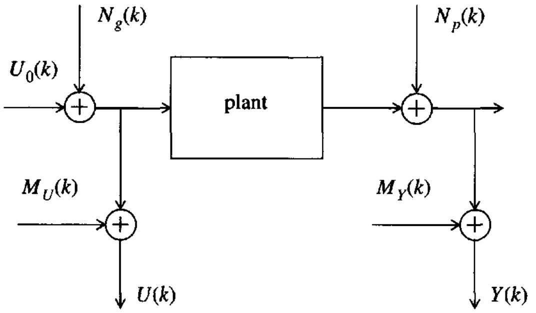


Figure 7-3. Frequency domain experiment: a single sine excitation $u_g(t) = A\sin(2\pi f_k t + \phi)$ is applied to the plant and the input-output spectra of the steady-state response are measured at frequency f_k. This experiment is repeated at F different frequencies. $N_g(k)$ is the generator noise, $M_U(k)$ and $M_Y(k)$ are the input and output measurement errors, and $N_p(k)$ is the process noise.

Due to the imperfections of the measurement devices, it is recommended not to use measurements at DC and in the neighborhood of the Nyquist frequency. Indeed, acquisition units mostly introduce DC offset errors and antialias protection is mostly guaranteed only up to about 80% of the Nyquist frequency. The measurements can also be the result of a linearization of a nonlinear system at an operating point. This will introduce DC values in the input and output signals that are not compatible with the linear model and, hence, should be removed.

An important question asked when (re)designing an experiment is: "What will happen with the estimates (uncertainty, bias, ...) if one gathered, for example, four times more data?" Ideally, we would like to answer this question for each finite value of F. Except for the (weighted) linear least squares, this is possible only for "sufficiently large" values of F. To analyze the stochastic properties of the estimators for F "sufficiently large" we will make a mental experiment where the number of frequencies F tends to infinity. For a frequency domain experiment this implies that the number of single sine measurements F tends to infinity, while for a time domain experiment this implies that the number of measured time domain samples N tends to infinity. Note that we do not consider time domain experiments ($N \to \infty$) with periodic signals containing a fixed number (independent of N) of frequencies F. Indeed, for such experiments the signal-to-noise ratio tends to infinity as $N \to \infty$ at the excited DFT frequencies (see Appendix 7.C), and, hence, all the estimators considered in this chapter would be consistent in a trivial manner. For random and normalized multisine ($F = O(N)$, see Definition 3.2) excitations the signal-to-noise ratio per spectral line remains an $O(N^0)$ (see Appendices 7.B and 7.C) so that consistency is a nontrivial issue.

7.3 PLANT MODEL

Unless mentioned otherwise, we will assume in this chapter that the parameterization of the plant model is identifiable (see Definition 5.8). It implies that the parameter vector θ contains only the free parameters of the model, for example, all the numerator and denominator coefficients of the rational form $G(\Omega, \theta) = B(\Omega, \theta)/A(\Omega, \theta)$ except $a_0 = 1$. Note, however, that from a numerical point of view it is often better to use the full overparameterized form in combination with dedicated numerical methods. Chapter 18 discusses this issue in detail.

For any parameterization of Sections 5.2 and 5.3 (rational form, partial fraction expansion, and state space representation) we can use the *output error*, which is the difference between the observed output $Y(k)$ and the modeled output $Y(\Omega_k, \theta)$. From transfer function models (5-32), (5-35), and (5-36) we get

$$Y(\Omega_k, \theta) = G(\Omega_k, \theta)U(k) \tag{7-7}$$

for periodic signals ($\Omega = z^{-1}$, s, $\sqrt{s}$ or $\tanh(\tau_R s)$) and

$$Y(\Omega_k, \theta) = G(\Omega_k, \theta)U(k) + T(\Omega_k, \theta) \tag{7-8}$$

for arbitrary excitations ($\Omega = z^{-1}$ or s).

For the *rational forms*, (5-20), (5-25), (5-37), and (5-40), it is convenient also to introduce the *equation error* $e(\Omega_k, \theta, Z(k))$, which is the difference between the left- and right-hand sides of transfer function models (5-32), (5-35), and (5-36) after multiplication by $A(\Omega_k, \theta)$. We get

$$e(\Omega_k, \theta, Z(k)) = A(\Omega_k, \theta)Y(k) - B(\Omega_k, \theta)U(k) \tag{7-9}$$

for periodic signals ($\Omega = z^{-1}$, s, $\sqrt{s}$ or $\tanh(\tau_R s)$) and

$$e(\Omega_k, \theta, Z(k)) = A(\Omega_k, \theta)Y(k) - B(\Omega_k, \theta)U(k) - I(\Omega_k, \theta) \tag{7-10}$$

for arbitrary excitations ($\Omega = z^{-1}$ or s). The equation error $e(\Omega_k, \theta, Z(k))$ is not exactly zero because the observations $Y(k)$ and $U(k)$ are disturbed by noise and θ does not equal the true value θ_0 (if it exists).

Note that (7-8) and (7-10) are valid only at a (sub)set of the DFT frequencies but (7-7) and (7-9) are also valid at arbitrary (not related to a DFT grid) frequencies.

7.4 ESTIMATION ALGORITHMS

Most algorithms discussed in this chapter minimize (in each step) a "quadratic-like" cost function $V(\theta, Z)$

$$V(\theta, Z) = \varepsilon^H(\theta, Z)\varepsilon(\theta, Z) = \sum_{k=1}^{F} |\varepsilon(\Omega_k, \theta, Z(k))|^2 \tag{7-11}$$

where $\varepsilon(\theta, Z) \in \mathbb{C}^F$ is some kind of measure of the difference between the measurements and the model. The residual $\varepsilon(\theta, Z) \in \mathbb{C}^F$ is a (non)linear vector function of the model parameters θ and the measurements Z. Note that $\varepsilon_{[k]}(\theta, Z) = \varepsilon(\Omega_k, \theta, Z(k))$ depends only on the measurements at frequency Ω_k.

A first important subclass of (7-11) consists of the cost functions $V(\theta, Z)$, which are *quadratic-in-the-measurements* Z. For these cost functions the residual $\varepsilon(\theta, Z)$ is linear in Z and can be written as

$$\varepsilon(\theta, Z) = \varepsilon(\theta, Z_0) + \Delta(\theta, N_Z) \tag{7-12}$$

with $\Delta_{[k]}(\theta, N_Z) = \Delta(\Omega_k, \theta, N_Z(k))$ and $\Delta(\Omega_k, \theta, 0) = 0$. Hence, (7-11) becomes

$$\begin{aligned} V(\theta, Z) &= V(\theta, Z_0) + v(\theta, N_Z) + 2\mathrm{Re}(\varepsilon^H(\theta, Z_0)\Delta(\theta, N_Z)) \\ v(\theta, N_Z) &= \Delta^H(\theta, N_Z)\Delta(\theta, N_Z) \end{aligned} \tag{7-13}$$

where $v(\theta, N_Z)$ represents that part of the cost function depending on N_Z only.

A second important subclass of (7-11) consists of the cost functions $V(\theta, Z) = f(\theta, \eta(Z), Z)$, which depend on an initial guess $\eta(Z)$ of the model parameters,

$$V(\theta, Z) = \varepsilon^H(\theta, Z)\varepsilon(\theta, Z) = \sum_{k=1}^{F} |\varepsilon(\Omega_k, \theta, \eta(Z), Z(k))|^2 \tag{7-14}$$

and which are quadratic-in-the-measurements Z when $\eta(Z)$ in (7-14) is replaced by a nonrandom vector η.

Often, a Newton-Gauss type of algorithm is used to find the minimizer $\hat{\theta}(Z)$ of (7-11). Rewriting (7-11) as $V(\theta, Z) = \varepsilon_{\mathrm{re}}^T(\theta, Z)\varepsilon_{\mathrm{re}}(\theta, Z)$, where $(\ )_{\mathrm{re}}$ stacks the real and imaginary parts on top of each other (see Section 13.8),

$$\varepsilon_{\mathrm{re}}(\theta, Z) = \begin{bmatrix} \mathrm{Re}(\varepsilon(\theta, Z)) \\ \mathrm{Im}(\varepsilon(\theta, Z)) \end{bmatrix} \tag{7-15}$$

the ith iteration step of this algorithm is given by (see also Section 1.5.1)

$$J_{\mathrm{re}}^T(\theta^{(i-1)}, Z)J_{\mathrm{re}}(\theta^{(i-1)}, Z)\Delta\theta^{(i)} = -J_{\mathrm{re}}^T(\theta^{(i-1)}, Z)\varepsilon_{\mathrm{re}}(\theta^{(i-1)}, Z) \tag{7-16}$$

with $\Delta\theta^{(i)} = \theta^{(i)} - \theta^{(i-1)}$ and $J(\theta, Z) = \partial\varepsilon(\theta, Z)/\partial\theta$ the Jacobian of the vector $\varepsilon(\theta, Z)$. Using complex numbers, (7-16) can be written as

$$\mathrm{Re}(J^H(\theta^{(i-1)}, Z)J(\theta^{(i-1)}, Z))\Delta\theta^{(i)} = -\mathrm{Re}(J^H(\theta^{(i-1)}, Z)\varepsilon(\theta^{(i-1)}, Z)) \tag{7-17}$$

If the algorithm converges to the global minimum, then $\hat{\theta}(Z) = \theta^{(\infty)}$. When identifying continuous-time systems in the s- and $\sqrt{s}$-domains, it is indispensable to scale the frequency axis (and, hence, also the parameters) to guarantee the numerical stability of the normal equations (7-16). Without scaling, identification in the s- and $\sqrt{s}$-domains is often impossible with the available computing precision, even for modest orders of the transfer function. Although the scale factor that minimizes the condition number of $J_{\mathrm{re}}(\theta^{(i-1)}, Z)$ is plant and model dependent, a good compromise is to use the arithmetic mean of the maximum and minimum angular frequencies in the frequency band of interest: $\omega_{\mathrm{scale}} = (\omega_{\mathrm{max}} + \omega_{\mathrm{min}})/2$. For example, the term $a_m s^m$ becomes $a_m \omega_{\mathrm{scale}}^m (s/\omega_{\mathrm{scale}})^m$ after scaling and $a_m \omega_{\mathrm{scale}}^m$ is estimated. The numerical stability can still be improved by solving the overdetermined set of equations

$$J_{\mathrm{re}}(\theta^{(i-1)}, Z)\Delta\theta^{(i)} = -\varepsilon_{\mathrm{re}}(\theta^{(i-1)}, Z) \tag{7-18}$$

instead of (7-16), for example, using the singular value decomposition or a QR factorization (see Section 13.13). The convergence region of the Newton-Gauss algorithm can be enlarged by using a Levenberg-Marquardt version of (7-16) and (7-18) (see Fletcher, 1991 and Section 7.L.4 of Appendix 7.L).

To study the asymptotic behavior of the identification algorithms, it is convenient to scale the cost function with the number of frequencies, $V_F(\theta, Z) = V(\theta, Z)/F$, $v_F(\theta, N_Z) = v(\theta, N_Z)/F$, and $f_F(\theta, \eta(Z), Z) = f(\theta, \eta(Z), Z)/F$. The expected values of the cost function $V_F(\theta) = \mathcal{E}\{V_F(\theta, Z)\}$ and its minimizer $\tilde{\theta}(Z_0)$ play an important role in the convergence analysis of the estimate $\hat{\theta}(Z)$. All the asymptotic properties ($F \to \infty$) of the estimate $\hat{\theta}(Z)$ will be formulated w.r.t. the minimizer $\tilde{\theta}(Z_0)$ of the expected value of the cost function. The conditions under which $\hat{\theta}(Z)$ converges to $\tilde{\theta}(Z_0)$ will be studied. This is a stochastic convergence problem that mainly depends on the disturbing noise properties. When model errors are present, $\tilde{\theta}(Z_0)$ will vary as the number of frequencies F increases. We may wonder then whether $\tilde{\theta}(Z_0)$ converges to some limit value $\theta_* = \lim_{F \to \infty} \tilde{\theta}(Z_0)$ which is the minimizer of the limit cost function $V_*(\theta) = \lim_{F \to \infty} V_F(\theta)$. This is a deterministic convergence problem that depends on the way data (frequencies) are added in the time or frequency domain experiment. The notations introduced are summarized in Table 7-1.

TABLE 7-1 Overview of Notations Frequently Used: $\eta(Z)$ is an (Initial) Estimate of the Model Parameters and η_* is its Limit Value

Cost function	$V_F(\theta, Z)$, $f_F(\theta, \eta(Z), Z)$	$V_F(\theta) = \mathcal{E}\{V_F(\theta, Z)\}$, $V_F(\theta) = \mathcal{E}\{f_F(\theta, \eta_*, Z)\}$	$V_*(\theta) = \lim_{F \to \infty} V_F(\theta)$
Minimizer	$\hat{\theta}(Z)$	$\tilde{\theta}(Z_0)$	θ_*

7.5 QUICK TOOLS TO ANALYZE ESTIMATORS

The minimum we can expect from a "sound" estimator is that in the noiseless case we get the true answer (correctness property). In the noisy case we should get asymptotically ($F \to \infty$) the true answer (consistency property) and hopefully a "small" uncertainty (efficiency property). We may also wonder whether the estimates depend on the particular parameter constraint chosen ($a_0 = 1$, or $\|\theta\|_2^2 = 1$, or ...), how fast the estimates converge, and what happens with the estimates if the true model does not belong to the considered model set. All these questions are thoroughly studied in this chapter.

Some of the previously raised questions can easily be analyzed using the following quick tools. The first step in the analysis consists of calculating the (equivalent) cost function $V(\theta, Z)$ of the identification method. Next we verify the following:

1. *(Asymptotic) correctness:* assuming that the true model belongs to the model set, the identification algorithm is (asymptotically) correct if it produces the true model for an (in)finite amount of noiseless ($N_Z = 0$) data. This is true if $V_F(\theta, Z_0)$ ($\lim_{F \to \infty} V_F(\theta, Z_0)$) is minimal in the true model parameters θ_0. All the identification algorithms of this chapter are correct for transfer function models (7-7) with $\Omega = z^{-1}$, s, $\sqrt{s}$, or $\tanh(\tau_R s)$ and (7-8) with $\Omega = z^{-1}$, where $G(\Omega, \theta)$ and $T(\Omega, \theta)$ can take any parameterization of Sections 5.2 and 5.3. They are asymptotically correct for continuous-time models using arbitrary excitations, model (7-8) with $\Omega = s$.

2. *Consistency:* the (equivalent) cost function minimized by most identification methods in this chapter is a quadratic function of the measurements Z. The expected value of such cost functions can be written as

$$V_F(\theta) = \mathcal{E}\{V_F(\theta, Z)\} = \mathcal{E}\{V_F(\theta, Z_0)\} + \mathcal{E}\{v_F(\theta, N_Z)\} \tag{7-19}$$

   (see (7-13), Z_0 and N_Z are independent). A necessary condition for consistency is that the limit of the expected value of the cost function $V_*(\theta) = \lim_{F\to\infty} V_F(\theta)$ is minimal in θ_0 (Theorem 15.15). It follows from (7-19) that this condition is satisfied if $\mathcal{E}\{v_F(\theta, N_Z)\}$ is a θ-independent constant. Hence, for correct methods we have $\tilde{\theta}(Z_0) = \theta_0$, while for asymptotically correct methods $\theta_* = \theta_0$. For cost functions of the form (7-14), we replace $\eta(Z)$ by its limit value η_* before taking the expected value of the cost function. The same analysis is then performed on $V_F(\theta) = \mathcal{E}\{f_F(\theta, \eta_*, Z)\}$.

3. *Convergence to the noiseless solution:* if model errors exist, for example, because of a wrong choice of the order of the numerator and/or denominator polynomials, or because a true linear lumped model simply does not exist, then $\hat{\theta}(Z)$ converges to $\tilde{\theta}(Z_0) \neq \theta_0$. Under some conditions, the value $\tilde{\theta}(Z_0)$ is independent of the noise level of the measurements. To verify this, we replace C_{N_Z} by $\upsilon^2 C_{N_Z}$ in the cost function (7-19), with υ a real number. If this transforms $\mathcal{E}\{V_F(\theta, Z_0)\}$ into $f(\upsilon^2)\mathcal{E}\{V_F(\theta, Z_0)\}$ and if $\mathcal{E}\{v_F(\theta, N_Z)\}$ is a θ-independent constant then $\tilde{\theta}(Z_0)$, and, hence, also $\theta_* = \lim_{F\to\infty} \tilde{\theta}(Z_0)$ (if it exists), is independent of the noise level υ. This is true for any υ and, hence, also for $\upsilon \to 0$, which defines, asymptotically, the noiseless solution. Note, however, that the noiseless solution $\tilde{\theta}(Z_0)$ defined in this way may still depend on the noise coloring and the noise covariance matrix C_{N_Z}, for example, the ratio of the output variance $\sigma_Y^2(k)$ to the input variance $\sigma_U^2(k)$ (see Section 7.11). For cost functions of the form (7-14), the analysis is performed on $V_F(\theta) = \mathcal{E}\{f_F(\theta, \eta_*, Z)\}$ and the same conclusions hold if η_*, the limit value of $\eta(Z)$, is independent of υ.

4. *Dependence on the parameter constraint:* from a numerical point of view it is also handy that the estimate of the plant transfer function $G(\Omega_k, \hat{\theta}(Z))$ is independent of the particular parameter constraint chosen, for example, $a_i = 1$, or $b_j = 1$, or $\|\theta\|_2^2 = 1$ ... Indeed, if we fix a zero coefficient to one, then the normal equations (7-16) become ill conditioned. To avoid this problem, it is better to use the constraint $\|\theta\|_2^2 = 1$ (see also Chapter 18). The estimated plant model $G(\Omega, \hat{\theta}(Z))$ is independent of the parameter constraint chosen if, for any $\lambda \neq 0$, $V_F(\lambda\theta, Z) = V_F(\theta, Z)$, with θ the full overparameterized form (proof: see Chapter 18).

5. *Numerical reliability of the normal equations:* the Hessian of the expected value of the cost function has full rank in the true parameter values: $\text{rank}(V_F''(\theta_0)) = \dim(\theta) =$ number of free model parameters (' is the derivative w.r.t. θ). If the Hessian is not of full rank, then the cost function cannot be approximated by a quadratic function in the neighborhood of the solution θ_0. This is problematic for most of the nonlinear minimization algorithms.

6. *Influence of the noise level and the model errors:* to study the influence of small measurement errors, we replace N_Z by υN_Z and C_{N_Z} by $\upsilon^2 C_{N_Z}$ and analyze the expression for $\upsilon \to 0$. Model errors are present if $e(\tilde{\theta}(Z_0), Z_0) \neq 0$. To study the influence of small model errors we replace $e(\tilde{\theta}(Z_0), Z_0)$ by $\mu e(\tilde{\theta}(Z_0), Z_0)$ and analyze the expressions for $\mu \to 0$.

7.6 ASSUMPTIONS

In this section we give an overview of all the assumptions required to analyze the asymptotic $(F \to \infty)$ behavior of the estimate $\hat{\theta}(Z)$. They are grouped per property in increasing order of complexity: stochastic convergence, stochastic convergence rate, systematic and stochastic errors, consistency, asymptotic bias, asymptotic normality, and asymptotic efficiency. Hereby, we make the distinction between a time and a frequency domain experiment because the signal and disturbing noise properties are easiest to describe in the respective domains. It allows the reader to verify, easily, what kind of assumptions are required for a particular property and experiment in each theorem of this chapter.

The cost function (7-11) and its higher order derivatives w.r.t. θ may not exist for some values of the model parameters θ. To avoid the resulting technical difficulties in the proof of the theorems, a regular set Θ_r of θ-values is constructed where $V_F(\theta, Z)$ and its higher order derivatives exist and are finite. By construction, we make this set closed and bounded compact. The minimizer of (7-11) is then defined as

$$\hat{\theta}(Z) = \arg\min_{\theta \in \Theta_r} V_F(\theta, Z) \tag{7-20}$$

(for the maximum likelihood estimation of ARMAX models the compactness assumption of the parameter space can be avoided, see Hannan and Deistler (1988)).

The properties of (7-20) will be studied using the results of Chapter 15 for Sections 7.8.2, 7.9, and 7.10; of Chapter 16 for Sections 7.8.3, 7.12, and 7.14; and of Chapter 17 for Section 7.11. The reader is referred to these chapters for detailed background information concerning the proof of the theorems. There she or he will also find answers to questions such as "Why do we need a particular assumption and what is it used for?" and "What is the main philosophy behind the proof of a particular property?" Other basic questions such as "Which statistical tools are available?" and "How should they be used?" are tackled in Chapter 14.

7.6.1 Stochastic Convergence

To show the convergence $(F \to \infty)$ of the estimator $\hat{\theta}(Z)$ (7-20) to $\tilde{\theta}(Z_0)$ we need conditions on the excitation signal, the disturbing noise, and the cost function. For example, the persistence of excitation Assumption 7.7 requires that the excitation signal satisfies, at least, the identifiability conditions of Section 5.5. Note that we do not require the existence of a true model.

Assumption 7.1 (Excitation Signal—Time Domain Experiment): The excitation $u(t)$ either is a normalized periodic signal (see Definitions 3.2, 3.3, and 3.4) or can be written at the sampling instances as filtered white noise $u(t) = H_u(q)e_u(t)$, where $H_u(z^{-1})$ is a stable rational filter. $e_u(t)$ is independently distributed and has stationary first- and second-order moments and uniformly bounded fourth-order moments. For periodic excitations the input-output signals of the steady-state response are observed over an integer number of periods. N samples of the input and output signals are transformed to the frequency domain using the DFT. $F \leq N/2 + 1$ DFT frequencies of the input-output DFT spectra are used for the identification. The number of selected frequencies F is proportional to N: $F = O(N)$.

In classical time domain system identification the excitation signal $u(t)$ should be *quasi-stationary* (Ljung, 1999), which means that

$$\begin{aligned} &\mathcal{E}\{u(t)\} = \mu_u(t) && |\mu_u(t)| \le c_1 < \infty \\ &\mathcal{E}\{u(t)u(r)\} = R_{uu}(t, r) && |R_{uu}(t, r)| \le c_2 < \infty \\ &R_{uu}(\tau) = \lim_{N \to \infty} \frac{1}{N}\sum\nolimits_{t=1}^{N} R_{uu}(t, t-\tau) && \end{aligned} \tag{7-21}$$

should be satisfied for any t, r, and τ, with c_1, c_2 constants independent of t, r. The class of excitation signals defined by Assumption 7.1 forms a subset of the class of quasi-stationary signals (7-21) and, hence, is less general (see Exercise 7.1). This restriction is the price to pay to allow noncausal filtering (removal of DFT frequencies) of the input and output DFT spectra. Note that Assumption 7.1 is easily met if the excitation stems from an arbitrary waveform generator.

Assumption 7.2 (Excitation Signal—Frequency Domain Experiment): The plant is measured in steady state with a single sine excitation. This experiment is repeated at F different frequencies $f_{\min} \le f_k \le f_{\max}$, $k = 1, 2, \ldots, F$, with $f_{\min}$ and $f_{\max} < \infty$ respectively the minimum and maximum excitation frequencies.

Assumption 7.3 (Disturbing Noise—Time Domain Experiment): At the sampling instances the disturbing time domain noise sources $n_y(t)$, $n_u(t)$ are jointly correlated filtered white noise sequences

$$\begin{bmatrix} n_y(t) \\ n_u(t) \end{bmatrix} = \begin{bmatrix} H_{11}(q) & H_{12}(q) \\ H_{21}(q) & H_{22}(q) \end{bmatrix} \begin{bmatrix} e_1(t) \\ e_2(t) \end{bmatrix} \quad \text{or } n_z(t) = H(q)e(t) \tag{7-22}$$

with $n_z^T(t) = [n_y(t)\ n_u(t)]$, $e^T(t) = [e_1(t)\ e_2(t)]$ and where $H(z^{-1})$ is a stable filter. $e(t)$ is independently distributed (over t and over its entries) with continuous probability density function, has stationary first- and second-order moments, uniformly bounded fourth-order moments, and is independent of the true (unknown) excitation $u_0(t)$. The frequency domain errors $N_Y(k)$, $N_U(k)$ are related to the time domain errors $n_y(t)$, $n_u(t)$ by the discrete Fourier transform: $N_Y(k) = \text{DFT}(n_y(t))$ and $N_U(k) = \text{DFT}(n_u(t))$.

Assumption 7.4 (Disturbing Noise—Frequency Domain Experiment): The frequency domain errors $N_Y(k)$, $N_U(k)$ are independent (over k), jointly correlated, zero mean random variables with uniformly bounded absolute moments of order four. $N_Y(k)$, $N_U(k)$ are independent of the true (unknown) excitation $U_0(k)$.

Assumption 7.5 (Frequency Domain Errors): The (co-)variances $\sigma_Y^2(k) = \text{var}(N_Y(k))$, $\sigma_U^2(k) = \text{var}(N_U(k))$, and $\sigma_{YU}^2(k) = \text{covar}(N_Y(k), N_U(k))$ of the frequency domain errors $N_Y(k)$, $N_U(k)$ are known.

Assumption 7.6 (Continuity Cost Function): The cost function $V_F(\theta, Z)$ is a continuous function of θ in the compact set Θ_r.

Assumption 7.7 (Persistence of Excitation): There exists an F_0 such that for any $F \ge F_0$, ∞ included, the expected value of the cost function $V_F(\theta) = \mathcal{E}\{V_F(\theta, Z)\}$ has a unique global minimum $\tilde{\theta}(Z_0)$, which is an interior point of Θ_r.

If $V_F(\theta)$ is not convex, then in the presence of model errors $V_F(\theta)$ can have more than one global minimum. An example of this can be found in Kabaila (1983). To handle these cases we restrict the compact set Θ_{r} in Assumption 7.7 such that $V_F(\theta)$ contains a unique global minimum in Θ_{r}.

7.6.2 Stochastic Convergence Rate

When designing a new time or frequency domain experiment based on the results of a previous experiment, one must choose the number of frequencies F. To make a motivated choice it is important to know how fast the difference $\hat{\theta}(Z) - \tilde{\theta}(Z_0)$ converges to zero as $F \to \infty$. To establish the convergence rate of $\hat{\theta}(Z)$ to $\tilde{\theta}(Z_0)$, we need suitable assumptions concerning the first- and second-order derivatives of the cost function w.r.t. θ. We also need a persistence-of-excitation condition that is stronger than Assumption 7.7. In addition to Assumptions 7.1 to 7.6, we require:

Assumption 7.8 (Continuity First- and Second-Order Derivatives Cost Function): The cost function $V_F(\theta, Z)$ has continuous first- and second-order derivatives w.r.t. θ in Θ_{r} for any value of F, ∞ included.

Assumption 7.9 (Persistence of Excitation): There exists an F_0 such that for any $F \geq F_0$, ∞ included, the Hessian of the expected value of the cost function is regular at the unique global minimizer $\tilde{\theta}(Z_0)$, which is an interior point of Θ_{r}: $c_1 I_{n_\theta} \leq V_F''(\tilde{\theta}(Z_0)) \leq c_2 I_{n_\theta}$ where $0 < c_1 \leq c_2 < \infty$ and c_1, c_2 are F-independent constants.

7.6.3 Systematic and Stochastic Errors

A more profound analysis makes it possible to distinguish between the asymptotic behavior of the stochastic and the systematic deviations in the residual $\hat{\theta}(Z) - \tilde{\theta}(Z_0)$. In addition to Assumptions 7.1 to 7.6, 7.8, and 7.9, we require:

Assumption 7.10 (Continuity Third-Order Derivative Cost Function): The cost function has continuous third-order derivatives w.r.t. θ in Θ_{r} for any value of F, ∞ included.

7.6.4 Asymptotic Normality

To calculate uncertainty regions with a given confidence level, we need the probability density function of the estimate $\hat{\theta}(Z)$. A good approximation can be found if the asymptotic distribution function of $\hat{\theta}(Z)$ is known. Whereas the consistency and convergence rate analysis of $\hat{\theta}(Z)$ requires finite moments of order 4, the convergence and the convergence rate analysis of the distribution function of $\hat{\theta}(Z)$ needs the existence of the moments of any order for a time domain experiment and of order 6 for a frequency domain experiment. In addition to Assumptions 7.1 to 7.6 and 7.8 to 7.10, we require:

Assumption 7.11 (Excitation Signal—Time Domain Experiment): The excitation signals $u(t)$ in Assumption 7.1 have finite moments of any order. The excitation noise $e_u(t)$ is independent and identically distributed.

Assumption 7.12 (Disturbing Noise—Time Domain Experiment): The disturbing noise $e(t)$ in Assumption 7.3 is independent and identically distributed with finite moments of any order.

For a frequency domain experiment, these conditions can be relaxed because the successive frequency measurements are independent (see Assumptions 7.2 and 7.4), whereas they are correlated for a time domain experiment (see Assumption 7.3).

Assumption 7.13 (Disturbing Noise—Frequency Domain Experiment): (a) For the asymptotic normality: the disturbing noise N_Z satisfies $\sum_{k=1}^{F} \mathrm{Cov}(N_Z(k)) = O(F)$ and has uniformly bounded absolute moments of order $4 + \varepsilon$ with $\varepsilon > 0$, for example, $\mathcal{E}\{|N_Y(k)|^{4+\varepsilon}\} \le c_1 < \infty$ with c_1 independent of F. (b) For the convergence rate: in addition, the disturbing noise N_z has uniformly bounded absolute moments of order six, for example, $\mathcal{E}\{|N_U(k)|^6\} \le c_2 < \infty$ with c_2 independent of F.

7.6.5 Deterministic Convergence

To study the deterministic convergence and the convergence rate of $\tilde{\theta}(Z_0)$ to θ_*, we must define the strategy of adding new frequencies to the data. We need this information because the model errors depend on the power spectrum of the excitation. In addition to Assumptions 7.1 to 7.6, 7.8, and 7.9, we require:

Assumption 7.14 (Strategy of Adding Frequencies): As $F \to \infty$ the frequencies f_k cover the frequency interval $[f_{\min}, f_{\max}]$ with a density function $n(f)$ defined as

$$n(f) = \lim_{\Delta f \to 0} \lim_{F \to \infty} \frac{N_F(f + \Delta f) - N_F(f)}{F \Delta f} \tag{7-23}$$

where $N_F(f)$ is the number of frequencies in the interval $[0, f]$ when the total number of frequencies is F. The density $n(f)$ is continuous with bounded second-order derivative w.r.t. f in $[f_{\min}, f_{\max}]$ except at a finite number of frequencies.

Special cases are a uniform ($n(f)$ independent of f) or a logarithmic ($n(f)$ is proportional to f^{-1}) distribution of the number frequencies in $[f_{\min}, f_{\max}]$.

Assumption 7.15 (Constraint on the Residual): The second-order derivatives w.r.t. the frequency f of the residual $\mathcal{E}\{|\varepsilon(\Omega(f), \theta, Z(f))|^2\}$ and its first- and second-order derivatives w.r.t. θ, are bounded in the frequency band $[f_{\min}, f_{\max}]$, except at a finite number of frequencies ($\Omega(f) = j2\pi f$, $e^{j2\pi f T_s}$, $\sqrt{j2\pi f}$ or $\tanh(\tau_R j2\pi f)$).

Assumption 7.15 puts some conditions on the limit power spectrum $|U_0(f)|^2$ or $S_{uu}(\omega)$ of the periodic or random excitation; it should be a continuous function of f with bounded second-order derivative.

7.6.6 Consistency

Contrary to the stochastic convergence, consistency can be shown only if a true linear model exists and if it belongs to the considered model set. It also imposes some conditions on the expected value of the cost function, which should be verified for each estimator. To study, under these conditions, the stochastic convergence, the stochastic convergence rate, the improved stochastic convergence rate, and the asymptotic normality, we require, in addition to the assumptions of Sections 7.6.1 to 7.6.4, the following:

Assumption 7.16 (Existence of a True Linear Plant Model): There is an identifiable parameterization $\theta_0 \in \Theta_r$ such that $G(\Omega_k, \theta_0)U_0(k)$, $G(z_k^{-1}, \theta_0)U_0(k) + T(z_k^{-1}, \theta_0)$, or $G(s_k, \theta_0)U_0(k) + T(s_k, \theta_0) + \delta(s_k)$ with $G(s, \theta_0)$ stable represents the true output $Y_0(k)$.

Assumption 7.17 (Consistency Condition on the Cost Function): The expected value of the cost function $V_F(\theta) = \mathcal{E}\{V_F(\theta, Z)\}$, or its limit value $V_*(\theta) = \lim\limits_{F \to \infty} V_F(\theta)$, is minimal in the true model parameters θ_0.

7.6.7 Asymptotic Bias

Speaking about systematic or bias errors makes sense only if a true model exists and if it belongs to the considered model set. Studying the bias is possible only if the expected value of the estimate $\hat{\theta}(Z)$ exists. To ensure the existence of the expected value, we remove "large," "highly improbable" values of $\hat{\theta}(Z)$. This results in the truncated estimate $\underline{\hat{\theta}}(Z)$, which is defined as

$$\underline{\hat{\theta}}(Z) = \begin{cases} \hat{\theta}(Z) & \left\|\hat{\theta}(Z) - \tilde{\theta}(Z_0)\right\|_2 \le L \\ 0 & \left\|\hat{\theta}(Z) - \tilde{\theta}(Z_0)\right\|_2 > L \end{cases} \tag{7-24}$$

where L is an (arbitrarily) large number $(0 < L < \infty)$ independent of F. Lemma 15.27 guarantees that there exists an F_0 such that for any $F \ge F_0$ $\underline{\hat{\theta}}(Z) = \hat{\theta}(Z)$ with probability one (in probability). We require that Assumptions 7.1 to 7.6, 7.8 to 7.10, 7.16, and 7.17 are valid.

7.6.8 Asymptotic Efficiency

A basic step in the analysis of the asymptotic efficiency of the estimate $\hat{\theta}(Z)$ is the calculation of the Fisher information matrix. It inherently assumes the existence of a true model and knowledge of the probability density function of the disturbing noise in the frequency domain. Therefore, in addition to Assumptions 7.1 to 7.6, 7.8 to 7.13, 7.16, and 7.17, we require:

Assumption 7.18 (Circular Complex Frequency Domain Errors): The frequency domain errors $N_Y(k)$, $N_U(k)$ are independent (over k), jointly correlated, zero mean, circular complex distributed random variables.

Assumption 7.19 (pdf Frequency Domain Errors): The observations Z_0 are deterministic and the frequency domain errors $N_Y(k)$, $N_U(k)$ are normally distributed random variables.

Assumption 7.20 (Efficiency Condition Frequency Domain Errors): The number of noncoherent noise sources equals 1. This is true if and only if one of the three following conditions is fulfilled for $k = 1, 2, \ldots, F$: (i) no input noise $\sigma_U^2(k) = 0$, (ii) no output noise $\sigma_Y^2(k) = 0$, or (iii) totally correlated input-output errors $\left|\sigma_{YU}^2(k)\right| / (\sigma_Y(k)\sigma_U(k)) = 1$.

For example, Assumption 7.20 is fulfilled in feedback when only process noise is present (no measurement errors and no controller noise, see Section 7.18 and Exercise 7.2).

7.7 ASYMPTOTIC PROPERTIES

In this section we give an overview and an elaborated discussion of the asymptotic properties of the minimizer $\hat{\theta}(Z)$ of cost functions $V_F(\theta, Z)$ which are quadratic-in-the-measurements Z. The overview starts with general estimators, proceeds with consistent estimators, and ends with the maximum likelihood estimator. Afterward, the results are generalized to cost functions of the form (7-14) that are nonquadratic in Z. In a first reading, one may skip Theorems 7.21 and 7.28 and go directly to the discussion of the properties.

Theorem 7.21 (Asymptotic Properties $\hat{\theta}(Z)$): Consider models (7-7) and (7-8) with any identifiable parameterization of Sections 5.2 and 5.3. Let $\hat{\theta}(Z)$ be the minimizer of a cost function $V_F(\theta, Z)$ of the form (7-11) that is quadratic-in-the-measurements Z. Under the assumptions of Section 7.6, the minimizer $\hat{\theta}(Z)$ has the following asymptotic ($F \rightarrow \infty$) properties,

1. *Stochastic convergence:* $\hat{\theta}(Z)$ converges strongly to $\tilde{\theta}(Z_0)$, the minimizer of $V_F(\theta) = \mathcal{E}\{ V_F(\theta, Z)\}$ (assumptions Section 7.6.1).
2. *Stochastic convergence rate:* $\hat{\theta}(Z)$ converges in probability at the rate $O_P(F^{-1/2})$ to $\tilde{\theta}(Z_0)$ (assumptions Section 7.6.2).
3. *Systematic and stochastic errors:* $\hat{\theta}(Z)$ converges in probability to $\tilde{\theta}(Z_0)$ with

$$\begin{aligned} \hat{\theta}(Z) &= \tilde{\theta}(Z_0) + \delta_\theta(Z) + b_\theta(Z) \\ \delta_\theta(Z) &= -V_F''^{-1}(\tilde{\theta}(Z_0)) V_F'^T(\tilde{\theta}(Z_0), Z) \end{aligned} \tag{7-25}$$

   where $\delta_\theta(Z) = O_p(F^{-1/2})$, with $\mathcal{E}\{ \delta_\theta(Z)\} = 0$, is the dominating stochastic error and where $b_\theta(Z) = O_P(F^{-1})$ contains the contribution of the systematic errors (assumptions Section 7.6.3).
4. *Asymptotic normality:* $\sqrt{F}(\hat{\theta}(Z) - \tilde{\theta}(Z_0))$ converges in law at the rate $O(F^{-1/2})$ to a Gaussian random variable with zero mean and covariance matrix $\mathrm{Cov}(\sqrt{F}\delta_\theta(Z))$

$$\begin{aligned} \mathrm{Cov}(\sqrt{F}\delta_\theta(Z)) &= V_F''^{-1}(\tilde{\theta}(Z_0)) Q_F(\tilde{\theta}(Z_0)) V_F''^{-1}(\tilde{\theta}(Z_0)) \\ Q_F(\tilde{\theta}(Z_0)) &= F\mathcal{E}\{ V_F'^T(\tilde{\theta}(Z_0), Z) V_F'(\tilde{\theta}(Z_0), Z)\} \end{aligned} \tag{7-26}$$

   (assumptions Section 7.6.4).
5. *Deterministic convergence:* $\tilde{\theta}(Z_0)$ converges to θ_*, the minimizer of

$$V_*(\theta) = \int_{f_{\min}}^{f_{\max}} \mathcal{E}\{ |\varepsilon(\Omega(f), \theta, Z(f))|^2\} n(f) df \tag{7-27}$$

   with $\Omega(f) = j2\pi f$, $e^{j2\pi f T_s}$, $\sqrt{j2\pi f}$, or $\tanh(\tau_R j2\pi f)$. The convergence rate is an $O(F^{-2})$ for frequency domain experiments and $O(F^{-1})$ for time domain experiments (assumptions Section 7.6.5).

   If in addition $V_F(\theta, Z)$ satisfies the consistency conditions then,
6. *Consistency:* $\hat{\theta}(Z)$ is strongly (weakly for model (7-8) with $\Omega = s$) consistent; replace in properties 1 to 4 $\tilde{\theta}(Z_0)$ ($\lim_{F \rightarrow \infty} \tilde{\theta}(Z_0) = \theta_*$ for model (7-8) with $\Omega = s$) by θ_0 (assumptions Section 7.6.6).

7. *Asymptotic bias:* The asymptotic bias $b_\theta = \mathcal{E}\{b_\theta(Z)\}$, and its derivative w.r.t. θ_0, $\partial b_\theta / \partial \theta_0$, of $\hat{\underline{\theta}}(Z)$ are an $O(F^{-1})$ ($O(F^{-1/2})$ for model (7-8) with $\Omega = s$ and random excitation) for all $\theta_0 \in \Theta_r$ (assumptions Section 7.6.7).
   If in addition $V_F(\theta, Z)$ is the maximum likelihood cost function then,
8. *Asymptotic efficiency:* The Gaussian maximum likelihood estimate $\hat{\theta}_{\mathrm{ML}}(Z)$ is asymptotically efficient: $\mathrm{Cov}(\delta_\theta(Z)) = Fi^{-1}(\theta_0)$ with $Fi(\theta_0) = FV_F''(\theta_0)$ the Fisher information matrix. Moreover, we have

$$\lim_{F \to \infty} (\mathrm{Cov}(\sqrt{F}\hat{\underline{\theta}}(Z)) - \mathrm{Cov}(\sqrt{F}\delta_\theta(Z))) = 0 \tag{7-28}$$

   (assumptions Section 7.6.8).

Proof. See Appendix 7.E. □

Corollary 7.22 (Asymptotic Properties $\hat{\theta}(Z)$—continued): Let $\hat{\theta}(Z)$ be the minimizer of a cost function $V_F(\theta, Z) = f_F(\theta, \eta(Z), Z)$ of the form (7-14) where $f_F(\theta, \eta, Z)$ is quadratic-in-the-measurements Z. Assume that the cost function $f_F(\theta, \eta, Z)$ and its third-order derivatives w.r.t. $x = [\theta^T \ \eta^T]^T$ are continuous and that $f_F(\theta, \eta, Z)$ fulfills the assumptions of Section 7.6. Define, furthermore, $g_F(\theta, \eta(Z), Z) = V_F'^T(\theta, Z)$ and $g_F(\theta, \eta) = \mathcal{E}\{g_F(\theta, \eta, Z)\}$. If Theorem 7.21 is valid for the (initial) estimate $\eta(Z)$, then the minimizer $\hat{\theta}(Z)$ has the asymptotic properties of Theorem 7.21 with the following three modifications:

1. To calculate $V_F(\theta)$ and $V_*(\theta)$ we first replace $\eta(Z)$ by its limit value η_* before taking the expected value, which gives

$$\begin{aligned} V_F(\theta) &= \frac{1}{F}\sum_{k=1}^{F} \mathcal{E}\{|\varepsilon(\Omega_k, \theta, Z(k), \eta_*)|^2\} \\ V_*(\theta) &= \int_{f_{\min}}^{f_{\max}} \mathcal{E}\{|\varepsilon(\Omega(f), \theta, Z(f), \eta_*)|^2\} n(f) df \end{aligned} \tag{7-29}$$

2. $\mathcal{E}\{\delta_\theta(Z)\}$ is not necessarily zero or may not even exist.
3. $\delta_\theta(Z)$ in the expression of the covariance matrix (7-26) is replaced by $d_\theta(Z)$

$$\begin{aligned} d_\theta(Z) &= -V_F''^{-1}(\tilde{\theta}(Z_0)) d_F(Z) \\ d_F(Z) &= g_F(\tilde{\theta}(Z_0), \eta_*, Z) + \frac{\partial g_F(\tilde{\theta}(z_0), \eta)}{\partial \eta_*} \delta_\eta(Z) \end{aligned} \tag{7-30}$$

   where $\delta_\eta(Z)$ is given by (7-25), and with $d_\theta(Z) = O_{\mathrm{p}}(F^{-1/2})$, $\mathcal{E}\{d_\theta(Z)\} = 0$.

Proof. See Appendix 7.F. □

We are ready now to answer the question we posed in Section 7.2: "What will happen with one's estimates (uncertainty, bias ...) if one gathered, for example, four times more data?" Property 1 ensures that $\hat{\theta}(Z)$ is likely to be closer to the minimizer $\tilde{\theta}(Z_0)$ of the expected value of the cost function. Property 2 tells us that $\hat{\theta}(Z)$ is likely to be two times closer to $\tilde{\theta}(Z_0)$. From property 3 it follows that the systematic and stochastic errors in the residual $\hat{\theta}(Z) - \tilde{\theta}(Z_0)$ are likely to decrease with a factor of 4 and 2 respectively. Finally, property 4 ensures that the distribution function of $\hat{\theta}(Z)$ is likely to be two times closer to a normal distribution. Similar results are obtained when no model errors are present $\tilde{\theta}(Z_0) = \theta_0$.

Expression (7-26) allows a theoretical calculation of the covariance matrix of the estimates in the presence of model errors. It requires, however, knowledge of the fourth-order moments of the noise and of the minimizer $\tilde{\theta}(Z_0)$ of the expected value of the cost function. Although $\tilde{\theta}(Z_0)$ can be approximated by the actual estimate $\hat{\theta}(Z)$, the fourth-order moments of the noise are mostly unknown. For the maximum likelihood estimator, the covariance expression (7-26) can be significantly simplified (only second-order moments of the noise are required) and a good approximation of the covariance matrix results as a by-product of the nonlinear minimization scheme (7-16) (see Section 7.11). Property 4 then makes it possible to calculate uncertainty regions around $\hat{\theta}(Z)$ that contain $\tilde{\theta}(Z_0)$ with some user-defined probability level. The same can be done for any model-related quantity (see Sections 14.2 and 17.4.7).

If model errors exist, then $\hat{\theta}(Z)$ converges to a value $\tilde{\theta}(Z_0) \neq \theta_0$ ($\theta_* \neq \theta_0$) that still depends on F. Property 5 guarantees that $\tilde{\theta}(Z_0)$ converges at the rate $O(F^{-1})$ or faster to its limit value θ_*, while according to property 2 the stochastic convergence rate of $\hat{\theta}(Z)$ to $\tilde{\theta}(Z_0)$ is an $O_p(F^{-1/2})$. Therefore, $\tilde{\theta}(Z_0)$ can be replaced everywhere by θ_* in properties 1 to 4. In case of model errors, we may also wonder whether $\hat{\theta}(Z)$ still converges to the same solution if the same experiment is repeated with a higher signal-to-noise ratio (lower noise levels). To verify this, we apply quick analysis tool number 3 (see Section 7.5) to the cost function. If so, then $\tilde{\theta}(Z_0)$ (θ_*) can be interpreted as the solution of the noiseless problem.

Property 6 guarantees that the estimate $\hat{\theta}(Z)$ converges to the true model parameters θ_0 for cost functions satisfying the consistency conditions 7.16 and 7.17. This does, however, not imply that the (equivalent) initial conditions in model (7-8) are consistently estimated. Indeed, the part of θ_0 corresponding to the (equivalent) initial conditions decreases to zero as $F^{-1/2}$ (use Lemma 5.5 taking into account that $F = O(N)$ for a time domain experiment), while the difference $\hat{\theta}(Z) - \theta_0$ is an $O_p(F^{-1/2})$. Hence, the relative difference $|(\hat{\theta}_{[i]}(Z) - \theta_{0[i]})/\theta_{0[i]}|$ between the estimated and the true initial conditions does not decrease to zero as the number of frequencies F increases to infinity, which shows that the initial conditions are not consistently estimated. This result can easily be understood in the time domain. The (equivalent) initial conditions (transient term $T(\Omega, \theta)$ in (7-8)) correspond to an exponentially decaying transient in the time domain. Observing the input and output signals during a longer period does not give more information about the transient, hence, it cannot be estimated consistently. For the same reason, properties 7 and 8 do not imply that the estimated equivalent initial conditions are asymptotically efficient and have an $O(F^{-1})$ bias. Although they cannot be estimated consistently, we still include the initial conditions in model (7-8) because it turns out that they improve the finite sample behavior (F is not "large") of the estimated plant model $G(\Omega, \hat{\theta}(Z))$. Note also that the influence of the transient term $T(\Omega, \theta)$ to the cost function $V_F(\theta, Z)$ is an $O_p(F^{-1})$ (see Appendix 7.D).

The asymptotic efficiency of the maximum likelihood estimator (property 8 of Theorem 7.21) has been shown under some restrictive noise assumptions (see Assumption 7.20); for example, the input must be known exactly. In general, the maximum likelihood solution is not asymptotically efficient. This is not in contradiction with the general properties of maximum likelihood estimators (Section 1.5.3) because the number of estimated parameters in the errors-in-variables problem increases with F (see Section 7.11).

For deterministic vectors $\eta(Z) = \eta_*$ the term $d_N(Z)$ in (7-30) reduces to $g_F(\tilde{\theta}(z_0), \eta_*, Z)$. Therefore, modification number 3 of Corollary 7.22 shows that in general the stochastic vector will increase the uncertainty of the estimates. If, however, $\partial g_F(\tilde{\theta}(z_0), \eta)/\partial \eta_* = o(F^0)$, then there is no asymptotic increase in uncertainty (see, for example, Section 7.12.3).

7.8 LINEAR LEAST SQUARES

7.8.1 Introduction

A reasonable measure of goodness of fit is to compare the observed output $Y(k)$ with the modeled output $Y(k, \theta)$ (7-7) or (7-8), where $G(\Omega, \theta)$ (and $T(\Omega, \theta)$) can take any parameterization of Section 5.2 (and Section 5.3). The plant model parameters are then obtained by minimizing the sum of the squared residuals

$$V_{\mathrm{NLS}}(\theta, Z) = \sum_{k=1}^{F} |Y(k) - Y(k, \theta)|^2 \tag{7-31}$$

w.r.t. to θ. Because $Y(k, \theta)$ is a nonlinear function of θ, the cost function (7-31) is a nonquadratic function of θ. All the estimation methods presented in this section try to minimize (7-31) by (successive) linear least squares approximation(s). The key idea is to make (7-31) quadratic in θ by parameterizing $G(\Omega, \theta)$ (and $T(\Omega, \theta)$) as a rational form $B(\Omega, \theta)/A(\Omega, \theta)$ (and $I(\Omega, \theta)/A(\Omega, \theta)$) and by multiplying each residual $Y(k) - Y(k, \theta)$ in the cost function (7-31) by $A(\Omega_k, \theta)$.

7.8.2 Linear Least Squares

Multiplying each residual $Y(k) - Y(k, \theta)$ in the cost function (7-31) by $A(\Omega_k, \theta)$ gives the linear least squares (LS) cost function

$$V_{\mathrm{LS}}(\theta, Z) = \sum_{k=1}^{F} |e(\Omega_k, \theta, Z(k))|^2 \tag{7-32}$$

with $e(\Omega_k, \theta, Z(k))$ the equation error (7-9) or (7-10). The linear least squares (LS) estimate $\hat{\theta}_{\mathrm{LS}}(Z)$ is found by minimizing (7-32) w.r.t. θ using the constraint $a_i = 1$ or $b_i = 1$. In Levi (1959) the linear least squares approach was applied for the first time to identify continuous-time models starting from transfer function measurements ((7-32) with equation error (7-9), $\Omega = s$, $Y(k) = G(s_k)$, and $U(k) = 1$). The linearization of the output error $Y(k) - Y(\Omega_k, \theta)$ has two major drawbacks when identifying continuous-time models ($\Omega = s$, $\sqrt{s}$, and $\tanh(\tau_{\mathrm{R}} s)$): the overemphasizing of high-frequency errors in (7-32) and the large dynamic range of the numbers in the normal equation (7-16). Indeed, $e(\Omega_k, \theta, Z(k))$ is a polynomial in Ω_k and, hence, the contribution of the disturbing noise at frequency Ω_k to the cost function increases with $|\Omega_k|^{2\max(n_a, n_b)}$. This may result in poor low-frequency fits (see Figure 7-4) and ill-conditioned normal equations for identification problems with a large dynamic frequency range. Similar problems occur for discrete-time models ($\Omega = z^{-1}$) when identified on a "small" part of the unit circle.

Because $V_{\mathrm{LS}}(\theta, Z)$ is quadratic-in-the-measurements Z, the asymptotic properties proved in Theorem 7.21, with $V_F(\theta, Z) = V_{\mathrm{LS}}(\theta, Z)/F$, are valid for $\hat{\theta}_{\mathrm{LS}}(Z)$. To reveal the major properties of $\hat{\theta}_{\mathrm{LS}}(Z)$ we use the quick analysis tools of Section 7.5. Taking the expected value of (7-32) gives (7-19) with

$$\mathcal{E}\{v_F(\theta, N_Z)\} = \frac{1}{F}\sum_{k=1}^{F} \sigma_e^2(\Omega_k, \theta) \tag{7-33}$$

(see Exercise 7.3). $\sigma_e^2(\Omega_k, \theta) = \mathrm{var}(e(\Omega_k, \theta, N_Z(k)))$ is the variance of the equation error where the measurements Z have been replaced by the noise on the measurements N_Z

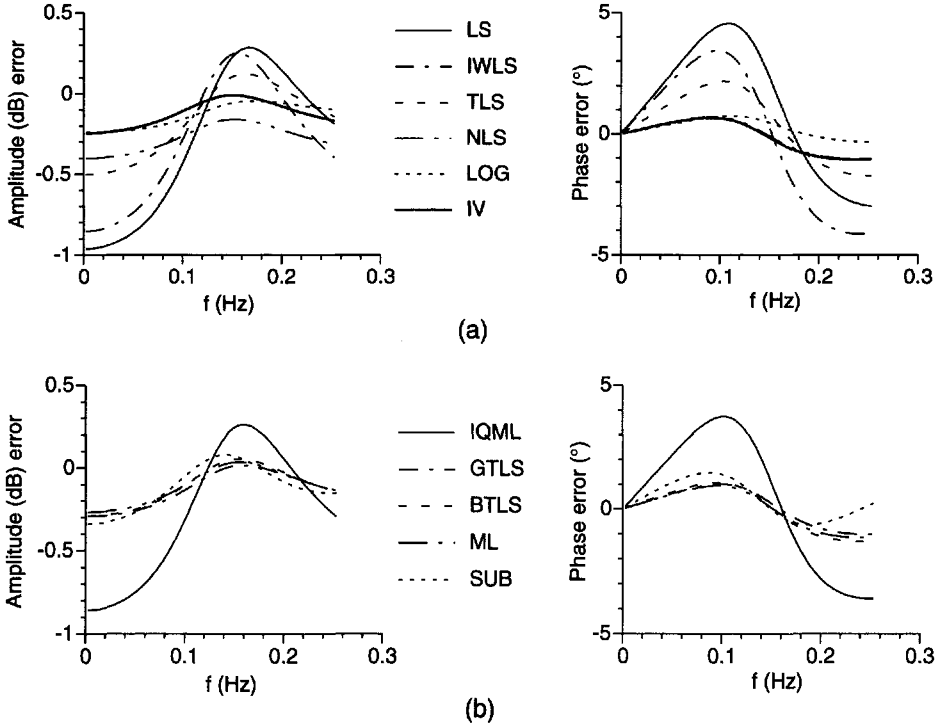


Figure 7-4. Second-order simulation example $G(s, \theta) = 1/(1 + s + s^2)$ defined in Appendix 7.A (see also Figure 7-1 on page 184). Left: difference between the estimated amplitude in dB and the true amplitude in dB, and right: phase error in degrees. (a) Estimators requiring no noise information, (b) estimators requiring the noise covariance.

$$\sigma_e^2(\Omega_k, \theta) = \sigma_Y^2(k)|A(\Omega_k, \theta)|^2 + \sigma_U^2(k)|B(\Omega_k, \theta)|^2 - 2\mathrm{Re}(\sigma_{YU}^2(k)A(\Omega_k, \theta)\bar{B}(\Omega_k, \theta)) \quad (7\text{-}34)$$

Applying quick tool 2 (see Section 7.5) to (7-33) shows that the linear least squares estimate $\hat{\theta}_{LS}(Z)$ is, in general, inconsistent because (7-34) is, in general, θ dependent. It is consistent if $\sigma_e^2(\Omega_k, \theta)$ is independent of θ, for example, no input noise ($\sigma_U^2(k) = 0$, $\sigma_{YU}^2(k) = 0$) and a polynomial plant model ($A(\Omega, \theta) = 1$). Replacing C_{N_Z} by $\upsilon^2 C_{N_Z}$ in the expected value of the cost function gives, taking into account (7-33),

$$V_F(\theta) = \mathcal{E}\{V_F(\theta, Z_0)\} + \upsilon^2\mathcal{E}\{v_F(\theta, N_Z)\} \quad (7\text{-}35)$$

It shows that in general $\tilde{\theta}_{LS}(Z_0)$ and θ_{*LS} depend on the disturbing noise level υ and, hence, cannot be considered as the noiseless solutions (see Section 7.5, quick tool 3). They are the noiseless solutions if $\sigma_e^2(\Omega_k, \theta)$ is independent of θ. From (7-32) it follows directly that $V_{LS}(\lambda\theta, Z) = \lambda^2 V_{LS}(\theta, Z)$ so that $\hat{\theta}_{LS}(Z)$ depends on the particular constraint chosen, for example, $a_i = 1$ or $b_i = 1$ (see Section 7.5, quick tool 4). This is illustrated in Figure 7-5. Note that on the average the estimate with $a_0 = 1$ is too small (underbiased), while the estimate with $b_0 = 1$ is too large (overbiased). This is in agreement with the results of De Moor et al. (1994). See Table 7-5 on page 238 for an overview of the properties of the LS estimator.

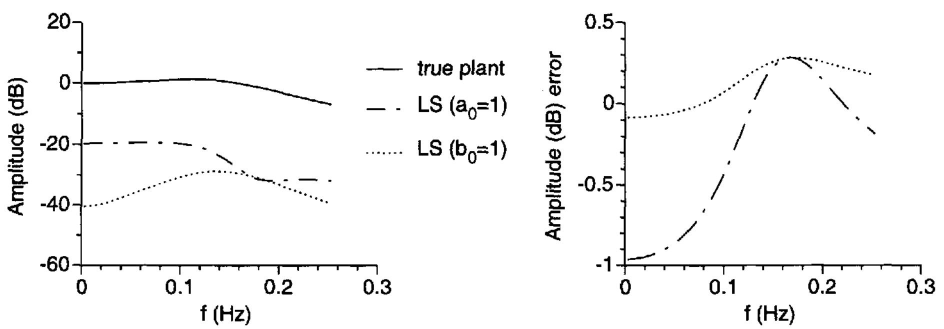


Figure 7-5. Second-order simulation example $G(s, \theta) = 1/(1+s+s^2)$ defined in Appendix 7.A (see also Figure 7-1 on page 184). Comparison of the linear least squares estimates using the constraint $a_0 = 1$ and the linear least squares estimates using the constraint $b_0 = 1$. Left figure, true plant model (solid line) and magnitude of the complex error between the estimated and the true plant model. Right figure, difference between the estimated amplitude in dB and the true amplitude in dB.

7.8.3 Iterative Weighted Linear Least Squares

To overcome the lack of sensitivity to low-frequency errors of the linear least squares estimator, the equation error $e(\Omega_k, \theta, Z(k))$ in (7-32) is divided by an initial guess of the denominator polynomial $A(\Omega_k, \theta^{(0)})$. The obtained weighted linear least squares estimate $\theta^{(1)}$ can be used to calculate a (hopefully) better estimate of the denominator polynomial $A(\Omega_k, \theta^{(1)})$ resulting in a (hopefully) better estimate $\theta^{(2)}$, and so on .... The ith step of the iterative procedure consists of minimizing

$$V^{(i)}_{\mathrm{IWLS}}(\theta^{(i)}, Z) = \sum_{k=1}^{F} \frac{|e(\Omega_k, \theta^{(i)}, Z(k))|^2}{|A(\Omega_k, \theta^{(i-1)})|^2} \tag{7-36}$$

with $e(\Omega_k, \theta, Z(k))$ the equation error (7-9) or (7-10), w.r.t. $\theta^{(i)}$ using the constraint $a_j = 1$ or $b_j = 1$. In most cases the linear least squares estimate is used as starting value $\theta^{(0)} = \hat{\theta}_{\mathrm{LS}}(Z)$, and when convergent the iterative weighted linear least squares (IWLS) estimate is $\hat{\theta}_{\mathrm{IWLS}}(Z) = \theta^{(\infty)}$. In Sanathanan and Koerner (1963) this iterative procedure was applied for the first time to identify continuous-time models starting from transfer function measurements ((7-36) with equation error (7-9), $\Omega = s$, $Y(k) = G(s_k)$ and $U(k) = 1$). From Figure 7-4 on page 200 it can be seen that the low-frequency errors of the IWLS fit are indeed smaller than those of the LS fit. When convergent ($\theta^{(i)} = \theta^{(i-1)}$ for i sufficiently large) the IWLS cost (7-36) tends to the nonlinear least squares cost (7-31). Although this property is very appealing, it does not guarantee that the global minima of both cost functions are the same. Therefore, one needs that the derivatives of these cost functions w.r.t. θ are asymptotically ($i \to \infty$) the same. In general, this is not true and, hence, $\hat{\theta}_{\mathrm{IWLS}}(Z) \neq \hat{\theta}_{\mathrm{NLS}}(Z)$. However, as the elementwise difference between the Jacobians is proportional to the equation error $e(\Omega_k, \theta^{(i-1)}, Z(k))$ (see Exercise 7.4), both estimates will coincide ($\hat{\theta}_{\mathrm{IWLS}}(Z) \approx \hat{\theta}_{\mathrm{NLS}}(Z)$) for "sufficiently high" signal-to-noise ratios and "sufficiently small" modeling errors, otherwise the difference may be large. This is illustrated by the "high noise" simulation example of Figure 7-4 (compare IWLS with NLS), and the "low noise" simulation example of Figure 7-8 on page 230 (compare IWLS to NLS).

Analysis of the statistical properties of the estimate $\theta^{(\infty)}$ is in general impossible. It is, however, feasible to analyze the properties of the first step of the iterative procedure (7-36). If the initial guess $\theta^{(0)}$ is deterministic and independent of the number of frequencies F, then Theorem 7.21 is valid and $\hat{\theta}_{\text{IWLS}}(Z) = \theta^{(1)}$ has asymptotic $(F \to \infty)$ properties similar to those of $\hat{\theta}_{\text{LS}}(Z)$ (see Section 7.8.2). If the choice $\theta^{(0)} = \hat{\theta}_{\text{LS}}(Z)$ is made, then the cost function (7-36) is no longer a quadratic function of the measurements Z. Indeed, $\hat{\theta}_{\text{LS}}(Z)$ depends on Z and appears in the denominator of (7-36). Although this complicates the analysis, it turns out that Theorem 7.21 is still valid for $\hat{\theta}_{\text{IWLS}}(Z) = \theta^{(1)}$ with three minor modifications (see Corollary 7.22). Hence, $\hat{\theta}_{\text{IWLS}}(Z) = \theta^{(1)}$ has the same asymptotic $(F \to \infty)$ properties as $\hat{\theta}_{\text{LS}}(Z)$. We conclude that in general the estimate $\hat{\theta}_{\text{IWLS}}(Z)$ is inconsistent, depends on the particular constraint chosen, and does not converge to a noiseless solution.

Many modifications of and extensions to the original method of Sanathanan and Koerner (1963) have been published. Almost all of them fit within the following (iterative) weighted least squares framework:

$$\sum_{k=1}^{F} W^2(\Omega_k, \theta^{(i-1)})|e(\Omega_k, \theta^{(i)}, Z(k))|^2 \tag{7-37}$$

where $W(\Omega_k, \theta^{(i-1)})$ is a well-chosen real weighting function (see Pintelon et al., 1994 for an overview). One particular weighting is interesting, namely

$$W(\Omega_k, \theta^{(i-1)}) = \frac{1}{|A(\Omega_k, \theta^{(i-1)})|^r} \text{ with } r \in [0, \infty) \tag{7-38}$$

Two special cases of (7-38) are the linear least squares method for $r = 0$ and the iterative weighted linear least squares method (7-36) for $r = 1$. Powers r, different from one, may result in smaller output errors $Y(k) - Y(k, \theta)$; for example, if the iterative scheme (7-36) does not converge, then relaxation $(r < 1)$ is helpful. In 't Mannetje (1973), the relaxation idea was applied for the first time to identify continuous-time models starting from transfer function measurements ((7-36) with equation error (7-9), $\Omega = s$, $Y(k) = G(s_k)$, and $U(k) = 1$). The asymptotic $(F \to \infty)$ properties of the minimizer of (7-37) are similar to those of $\hat{\theta}_{\text{IWLS}}(Z)$ (7-36), so that in general the minimizers of (7-37) and (7-31) are different. See Table 7-5 on page 238 for an overview of the properties of the IWLS estimator.

7.8.4 A Simple Example

Consider the identification of an integrator $G(s, \theta) = b_0/(a_1 s)$, starting from frequency response data $G(s_k) = G_0(s_k) + N_G(k)$, perturbed with independent (over the frequency), zero mean, circular complex noise $N_G(k)$ with variance $\text{var}(N_G(k)) = \sigma^2$ and finite fourth-order moments. The iterative weighted linear least squares estimate (7-37) is calculated using the weight (7-38) and the constraint $b_0 = 1$

$$(\hat{a}_1)_{\text{IWLS}} = \frac{\sum_{k=1}^{F} |s_k|^{-2r}\text{Re}(s_k G(s_k))}{\sum_{k=1}^{F} |s_k|^{2(1-r)}|G(s_k)|^2} \tag{7-39}$$

Applying the strong law of large numbers (see Section 14.9, version 2) to the numerator and denominator of (7-39) and the interchangeability property of the almost sure limit and a continuous function (see Section 14.8, property 1), we find

$$\underset{F \to \infty}{\text{a.s.lim}}(\hat{a}_1)_{\text{IWLS}} = (a_1)_0 \frac{1}{1 + \lim\limits_{F \to \infty} \frac{\sum_{k=1}^{F} |s_k|^{-2r} \sigma^2 / |G_0(s_k)|^2}{\sum_{k=1}^{F} |s_k|^{-2r}}} \tag{7-40}$$

with $(a_1)_0$ the true value. As predicted by the theory (apply quick tool number 2), it clearly follows from (7-40) that $(\hat{a}_1)_{\text{IWLS}}$ and, hence, also $G(s_k, \hat{\theta}_{\text{IWLS}}(Z))$ are inconsistent estimates. Taking, for example, $F = 100$ angular frequencies equally spaced between 0.1 and 2 and $\sigma^2 = 0.5$; the right-hand side of (7-40) is then equal to $0.587(a_1)_0$ and $0.916(a_1)_0$ for, respectively, $r = 0$ (LS solution (7-32)) and $r = 1$ (IWLS solution (7-36)). It shows that weighting the linear least squares residual with an initial guess of the denominator polynomial indeed improves the estimates. In this numerical example, values of $r > 1$ give even better results compared with $r = 1$.

Making the same calculations for the IWLS estimate with constraint $a_1 = 1$, we get

$$\begin{aligned} (\hat{b}_0)_{\text{IWLS}} &= \frac{\sum_{k=1}^{F} |s_k|^{-2r} \text{Re}(s_k G(s_k))}{\sum_{k=1}^{F} |s_k|^{-2r}} \\ \underset{F \to \infty}{\text{a.s.lim}}(\hat{b}_0)_{\text{IWLS}} &= \frac{\lim\limits_{F \to \infty} \frac{1}{F} \sum_{k=1}^{F} |s_k|^{-2r} \text{Re}(s_k G_0(s_k))}{\lim\limits_{F \to \infty} \frac{1}{F} \sum_{k=1}^{F} |s_k|^{-2r}} = (b_0)_0 \end{aligned} \tag{7-41}$$

with $(b_0)_0$ the true value. As predicted by the theory (apply quick tool number 2) $(\hat{b}_0)_{\text{IWLS}}$ and, hence, also $G(s_k, \hat{\theta}_{\text{IWLS}}(Z))$ are consistent estimates. It illustrates nicely the dependence of $G(s_k, \hat{\theta}_{\text{IWLS}}(Z))$ on the parameter constraint used (quick tool number 4).

Putting $r = 0$ in (7-39) to (7-41) shows that the same conclusions hold for the least squares estimate $(\hat{a}_0)_{\text{LS}}$ and $(\hat{b}_0)_{\text{LS}}$.

7.9 NONLINEAR LEAST SQUARES

7.9.1 Output Error

The nonlinear least squares (NLS) estimator $\hat{\theta}_{\text{NLS}}(Z)$ minimizes the sum of the squared residuals between the observed output $Y(k)$ and the modeled output $Y(k, \theta)$ (7-7) or (7-8), where $G(\Omega, \theta)$ (and $T(\Omega, \theta)$) can take any parameterization of Section 5.2 (and Section 5.3)

$$V_{\text{NLS}}(\theta, Z) = \sum_{k=1}^{F} |Y(k) - Y(k, \theta)|^2 \tag{7-42}$$

The Newton-Gauss minimization scheme (7-18) is used to calculate $\hat{\theta}_{\text{NLS}}(Z)$, and as with most nonlinear minimization problems, the method may converge to a local minimum of (7-42) ($\theta^{(\infty)} \neq \hat{\theta}_{\text{NLS}}(Z)$). Therefore, it is important to have starting values of "sufficiently high" quality. The (iterative) weighted linear least squares solution (7-36) can be used for this purpose. In Van den Enden et al. (1977) and Van den Enden and Leenknegt (1986) this scheme was used for the first time to identify respectively continuous-time and discrete-time models starting from transfer function measurements ((7-42) with output model (7-7), $\Omega = s$ or z^{-1}, $Y(k) = G(s_k)$, and $U(k) = 1$).

Because $V_{\text{NLS}}(\theta, Z)$ is quadratic-in-the-measurements Z, the asymptotic properties proved in Theorem 7.21, with $V_F(\theta, Z) = V_{\text{NLS}}(\theta, Z)/F$, are valid for $\hat{\theta}_{\text{NLS}}(Z)$. We use the quick analysis tools of Section 7.5 to reveal the major properties of $\hat{\theta}_{\text{NLS}}(Z)$. Taking the expected value of (7-42) gives (7-19) with

$$\mathcal{E}\{\, v_F(\theta, N_Z)\} = \frac{1}{F}\sum\nolimits_{k=1}^{F} \sigma_Y^2(\Omega_k, \theta) \tag{7-43}$$

(see Exercise 7.5). $\sigma_Y^2(\Omega_k, \theta)$ is the variance of the output error where the measurements Z have been replaced by the noise on the measurements N_Z

$$\sigma_Y^2(\Omega_k, \theta) = \sigma_Y^2(k) + \sigma_U^2(k)|G(\Omega_k, \theta)|^2 - 2\text{Re}(\sigma_{YU}^2(k)\bar{G}(\Omega_k, \theta)) \tag{7-44}$$

Applying quick tool 2 (see Section 7.5) to (7-43) shows that in general the nonlinear least squares estimate $\hat{\theta}_{\text{NLS}}(Z)$ is inconsistent. It is consistent if $\sigma_Y^2(\Omega_k, \theta)$ is independent of θ, which is the case for transfer function measurements (7-2) ($Y(k) = G(\Omega_k)$, $U(k) = 1$, $\sigma_U^2(k) = 0$, and $\sigma_{YU}^2(k) = 0$) or input-output measurements (7-1) with exactly known input ($\sigma_U^2(k) = 0$, $\sigma_{YU}^2(k) = 0$). Replacing C_{N_Z} by $\upsilon^2 C_{N_Z}$ in the expected value of the cost function gives, taking into account (7-43),

$$V_F(\theta) = \mathcal{E}\{\, V_F(\theta, Z_0)\} + \upsilon^2 \mathcal{E}\{\, v_F(\theta, N_Z)\} \tag{7-45}$$

It shows that in general $\tilde{\theta}_{\text{NLS}}(Z_0)$ and $\theta_{*\text{NLS}}$ depend on the disturbing noise level υ and, hence, cannot be considered as the noiseless solutions (see Section 7.5, quick tool 3). They are the noiseless solutions for transfer function measurements and input-output measurements with exactly known input, because $\sigma_Y^2(\Omega_k, \theta)$ and, hence, also $\mathcal{E}\{\, v_F(\theta, N_Z)\}$ are then independent of θ. From (7-42) it follows immediately that $V_{\text{NLS}}(\lambda\theta, Z) = V_{\text{NLS}}(\theta, Z)$ so that $\hat{\theta}_{\text{NLS}}(Z)$ is independent of the particular parameter constraint $a_i = 1$, $b_i = 1$ or $\|\theta\|_2^2 = 1$ chosen (see Section 7.5, quick tool 4).

We conclude from the previous discussion that the NLS estimator is inconsistent for noisy input-output measurements, while it is consistent for transfer function measurements. This suggests that for transfer function model (7-7) the bias in the estimates could be removed if the input-output measurements (7-1) are transformed into a transfer function measurement (7-2) with $G(k) = Y(k)/U(k)$. The nonlinear least squares estimate then minimizes

$$V_{\text{NLS}}(\theta) = \sum\nolimits_{k=1}^{F} |Y(k)/U(k) - G(\Omega_k, \theta)|^2 \tag{7-46}$$

w.r.t. θ. From a theoretical point of view the minimizer of (7-46) is inconsistent because the mean value of the noise on $Y(k)/U(k)$ is not zero,

$$\begin{aligned} Y(k)/U(k) &= G_0(\Omega_k) + N_G(k) \\ N_G(k) &= G_0(k)\left(\frac{1 + N_Y(k)/Y_0(k)}{1 + N_U(k)/U_0(k)} - 1\right) \end{aligned} \tag{7-47}$$

with $\mathcal{E}\{N_G(k)\} \neq 0$. Moreover, the moments of order 2 and higher of $N_G(k)$ do not exist (Guillaume et al., 1996a). We first study the bias term as a function of the signal-to-noise ratios and next tackle the nonexistence of the higher order moments.

For zero mean, circular complex distributed errors $N_Y(k)$, $N_U(k)$ (Assumption 7.18) with even probability density function the bias $\mathcal{E}\{N_G(k)\}$ is a function of the fourth-order moments of the noise (see Appendix 7.G). Assume now that the input-output errors are linearly correlated,

$$\begin{aligned} N_Y(k) &= N(k) + \rho(k)\frac{\sigma_U(k)\sigma_Y(k)}{\sigma_V^2(k)}N_V(k) \\ N_U(k) &= M(k) + N_V(k) \end{aligned} \tag{7-48}$$

where $N(k)$, $M(k)$, and $N_V(k)$ are mutually independent random variables, and with $\rho(k) = \sigma_{YU}^2(k)/(\sigma_U(k)\sigma_Y(k))$ the correlation coefficient. Note that a correlation of the form (7-48) occurs, for example, in linear feedback systems (see Section 7.18). If $N_Y(k)$, $N_U(k)$ are, in addition, circular complex normally distributed (Assumptions 7.18 and 7.19), then an analytic expression can be found for the relative bias $b(k) = \mathcal{E}\{N_G(k)\}/G_0(\Omega_k)$ (see Appendix 7.G)

$$b(k) = -\exp(-|U_0(k)|^2/\sigma_U^2(k))\left(1 - \rho(k)\frac{U_0(k)/\sigma_U(k)}{Y_0(k)/\sigma_Y(k)}\right) \text{ for } k \neq 0, N/2 \tag{7-49}$$

For uncorrelated input-output errors, $\rho(k) = 0$, (7-49) reduces to a real number

$$b(k) = -\exp(-|U_0(k)|^2/\sigma_U^2(k)) \tag{7-50}$$

and, hence, the bias does not affect the phase. From (7-49) it follows that the relative bias $|b(k)|$ is maximal for totally correlated input-output errors, $|\rho(k)| = 1$ and $\angle\rho(k) = \pi + \angle G_0(\Omega_k)$,

$$\max_{\rho(k)}|b(k)| = \exp(-|U_0(k)|^2/\sigma_U^2(k))\left(1 + \frac{|U_0(k)|/\sigma_U(k)}{|Y_0(k)|/\sigma_Y(k)}\right) \tag{7-51}$$

The relative bias $|b(k)|$ (7-50) is smaller than 5×10^{-5} for signal-to-noise ratios $|U_0(k)|/\sigma_U(k)$ larger than 10 dB, and the maximal relative bias (7-51) is smaller than 1×10^{-4} if the worst case input and output signal-to-noise ratios $|U_0(k)|/\sigma_U(k)$, $|Y_0(k)|/\sigma_Y(k)$ are larger than 10 dB.

To ensure the existence of the higher order moments of $N_G(k)$ we exclude large, highly improbable values of $G(\Omega_k) = Y(k)/U(k)$. Define the truncated ratio $\underline{G}(\Omega_k)$ as

$$\underline{G}(\Omega_k) = \begin{cases} Y(k)/U(k) & |U(k)/U_0(k)| \geq L \\ 0 & |U(k)/U_0(k)| < L \end{cases} \tag{7-52}$$

with L an arbitrarily small number. Note that this is exactly what we do in practice: if the ratio $Y(k)/U(k)$ is unacceptably large, then we reject it. For input signal-to-noise ratios larger than 10 dB and $L = 1\times10^{-3}$ the change in bias of $\underline{G}(\Omega_k)$ w.r.t. $G(\Omega_k)$ is negligible and the variance of the truncated estimate is in good approximation given by the variance obtained via linearization (see (2-25) and Guillaume et al., 1996a)

$$\sigma_G^2(k) = |G_0(\Omega_k)|^2[\sigma_Y^2(k)/|Y_0(k)|^2 + \sigma_U^2(k)/|U_0(k)|^2 - 2\mathrm{Re}(\sigma_{YU}^2(k)/(Y_0(k)\overline{U}_0(k)))] \quad (7\text{-}53)$$

Hence, from a practical point of view, we may say that $N_G(k)$ has zero mean with existing higher order moments and that Assumption 7.4 is valid for $N_G(k)$ if $N_U(k)$, $N_Y(k)$ satisfy Assumptions 7.18 and 7.19. Because the cost function (7-46) is quadratic-in-the-measurements $Y(k)/U(k)$, we conclude that Theorem 7.21 is "practically valid" for the estimate $\hat{\theta}_{\mathrm{NLS}}(Z)$ if the worst case input and output signal-to-noise ratio is at least 10 dB. Figure 7-6 shows that the errors of the NLS-I/O estimate (7-42) based on the input-output spectra are larger than those of the NLS-FRF estimate based on the frequency response function (7-46). As predicted by the theory, the NLS-I/O estimate is biased while the NLS-FRF estimate is "practically" consistent (compare NLS-FRF of Figure 7-6 to ML of Figure 7-4 on page 200). See Table 7-5 on page 238 for an overview of the properties of the NLS-FRF and NLS-IO estimators.

7.9.2 Logarithmic Least Squares

For frequency response functions with a large dynamic range, the nonlinear least squares estimator (7-42) of rational transfer function model (7-7) parameterized in powers of Ω_k (see (5-20)) may become ill conditioned. The dynamic range of the frequency response function can be limited by taking the natural logarithm of the model equation $Y(k) = G(\Omega_k, \theta)U(k)$ giving $\ln(Y(k)/U(k)) = \ln(G(\Omega_k, \theta))$. The logarithmic least squares (LOG) estimator then minimizes

$$V_{\mathrm{LOG}}(\theta, Z) = \sum_{k=1}^{F} |\ln(Y(k)/U(k)) - \ln(G(\Omega_k, \theta))|^2 \quad (7\text{-}54)$$

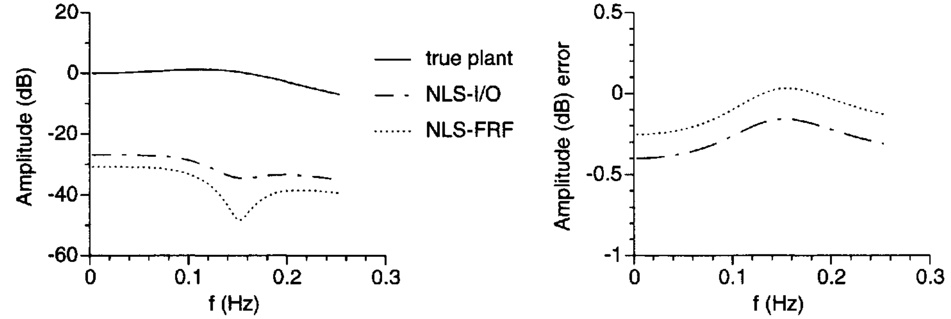


Figure 7-6. Second-order simulation example $G(s, \theta) = 1/(1 + s + s^2)$ defined in Appendix 7.A (see also Figure 7-1 on page 184). Comparison of the nonlinear least squares estimates using the input-output spectra $Y(k)$, $U(k)$ (NLS-I/O) and the nonlinear least squares estimates using the frequency response function $G(k) = Y(k)/U(k)$ (NLS-FRF). Left, true plant model (solid line) and magnitude of the complex error between the estimated and the true plant model. Right, difference between the estimated amplitude in dB and the true amplitude in dB.

w.r.t. θ. Besides its improved numerical stability (Sidman et al., 1991), the logarithmic least squares estimate $\hat{\theta}_{\mathrm{LOG}}(Z)$ is particularly robust with respect to outliers in the measurements (Guillaume et al., 1995). Good starting values for the LOG estimator are the LS (7-32) and the IWLS (7-36) estimates.

From a theoretical point of view the logarithmic least squares estimator is inconsistent because the noise on $\ln(Y(k)/U(k))$ has no zero mean,

$$\begin{aligned} \ln(Y(k)/U(k)) &= \ln(G_0(\Omega_k)) + N(k) \\ N(k) &= \ln(1 + N_Y(k)/Y_0(k)) - \ln(1 + N_U(k)/U_0(k)) \end{aligned} \tag{7-55}$$

with $\mathcal{E}\{N(k)\} \neq 0$. The higher order moments of $N(k)$, however, do exist (Guillaume et al., 1996a). For zero mean, circular complex distributed errors $N_Y(k)$, $N_U(k)$ (Assumption 7.18) with even probability density function, the bias $b(k) = \mathcal{E}\{N(k)\}$ is a function of the fourth-order moments of the noise (see Appendix 7.G). If the errors are, in addition, normally distributed (Assumption 7.19), then an analytic expression can be found for $b(k)$ (see Appendix 7.G)

$$b(k) = \frac{1}{2}\mathrm{Ei}\left(-\frac{|U_0(k)|^2}{\sigma_U^2(k)}\right) - \frac{1}{2}\mathrm{Ei}\left(-\frac{|Y_0(k)|^2}{\sigma_Y^2(k)}\right) \text{ for } k \neq 0, N/2 \tag{7-56}$$

with Ei(.) the exponential integral function (Gradshteyn and Ryzhik, 1980). Note that this expression is also valid for correlated input-output errors. It follows that the maximum bias error $|b(k)|$ is smaller than 2×10^{-6} for signal-to-noise-ratios $|Y_0(k)|/\sigma_Y(k)$ and $|U_0(k)|/\sigma_U(k)$ larger than 10 dB (see also Figure 2-18 on page 51). Hence, from a practical point of view, we may say that $N(k)$ has zero mean and that Assumption 7.4 is valid for $N(k)$ if $N_Y(k)$ and $N_U(k)$ satisfy Assumptions 7.18 and 7.19. Because the cost function (7-54) is quadratic-in-the-measurements $\ln(Y(k)/U(k))$, we conclude that Theorem 7.21, with $V_F(\theta, Z) = V_{\mathrm{LOG}}(\theta, Z)/F$, is "practically valid" for the logarithmic least squares estimate $\hat{\theta}_{\mathrm{LOG}}(Z)$ if the worst case signal-to-noise ratio is at least 10 dB. The expected value of (7-54) then equals (7-19) with

$$\mathcal{E}\{ v_F(\theta, N_Z)\} = \frac{1}{F}\sum_{k=1}^{F} \mathcal{E}\{ |\ln(1 + N_Y(k)/Y_0(k)) - \ln(1 + N_U(k)/U_0(k))|^2\} \tag{7-57}$$

Because $\mathcal{E}\{ v_F(\theta, N_Z)\}$ is independent of θ, $\hat{\theta}_{\mathrm{LOG}}(Z)$ is "practically consistent" and $\tilde{\theta}_{\mathrm{LOG}}(Z_0)$, $\theta_{*\mathrm{LOG}}$ are "practically" the noiseless solutions in case model errors are present (apply quick tools number 2 and 3 of Section 7.5). From (7-54) it follows that $V_{\mathrm{LOG}}(\lambda\theta, Z) = V_{\mathrm{LOG}}(\theta, Z)$, and, hence, $\hat{\theta}_{\mathrm{LOG}}(Z)$ is independent of the particular, chosen parameter constraint $a_i = 1$, $b_i = 1$, or $\|\theta\|_2^2 = 1$ (see Section 7.5, quick tool 4). See Table 7-5 on page 238 for an overview of the properties of the LOG estimator.

7.9.3 A Simple Example—Continued

We use the example of Section 7.8.4 to calculate the nonlinear least squares estimate (7-42) of the integrator model $G(s, \theta) = b_0/(a_1 s)$, using the constraint $b_0 = 1$. Making similar calculations as in Section 7.8.4, we get

$$(\hat{a}_1)_{\mathrm{NLS}} = \frac{\sum_{k=1}^{F} |s_k|^{-2}}{\sum_{k=1}^{F} \mathrm{Re}(G(s_k)/\bar{s}_k)}$$

$$\underset{F \to \infty}{\mathrm{a.s.lim}}(\hat{a}_1)_{\mathrm{NLS}} = \frac{\lim_{F \to \infty} \frac{1}{F} \sum_{k=1}^{F} |s_k|^{-2}}{\lim_{F \to \infty} \frac{1}{F} \sum_{k=1}^{F} \mathrm{Re}(G_0(s_k)/\bar{s}_k)} = (a_1)_0$$

which shows that the nonlinear least squares estimator $(\hat{a}_1)_{\mathrm{NLS}}$ and, hence, $G(s_k, \hat{\theta}_{\mathrm{NLS}}(Z))$ are, indeed, consistent for transfer function measurements. It is easy to verify that the NLS estimate using the constraint $a_0 = 1$ equals $(\hat{b}_1)_{\mathrm{NLS}} = 1/(\hat{a}_1)_{\mathrm{NLS}}$. Hence, $(\hat{b}_1)_{\mathrm{NLS}}$ and $G(s_k, \hat{\theta}_{\mathrm{NLS}}(Z))$ are consistent estimates, which illustrates the independence of $G(s_k, \hat{\theta}_{\mathrm{NLS}}(Z))$ on the parameter constraint used (quick tool number 4).

7.10 TOTAL LEAST SQUARES

7.10.1 Introduction

The total least squares (TLS) approach requires a model equation that is linear in the model parameters θ. Transfer function models (7-7) and (7-8), where $G(\Omega, \theta)$ and $T(\Omega, \theta)$ are parameterized as rational forms (5-20), (5-25) and (5-37), (5-40), can be made linear in θ by multiplication with the denominator polynomial $A(\Omega_k, \theta)$. This is not the case for the other parameterizations and, therefore, the TLS estimators can only be applied to rational forms without delay. Hence, the linear set of equations that needs to be solved in total least square sense is

$$e(\Omega_k, \theta, Z(k)) \approx 0 \qquad k = 1, 2, \ldots, F \tag{7-58}$$

with $e(\Omega_k, \theta, Z(k))$ the equation error (7-9) or (7-10). They can be written as

$$J(Z)\theta \approx 0 \text{ or } J_{\mathrm{re}}(Z)\theta \approx 0 \tag{7-59}$$

with $J(Z) = \partial e(\theta, Z)/\partial \theta$ the Jacobian of the vector $e(\theta, Z)$ ($e_{[k]}(\theta, Z) = e(\Omega_k, \theta, Z(k))$), and where $(\ )_{\mathrm{re}}$ stacks the real and imaginary parts on top of each other

$$J_{\mathrm{re}}(Z) = \begin{bmatrix} \mathrm{Re}(J(Z)) \\ \mathrm{Im}(J(Z)) \end{bmatrix} \tag{7-60}$$

(see Section 13.8). Operation (7-60) is necessary to ensure that the solution θ is real. A left and a right weighting can be applied to (7-59)

$$(WJ(Z)C^{-1})(C\theta) \approx 0 \text{ or } (W_{\mathrm{Re}}J_{\mathrm{re}}(Z)C^{-1})(C\theta) \approx 0 \tag{7-61}$$

where $W \in \mathbb{C}^{F \times F}$ and $C \in \mathbb{R}^{n_\theta \times n_\theta}$ are regular matrices and where $(WJ(Z))_{\mathrm{re}} = W_{\mathrm{Re}}J_{\mathrm{re}}(Z)$ with

$$W_{\text{Re}} = \begin{bmatrix} \text{Re}(W) & -\text{Im}(W) \\ \text{Im}(W) & \text{Re}(W) \end{bmatrix} \tag{7-62}$$

(see Lemma 13.4). A diagonal left weighting matrix W influences each row of $J(Z)$ separately and makes it possible to introduce a frequency-dependent weighting of the residuals $e(\Omega_k, \theta, Z(k))$. The right weighting matrix C influences each row of $J(Z)$ in exactly the same way and, hence, will not introduce a frequency-dependent weighting of the residuals $e(\Omega_k, \theta, Z(k))$. It can be used to influence the noise characteristics of $J(Z)$ (see Section 7.10.3).

The total least squares solution of the weighted problem (7-61) tries to find a modified matrix $\tilde{J}_{\text{re}}$, which is as close as possible to $J_{\text{re}}(Z)$ (in Frobenius norm, see Section 13.3), and a vector θ satisfying $\tilde{J}_{\text{re}}\theta = 0$. The unknown parameters in the total least squares problem are, hence, the matrix $\tilde{J}_{\text{re}}$ ($2Fn_\theta$ real parameters) and the model parameters θ (n_θ real numbers). These parameters are related to each other by the model equation $\tilde{J}_{\text{re}}\theta = 0$ ($2F$ real equations), so that the total number of free parameters equals $(2F+1)n_\theta - 2F$. This should be compared with the measured matrix $J_{\text{re}}(Z)$ ($2Fn_\theta$ real numbers), which gives a redundancy of $2F - n_\theta$. It shows that increasing F will (most probably) give more information about θ, but not about $\tilde{J}_{\text{re}}$. Indeed, no additional information can be accumulated about $2Fn_\theta$ real parameters starting from $2F$ real measurements.

The matrix $\tilde{J}_{\text{re}}$ and the vector θ are the solution of

$$\arg\min_{\tilde{J}_{\text{re}},\,\theta} \left\| W_{\text{Re}}(J_{\text{re}}(Z) - \tilde{J}_{\text{re}})C^{-1} \right\|_F^2 \text{ subject to } \tilde{J}_{\text{re}}\theta = 0 \text{ and } \|C\theta\|_2^2 = 1 \tag{7-63}$$

(Van Huffel and Vandewalle, 1991). After elimination of $\tilde{J}_{\text{re}}$ in (7-63), we get the following equivalences.

Lemma 7.23 (Total Least Squares Solution—Equivalences): The total least squares problem (7-63) is equivalent to

1. $\arg\min_\theta \|WJ(Z)\theta\|_2^2 / \|C\theta\|_2^2$
2. $\arg\min_\theta \|WJ(Z)\theta\|_2^2$ subject to $\|C\theta\|_2^2 = 1$
3. finding the eigenvector θ corresponding to the smallest generalized eigenvalue λ of the generalized eigenvalue problem $(W_{\text{Re}}J_{\text{re}}(Z))^T(W_{\text{Re}}J_{\text{re}}(Z))\theta = \lambda C^T C\theta$

Proof. See Appendix 7.H. □

Although we have assumed, during the proof, that the matrices W and C are nonsingular, it follows from Lemma 7.23 that the TLS solution remains well defined for singular weighting matrices W and C. In these cases, we take Lemma 7.23 as a definition of the total least squares solution. Equivalences 1 and 2 of Lemma 7.23 (nonlinear minimization of a cost function) are used to analyze the asymptotic properties ($F \to \infty$) of the TLS solution, while equivalence 3 is used to calculate the solution. The generalized eigenvalue problem (equivalence 3 of Lemma 7.23) can be calculated in a numerical stable way, even when C is singular, through the generalized singular value decomposition (GSVD) of the matrix pair $(W_{\text{Re}}J_{\text{re}}(Z), C)$ (see Section 13.4.2). The TLS solution is then the generalized right singular vector corresponding to the smallest generalized singular value of $(W_{\text{Re}}J_{\text{re}}(Z), C)$. When $C = I_{n_\theta}$ then the generalized eigenvalue problem reduces to an ordinary eigenvalue problem, which is solved in a numerically stable way through the singular value decomposition

(SVD) of the matrix $W_{\mathrm{Re}}J_{\mathrm{re}}(Z)$ (see Section 13.4.2). The TLS solution is then the right singular vector corresponding to the smallest singular value of $W_{\mathrm{Re}}J_{\mathrm{re}}(Z)$.

7.10.2 Total Least Squares

Putting $W = I_F$ and $C = I_{n_\theta}$ in (7-63) gives the total least squares estimate $\hat{\theta}_{\mathrm{TLS}}(Z)$. According to equivalence 2 of Lemma 7.23, $\hat{\theta}_{\mathrm{TLS}}(Z)$ is the minimizer of

$$V_{\mathrm{TLS}}(\theta, Z) = \sum_{k=1}^{F} |e(\Omega_k, \theta, Z(k))|^2 \text{ subject to } \|\theta\|_2^2 = 1 \tag{7-64}$$

with $e(\Omega_k, \theta, Z(k))$ the equation error (7-9) or (7-10) (proof: see Appendix 7.J). It shows that the total least squares solution (7-64) is nothing other than the linear least squares solution (7-32) with parameter constraint $\|\theta\|_2^2 = 1$. Hence, $\hat{\theta}_{\mathrm{TLS}}(Z)$ has the same asymptotic properties ($F \to \infty$) as $\hat{\theta}_{\mathrm{LS}}(Z)$: in general, $\hat{\theta}_{\mathrm{TLS}}(Z)$ is inconsistent and $\tilde{\theta}_{\mathrm{TLS}}(Z_0)$, $\theta_{*\mathrm{TLS}}$ depend on the signal-to-noise ratio. To reveal when $\hat{\theta}_{\mathrm{TLS}}(Z)$ is consistent, we use equivalence 1 of Lemma 7.23

$$V_{\mathrm{TLS}}(\theta, Z) = \|J(Z)\theta\|_2^2/\|\theta\|_2^2 = \sum_{k=1}^{F} |e(\Omega_k, \theta, Z(k))|^2/\|\theta\|_2^2 \tag{7-65}$$

Taking the expected value of (7-65) gives (7-19) with $V_F(\theta) = \mathcal{E}\{V_{\mathrm{TLS}}(\theta, Z)\}/F$, and

$$\mathcal{E}\{v_F(\theta, N_Z)\} = \frac{1}{F}\theta^T C_J \theta/\|\theta\|_2^2 = \frac{1}{F}\sum_{k=1}^{F} \sigma_e^2(\Omega_k, \theta)/\|\theta\|_2^2 \tag{7-66}$$

where $\sigma_e^2(\Omega_k, \theta)$ is defined in (7-34) and where

$$C_J = \mathcal{E}\{j_{\mathrm{re}}^T(N_Z)j_{\mathrm{re}}(N_Z)\} = \mathcal{E}\{\mathrm{Re}(j^H(N_Z)j(N_Z))\} \text{ with } j(N_Z) = J(Z) - J(Z_0) \tag{7-67}$$

is the column covariance matrix of $j_{\mathrm{re}}(N_Z)$ (see Appendix 7.I). Note that $j(N_Z) \neq J(N_Z)$ for model (7-10). Applying quick analysis tool number 2 (see Section 7.5) to (7-66) shows that the total least squares estimator $\hat{\theta}_{\mathrm{TLS}}(Z)$ is consistent if C_J is proportional to I_{n_θ}: $C_J = \sigma^2 I_{n_\theta}$.

Like the LS estimate, the total least squares solution can be improved by adding an appropriate frequency dependent. The TLS version of (7-37) is found by making the choice $C = I_{n_\theta}$ and

$$W = \mathrm{diag}(W(\Omega_1, \theta^{(i-1)}), W(\Omega_2, \theta^{(i-1)}), \ldots, W(\Omega_F, \theta^{(i-1)})) \tag{7-68}$$

with $W(\Omega_k, \theta^{(i-1)}) \in \mathbb{R}$, in (7-63) (proof: see Appendix 7.J). The weighted total least squares solution is calculated as the right singular vector corresponding to the smallest singular value of $W_{\mathrm{Re}}J_{\mathrm{re}}(Z)$.

7.10.3 Generalized Total Least Squares

The total least squares estimator (7-64) is inconsistent because the column covariance matrix C_J (7-67) is different from $\sigma^2 I_{n_\theta}$ (see Section 7.10.2). Taking as right weighting C a square root of C_J

$$C = C_J^{1/2} \text{ such that } C^T C = C_J \tag{7-69}$$

(see Section 13.4.4), then the column covariance matrix of $j_{\mathrm{re}}(N_Z)C^{-1}$, with $j(N_Z) = J(Z) - J(Z_0)$, becomes

$$\mathcal{E}\{C^{-T} j_{\mathrm{re}}^T(N_Z) j(N_Z) C^{-1}\} = C^{-T} C_J C^{-1} = I_{n_\theta} \tag{7-70}$$

It shows that the total least squares estimator can be made consistent by an appropriate choice of the right weighting matrix C. Note that the calculation of C requires knowledge of the noise (co)variances (Assumption 7.5).

Putting $W = I_F$ and $C = C_J^{1/2}$ in (7-63) gives the generalized total least squares (GTLS) estimate $\hat{\theta}_{\mathrm{GTLS}}(Z)$. According to equivalence 1 of Lemma 7.23, $\hat{\theta}_{\mathrm{GTLS}}(Z)$ is the minimizer of

$$V_{\mathrm{GTLS}}(\theta, Z) = \frac{\sum_{k=1}^{F} |e(\Omega_k, \theta, Z(k))|^2}{\sum_{k=1}^{F} \sigma_e^2(\Omega_k, \theta)} \tag{7-71}$$

with $\sigma_e^2(\Omega_k, \theta) = \mathrm{var}(e(\Omega_k, \theta, N_Z(k)))$ (see (7-34)) and $e(\Omega_k, \theta, Z(k))$ the equation error (7-9) or (7-10) (proof: see Appendix 7.J). In Swevers et al. (1992) the generalized total least squares method was applied for the first time to identify discrete-time models from noisy input-output measurements ((7-71) with equation error (7-9) and $\Omega = z^{-1}$). Due to the equal weighting of the residuals $e(\Omega_k, \theta, Z(k))$ over all frequencies in (7-71), the GTLS estimate suffers from the same problem as the LS and TLS estimates: it overemphasizes the high-frequency errors. Although this effect is not apparent in the second-order simulation example (see Figure 7-4 on page 200), it is visible on more complex systems (see Figure 7-8 on page 230).

Because $V_{\mathrm{GTLS}}(\theta, Z)$ is quadratic-in-the-measurements Z, Theorem 7.21, with $V_F(\theta, Z) = V_{\mathrm{GTLS}}(\theta, Z)$, is valid for $\hat{\theta}_{\mathrm{GTLS}}(Z)$. Due to the denominator in (7-71), the expression for the limit cost $V_*(\theta)$ in property 5 is somewhat more complicated (see Exercise 7.6). Taking the expected value of (7-71) gives (7-19) with

$$\mathcal{E}\{v_F(\theta, Z)\} = 1 \tag{7-72}$$

As $\mathcal{E}\{v_F(\theta, Z)\}$ is independent of θ, the generalized total least squares estimate $\hat{\theta}_{\mathrm{GTLS}}(Z)$ is consistent, and $\hat{\theta}_{\mathrm{GTLS}}(Z_0)$, $\theta_{*\mathrm{GTLS}}$ are the noiseless solutions when model errors are present (apply quick analysis tools number 2 and 3 of Section 7.5). From (7-71), it follows that $V_{\mathrm{GTLS}}(\lambda\theta, Z) = V_{\mathrm{GTLS}}(\theta, Z)$ so that $\hat{\theta}_{\mathrm{GTLS}}(Z)$ is independent of the particular, chosen constraint $a_i = 1$, $b_i = 1$, or $\|\theta\|_2^2 = 1$ (quick tool number 4). See Table 7-5 on page 238 for an overview of the properties of the GTLS estimator.

To deemphasize the high frequency errors in (7-71), a left weighting matrix W should be added, and at the same time, to keep the consistency, the right weighting C should be adapted. For example, the choice,

$$W = \mathrm{diag}(W(\Omega_1), W(\Omega_2), \ldots, W(\Omega_F)) \text{ with } W(\Omega_k) \in \mathbb{R} \tag{7-73}$$

$$C = C_{WJ}^{1/2} \text{ such that } C^T C = C_{WJ} = \mathcal{E}\{\mathrm{Re}((Wj(N_Z))^H (Wj(N_Z)))\} \tag{7-74}$$

in (7-63), with C_{WJ} the column covariance matrix of $W_{\text{Re}}j_{\text{re}}(N_Z)$ (see (7-67)), gives the following weighted generalized total least squares cost function:

$$V_{\text{WGTLS}}(\theta, Z) = \frac{\sum_{k=1}^{F} W^2(\Omega_k)|e(\Omega_k, \theta, Z(k))|^2}{\sum_{k=1}^{F} W^2(\Omega_k)\sigma_e^2(\Omega_k, \theta)} \tag{7-75}$$

(see Appendix 7.J). Although the weight $W(\Omega_k)$ does not affect the consistency of the weighted generalized total least squares estimate $\hat{\theta}_{\text{WGTLS}}(Z)$, it can seriously influence its uncertainty. A motivated choice will be presented in Section 7.12.3. Apart from this effect, $\hat{\theta}_{\text{WGTLS}}(Z)$ has the same asymptotic properties as $\hat{\theta}_{\text{GTLS}}(Z)$. The estimate $\hat{\theta}_{\text{WGTLS}}(Z)$ is calculated as the generalized right singular vector corresponding to the smallest generalized singular value of the matrix pair $(W_{\text{Re}}J_{\text{re}}(Z), C_{WJ}^{1/2})$. Note that the column covariance matrix C_{WJ} in (7-74) is singular under Assumption 7.20(i) or 7.20(ii) (see Appendix 7.K).

7.11 MAXIMUM LIKELIHOOD

7.11.1 The Maximum Likelihood Solution

To construct the maximum likelihood solution, starting from the frequency domain data (7-1) or (7-2), we need the probability density function (pdf) of the frequency domain errors $N_Z(k) = [N_Y(k)\ N_U(k)]^T$, $k = 1, 2, \ldots, F$. For a frequency domain experiment $N_Z(k)$ is independent over k (Assumption 7.4), while for a time domain experiment $N_Z(k)$ is asymptotically ($F \to \infty$) independent over k and circular complex normally distributed (see Sections 7.6.1 and 14.16). Therefore, it is reasonable to construct the maximum likelihood (ML) solution under the assumption that $N_Z(k)$ is independent (over k) circular complex normally distributed with known covariance matrix (Assumptions 7.5, 7.18, and 7.19). We also assume that the true excitation $U_0(k)$ and, hence, also the true response $Y_0(k)$ are deterministic (Assumption 7.19).

Because the true input $U_0(k)$ and output $Y_0(k)$ DFT spectra in (7-1) are unknown, they should be estimated and parameterized as $U_p(k)$, $Y_p(k)$. The unknown parameters in the errors-in-variables approach are, hence, the unknown input $U_p(k)$ and output $Y_p(k)$ DFT spectra ($4F$ real numbers) and the model parameters θ (n_θ real numbers). These parameters are related to each other by the model equations

$$e(\Omega_k, \theta, Z_p(k)) = 0 \qquad k = 1, 2, \ldots, F \tag{7-76}$$

with $e(\Omega_k, \theta, Z(k))$ the equation error (7-9) or (7-10) ($2F$ real equations); thus, the total number of free parameters equals $2F + n_\theta$. This should be compared with the number of measured input $U(k)$ and output $Y(k)$ spectra ($4F$ real numbers), which gives a redundancy of $2F - n_\theta$. It shows that increasing F will (most probably) give more information about θ but not about $U_p(k)$ and $Y_p(k)$. Indeed, four new real parameters are added for each frequency.

Under Assumptions 7.5, 7.18, and 7.19 the negative log-likelihood function is

$$\begin{aligned} -\ln f_{N_Z}(Z, Z_p, \theta) &= (Z - Z_p)^H C_{N_Z}^{+}(Z - Z_p) + c \\ C_{N_Z} &= \text{diag}(\text{Cov}(N_Z(1)), \text{Cov}(N_Z(2)), \ldots, \text{Cov}(N_Z(F))) \end{aligned} \tag{7-77}$$

with + the Moore-Penrose pseudoinverse and c a constant, independent of Z_{p} and θ (see Appendix 7.L). (7-77) should be minimized w.r.t. Z_{p} and θ subject to the constraints (7-76). This constrained minimization problem can be solved using Lagrange multipliers $\lambda \in \mathbb{C}^F$

$$(Z - Z_{\text{p}})^H C_{N_Z}^{+}(Z - Z_{\text{p}}) + \text{Re}(\lambda^H e(\theta, Z_{\text{p}})) \tag{7-78}$$

Elimination of Z_{p} in (7-78) gives the maximum likelihood cost function

$$V_{\text{ML}}(\theta, Z) = \sum\nolimits_{k=1}^{F} \frac{|e(\Omega_k, \theta, Z(k))|^2}{\sigma_e^2(\Omega_k, \theta)} \tag{7-79}$$

(see Appendix 7.L) with $e(\Omega_k, \theta, Z(k))$ the equation error (7-9) or (7-10) and $\sigma_e^2(\Omega_k, \theta)$ the variance of the equation error where the measurements Z have been replaced by the noise on the measurements N_Z; see (7-34). If DC (Ω_0) and Nyquist ($\Omega_{N/2}$) are present in the data, then

$$\frac{1}{2}\frac{|e(\Omega_0, \theta, Z(0))|^2}{\sigma_e^2(\Omega_0, \theta)} + \frac{1}{2}\frac{|e(\Omega_{N/2}, \theta, Z(N/2))|^2}{\sigma_e^2(\Omega_{N/2}, \theta)} \tag{7-80}$$

should be added to the cost function (7-79) (see Appendix 7.L). Dividing the numerator and denominator of each term in the sum (7-79) by $|A(\Omega_k, \theta)|^2$ gives

$$V_{\text{ML}}(\theta, Z) = \sum\nolimits_{k=1}^{F} \frac{|Y(k) - Y(\Omega_k, \theta)|^2}{\sigma_Y^2(\Omega_k, \theta)} \tag{7-81}$$

with $\sigma_Y^2(\Omega_k, \theta)$ the variance of the output error, where the measurements Z have been replaced by the noise on the measurements N_Z; see (7-44). Under this form, it is suitable for any parameterization of the transfer function model (see Sections 5.2 and 5.3). Cost functions (7-79) and (7-81) can also be written as

$$V_{\text{ML}}(\theta, Z) = \sum\nolimits_{k=1}^{F} |\varepsilon(\Omega_k, \theta, Z(k))|^2 \tag{7-82}$$

where $\varepsilon(\Omega_k, \theta, Z(k))$ are the respective weighted residuals,

$$\varepsilon(\Omega_k, \theta, Z(k)) = e(\Omega_k, \theta, Z(k))/\sigma_e(\Omega_k, \theta) \tag{7-83}$$

$$\varepsilon(\Omega_k, \theta, Z(k)) = (Y(k) - Y(\Omega_k, \theta))/\sigma_Y(\Omega_k, \theta) \tag{7-84}$$

with $\text{var}(\varepsilon(\Omega_k, \theta, N_Z(k))) = 1$. The maximum likelihood estimate $\hat{\theta}_{\text{ML}}(Z)$ is the minimizer of (7-82) (see Appendix 7.L, Section 7.L.4 for the numerical implementation).

Using $\hat{\theta}_{\text{ML}}(Z)$, the maximum likelihood estimates $\hat{U}_{\text{ML}}(k)$ and $\hat{Y}_{\text{ML}}(k)$ of the input and output DFT spectra can be calculated, namely

$$\begin{aligned} \hat{Y}_{\mathrm{ML}}(k) &= Y(k) - (\sigma_Y^2(k)\bar{A}(\Omega_k, \hat{\theta}) - \sigma_{YU}^2(k)\bar{B}(\Omega_k, \hat{\theta}))\frac{e(\Omega_k, \hat{\theta}, Z(k))}{\sigma_e^2(\Omega_k, \hat{\theta})} \\ \hat{U}_{\mathrm{ML}}(k) &= U(k) - (\bar{\sigma}_{YU}^2(k)\bar{A}(\Omega_k, \hat{\theta}) - \sigma_U^2(k)\bar{B}(\Omega_k, \hat{\theta}))\frac{e(\Omega_k, \hat{\theta}, Z(k))}{\sigma_e^2(\Omega_k, \hat{\theta})} \end{aligned} \tag{7-85}$$

with $\hat{\theta} = \hat{\theta}_{\mathrm{ML}}(Z)$ (see Appendix 7.L). If the input is known ($\sigma_U^2(k) = 0$ and $\sigma_{YU}^2(k) = 0$), then (7-85) reduces to

$$\begin{aligned} \hat{Y}_{\mathrm{ML}}(k) &= G(\Omega_k, \hat{\theta}_{\mathrm{ML}}(Z))U_0(k) \;\; (+\, T(\Omega_k, \hat{\theta}_{\mathrm{ML}}(Z))) \\ \hat{U}_{\mathrm{ML}}(k) &= U_0(k) \end{aligned} \tag{7-86}$$

and the ML estimate $\hat{Y}_{\mathrm{ML}}(k)$ is nothing other than the output predicted by the model.

7.11.2 Discussion

The maximum likelihood solution (7-82) weights the equation or output error at each frequency Ω_k with its measurement uncertainty, so that frequency bands with high-quality measurements ($\sigma_Y^2(k)$ and $\sigma_U^2(k)$ are "small") contribute more to the ML cost than frequency bands with poor-quality measurements ($\sigma_Y^2(k)$ and $\sigma_U^2(k)$ are "large"). Hence, in a natural way, the ML cost gives much confidence to accurate measurements while it rejects noisy measurements. Inspection of the variance of the output error (7-44) leads to the following observations:

1. In the uncorrelated case ($\sigma_{YU}^2(k) = 0$) the relative importance of the input disturbance w.r.t. the output disturbance is given by the model-dependent ratio

$$\frac{|G(\Omega_k, \theta)|^2\sigma_U^2(k)}{\sigma_Y^2(k)} \tag{7-87}$$

2. The significance of the correlation between the input and output disturbances is assessed by the model-dependent ratio

$$\rho(k) = \frac{-2\mathrm{Re}(\sigma_{YU}^2(k)\bar{G}(\Omega_k, \theta))}{\sigma_Y^2(k) + |G(\Omega_k, \theta)|^2\sigma_U^2(k)} \tag{7-88}$$

3. If the measurement errors $M_Y(k)$ and $M_U(k)$ in Figure 7-3 on page 186 are uncorrelated, then the sign of $\rho(k)$ in (7-88) determines the behavior of the generator noise $N_g(k)$. If $\rho(k) < 0$, then the variance $\sigma_Y^2(\Omega_k, \theta)$ (see (7-44)) is decreased w.r.t. the uncorrelated case ($\sigma_{YU}^2(k) = 0$), which means that $N_g(k)$ contributes constructively to the excitation signal $U_0(k) + N_g(k)$ at frequency Ω_k. If $\rho(k) > 0$, then the variance $\sigma_Y^2(\Omega_k, \theta)$ is increased w.r.t. the uncorrelated case, which means that $N_g(k)$ acts as a disturbing noise source at frequency Ω_k.

If the Assumptions 7.5, 7.18, and 7.19 made to construct (7-82) are not fulfilled, for example, the errors $N_Z(k)$ are not normally distributed, then (7-82) is no longer the maximum likelihood solution of the problem. The same is true if the excitation is not deterministic. If the errors $N_Z(k)$ are non-Gaussian, independent (over k), circular complex distributed random variables, then (7-82) is a Markov estimator (see Section 17.2.2), for which all the results of Chapter 17 apply. If the errors are not circular complex distributed,

$\mathcal{E}\{N_Z(k)N_Z^T(k)\} \neq 0$, then $\text{Cov}(N_Z(k))$ does not contain all the information included in $\text{Cov}((N_Z(k))_{\text{re}})$, and (7-82) is no longer the Markov solution of the problem (see Exercise 7.7, and Section 17.2). In that case (7-82) is just a weighted nonlinear least squares solution. With some misuse of terminology $\hat{\theta}_{\text{ML}}(Z)$ will, independent of the true noise properties, denote the minimizer of (7-82).

7.11.3 Asymptotic Properties

The general maximum likelihood properties listed in Section 1.5.3 are NOT VALID for the maximum likelihood solution (7-82) of the errors-in-variables problem. Indeed, they have been shown under the assumption that the number of estimated parameters does not increase with the amount of data, while the number of free parameters in the errors-in-variables problem is $2F + n_\theta$ and increases with the number of frequencies F. Therefore, even under the ideal Assumptions 7.5, 7.18, and 7.19, the consistency, asymptotic normality, and asymptotic efficiency still have to be proved, and it is not self-evident at all that the ML solution (7-82) will have nice asymptotic ($F \to \infty$) properties. We will first study the properties of $\hat{\theta}_{\text{ML}}(Z)$ under less restrictive noise assumptions than those made to construct the ML solution.

Because $V_{\text{ML}}(\theta, Z)$ is quadratic-in-the-measurements Z, Theorem 7.21, with $V_F(\theta, Z) = V_{\text{ML}}(\theta, Z)/F$, is valid for $\hat{\theta}_{\text{ML}}(Z)$. Taking the expected value of (7-82) gives (7-19) with

$$\mathcal{E}\{v_F(\theta, Z)\} = 1 \tag{7-89}$$

It shows that $\hat{\theta}_{\text{ML}}(Z)$ is consistent and, if there are model errors, that $\tilde{\theta}_{\text{ML}}(Z_0)$, $\theta_{*\text{ML}}$ are the noiseless solutions (apply quick analysis tools number 2 and 3 of Section 7.5). The noiseless solutions are obtained by decreasing the input and output noise levels simultaneously to zero while maintaining the ratios $\sigma_Y^2(k)/\sigma_U^2(k)$ and $\sigma_{YU}^2(k)/\sigma_U^2(k)$ constant (see quick tool number 3). Changing the ratios $\sigma_Y^2(k)/\sigma_U^2(k)$ and $\sigma_{YU}^2(k)/\sigma_U^2(k)$ introduces a frequency-dependent modification of $\sigma_e^2(\Omega_k, \theta)$ or $\sigma_Y^2(\Omega_k, \theta)$ in the cost function (7-82) and, hence, changes the noiseless solutions. We also have $V_{\text{ML}}(\lambda\theta, Z) = V_{\text{ML}}(\theta, Z)$ (see (7-82)) so that $\hat{\theta}_{\text{ML}}(Z)$ is independent of the particular constraint chosen, for example, $a_i = 1$, $b_i = 1$, or $\|\theta\|_2^2 = 1$ (quick tool number 4). We conclude that $\hat{\theta}_{\text{ML}}(Z)$ is, in general, consistent and asymptotically normally distributed. From property 8 of Theorem 7.21, it follows that $\hat{\theta}_{\text{ML}}(Z)$ is, in general, inefficient. It is asymptotically efficient only if the input-output disturbances stem from one noncoherent noise source (see Assumption 7.20).

It can be seen from (7-85) that the estimates $\hat{U}_{\text{ML}}(k)$ and $\hat{Y}_{\text{ML}}(k)$ of the input and output DFT spectra are in general inconsistent, even if $\hat{\theta}_{\text{ML}}(Z)$ is consistent. This can easily be understood as follows: making more measurements (increasing F) will not increase the knowledge of the input and output DFT spectra at one particular frequency (no noise averaging effect occurs). Because they are inconsistent, it makes no sense to calculate, for example, an "improved" frequency response function estimate using $\hat{U}_{\text{ML}}(k)$ and $\hat{Y}_{\text{ML}}(k)$. If the input is known and $\hat{\theta}_{\text{ML}}(Z)$ is consistent, then $\hat{Y}_{\text{ML}}(k)$ is consistent (see (7-86)). Similarly, if the output is known and $\hat{\theta}_{\text{ML}}(Z)$ is consistent, then $\hat{U}_{\text{ML}}(k)$ is consistent.

As the properties of $\hat{\theta}_{\text{ML}}(Z)$ are also valid under the more restrictive Assumptions 7.5, 7.18, and 7.19, it follows from Theorem 7.21 that the maximum likelihood estimator ((7-82) with Assumptions 7.5, 7.18, and 7.19) is consistent and asymptotically normally distributed but that it is not asymptotically efficient (note the difference from the general maximum likelihood properties of Section 1.5.3). An inefficiency term is present; it tends to zero as the noise level v tends to zero

$$\mathrm{Cov}(\delta_\theta(Z)) = Fi^{-1}(\theta_0)(I_{n_\theta} + O(\upsilon)) \tag{7-90}$$

(see Appendix 7.E —asymptotic efficiency). For errors N_Z with an even pdf, the deviation in (7-90) is an $O(\upsilon^2)$. In practice the inefficiency term will be neglected when calculating the covariance matrix of the estimates (see Section 7.11.4). The ML estimator is asymptotically efficient if only one noncoherent disturbing noise source is present (Theorem 7.21). This corresponds to the case where the total number of estimated parameters does not increase with F (see Appendix 7.M), thus the general maximum likelihood properties of Section 1.5.3 are valid. Note that the consistency and asymptotic normality properties of the ML estimator ((7-82) with Assumptions 7.5, 7.18, and 7.19) have been shown in Theorem 7.21 under much less restrictive noise assumptions than those made to construct the ML solution. The errors $N_Z(k)$ may be non-Gaussian, correlated over the frequencies k, and noncircular complex $\mathcal{E}\{N_Z(k)N_Z^T(k)\} \neq 0$. It shows the *robustness* of the consistency and asymptotic normality properties of the ML estimator w.r.t. Assumptions 7.5, 7.18, and 7.19. See Table 7-5 on page 238 for an overview of the properties of the ML estimator.

7.11.4 Calculation of Uncertainty Bounds

According to property 3 of Theorem 7.21, the covariance matrix of the truncated estimator $\hat{\underline{\theta}}_{\mathrm{ML}}(Z)$ (see (7-24)) is asymptotically $(F \to \infty)$ given by expression (7-26)

$$\mathrm{Cov}(\hat{\underline{\theta}}_{\mathrm{ML}}(Z)) = \mathrm{Cov}(\delta_\theta(Z))(I_{n_\theta} + O(F^{-1/2})) \tag{7-91}$$

(see Theorem 15.30). Expression (7-26) for $\mathrm{Cov}(\delta_\theta(Z))$ is not really tractable because it requires, for example, the third- and fourth-order moments of the noise, which are mostly unknown. An approximation for "small" model errors $(\mu \to 0)$ and "large" signal-to-noise ratios $(\upsilon \to 0)$ can be calculated. Applying quick tool number 6 of Section 7.5 to (7-26) yields

$$\begin{aligned} \mathrm{Cov}(\delta_\theta(Z)) &= C_\theta(I_{n_\theta} + O(\upsilon) + O(\mu) + O(\mu^2\upsilon^{-2})\lambda(Z_0)) \\ C_\theta &= \left[\mathcal{E}\{\,2\mathrm{Re}\left(\left(\frac{\partial \varepsilon(\theta, Z_0)}{\partial \tilde{\theta}_{\mathrm{ML}}(Z_0)}\right)^H \left(\frac{\partial \varepsilon(\theta, Z_0)}{\partial \tilde{\theta}_{\mathrm{ML}}(Z_0)}\right)\right)\}\right]^{-1} = \upsilon^2 O(F^{-1}) \end{aligned} \tag{7-92}$$

where $\lambda(Z_0) = 1$ for random Z_0 and $\lambda(Z_0) = 0$ for deterministic Z_0 (see Exercise 17.10). If model errors are present $(\mu \neq 0)$, then the uncertainty of the estimated model parameters (7-92) does not decrease to zero for random excitations $(\lambda(Z_0) = 1)$ as the noise level υ tends to zero. To calculate (7-92) we need the true observations Z_0 and the minimizer $\tilde{\theta}_{\mathrm{ML}}(Z_0)$ of the expected value of the cost function, which are not available. An approximation is calculated by replacing Z_0 by Z and $\tilde{\theta}_{\mathrm{ML}}(Z_0)$ by $\hat{\underline{\theta}}_{\mathrm{ML}}(Z)$, giving

$$\mathrm{Cov}(\hat{\underline{\theta}}_{\mathrm{ML}}(Z)) \approx \left[2\mathrm{Re}\left(\left(\frac{\partial \varepsilon(\theta, Z)}{\partial \hat{\underline{\theta}}_{\mathrm{ML}}(Z)}\right)^H \left(\frac{\partial \varepsilon(\theta, Z)}{\partial \hat{\underline{\theta}}_{\mathrm{ML}}(Z)}\right)\right)\right]^{-1} \tag{7-93}$$

Note that the expression between brackets in (7-93) equals, within a factor of 2, the matrix of the normal equation in the last Newton-Gauss step (7-17). Together with property 4 of Theorem 7.21 and the results of Section 14.2, (7-93) allows the calculation of uncertainty regions with a given confidence level for any model-related quantity (see also Section 17.4.7).

7.12 APPROXIMATE MAXIMUM LIKELIHOOD

7.12.1 Introduction

Compared with the maximum likelihood solution, the iterative weighted linear least squares (IWLS) and weighted generalized total least squares (WGTLS) estimators have a big advantage as global minimizers. Their noise sensitivity can, however, be poor. The basic idea of this section is to construct estimators that combine the global minimization properties of the IWLS and WGTLS estimators with the good statistical properties of the ML estimator. The key to the solution of this problem is an appropriate choice of the frequency-dependent weighting. Comparing the IWLS and WGTLS cost functions (7-37) and (7-75) with the maximum likelihood solution (7-79) suggests that the "optimal" weighting is $W(\Omega_k) = \sigma_e^{-1}(\Omega_k, \theta)$. Because θ is unknown, it should be reconstructed iteratively as

$$W(\Omega_k, \theta^{(i-1)}) = \sigma_e^{-1}(\Omega_k, \theta^{(i-1)}) \tag{7-94}$$

The weighting (7-94) can even be relaxed as in (7-38)

$$W(\Omega_k, \theta^{(i-1)}) = \sigma_e^{-r}(\Omega_k, \theta^{(i-1)}) \text{ with } r \in [0, 1] \tag{7-95}$$

Special cases are no weighting, $r = 0$, and "full" weighting, $r = 1$.

Just as in Sections 7.8 and 7.10, the estimators of this section require that the plant transfer function $G(\Omega, \theta)$ and the transient term $T(\Omega, \theta)$ are parameterized as rational forms $B(\Omega, \theta)/A(\Omega, \theta)$ (see (5-20), (5-25)) and $I(\Omega, \theta)/A(\Omega, \theta)$ (see (5-37), (5-40)), respectively.

7.12.2 Iterative Quadratic Maximum Likelihood

Making the choice (7-95) in the IWLS cost function (7-37) gives the iterative quadratic maximum likelihood method,

$$V_{\text{IQML}}(\theta^{(i)}, Z) = \sum_{k=1}^{F} \frac{|e(\Omega_k, \theta^{(i)}, Z(k))|^2}{\sigma_e^{2r}(\Omega_k, \theta^{(i-1)})} \tag{7-96}$$

with $e(\Omega_k, \theta, Z(k))$ the equation error (7-9) or (7-10). If convergent ($\theta^{(i)} = \theta^{(i-1)}$ for i sufficiently large), the "full" IQML cost ((7-96) with $r = 1$) tends to the ML cost (7-79). This does not, however, imply that $\hat{\theta}_{\text{IQML}}(Z) = \hat{\theta}_{\text{ML}}(Z)$. Indeed, therefore, one needs that the derivatives of both cost functions w.r.t. θ are the same. This is not the case here so that $\hat{\theta}_{\text{IQML}}(Z) \neq \hat{\theta}_{\text{ML}}(Z)$. However, because the elementwise difference between both Jacobians is proportional to the residual $\varepsilon(\Omega_k, \theta^{(i-1)}, Z(k))$ (7-83) (see Exercise 7.8), both estimates will coincide ($\hat{\theta}_{\text{IQML}}(Z) \approx \hat{\theta}_{\text{ML}}(Z)$) for "sufficiently high" signal-to-noise ratios and "sufficiently small" modeling errors; otherwise the difference may be large. This is illustrated by the "high noise" simulation example of Figure 7-4 on page 200 (compare IQML and ML) and the "low noise" simulation example of Figure 7-8 on page 230 (compare IQML and ML). We conclude that the IQML estimator (7-96) is related to the ML solution (7-79) as the IWLS estimator (7-36) to the nonlinear least squares solution (7-42).

Because (7-96) is a special case of (7-37), the estimate $\hat{\theta}_{\text{IQML}}(Z)$ has the same asymptotic ($F \to \infty$) properties as $\hat{\theta}_{\text{IWLS}}(Z)$ (see Section 7.8.3): $\hat{\theta}_{\text{IQML}}(Z)$ is inconsistent, depends on the particular constraint chosen, and does not converge to a noiseless solution. See Table 7-5 on page 238 for an overview of the properties of the IQML estimator.

7.12.3 Bootstrapped Total Least Squares

Making the choice (7-95) in the WGTLS estimator (7-75) gives the bootstrapped total least squares (BTLS) method

$$V_{\text{BTLS}}(\theta^{(i)}, Z) = \frac{\sum_{k=1}^{F} \frac{|e(\Omega_k, \theta^{(i)}, Z(k))|^2}{\sigma_e^{2r}(\Omega_k, \theta^{(i-1)})}}{\sum_{k=1}^{F} \frac{\sigma_e^2(\Omega_k, \theta^{(i)})}{\sigma_e^{2r}(\Omega_k, \theta^{(i-1)})}} \tag{7-97}$$

with $\sigma_e^2(\Omega_k, \theta) = \text{var}(e(\Omega_k, \theta, N_Z(k)))$ (see (7-34)) and $e(\Omega_k, \theta, Z(k))$ the equation error (7-9) or (7-10). Relaxation of the weighting ($r < 1$) may be necessary if a lowly damped pole and zero are very close (relative to the spacing of the frequency grid) to each other. If convergent ($\theta^{(i)} = \theta^{(i-1)}$ for i sufficiently large), the "full" BTLS cost ((7-97) with $r = 1$) tends to the ML cost (7-79). The Jacobians of both estimators are, however, different, even for $i \to \infty$, and therefore $\hat{\theta}_{\text{BTLS}}(Z) \neq \hat{\theta}_{\text{ML}}(Z)$. Likewise, for IQML (see Section 7.12.2), the elementwise difference between both Jacobians is proportional to ML residual $\varepsilon(\Omega_k, \theta^{(i-1)}, Z(k))$ (7-83). In practice, it turns out that the difference is small for large signal-to-noise ratios such that the bootstrapped total least squares estimate $\hat{\theta}_{\text{BTLS}}(Z)$ is mostly (very) close to the maximum likelihood estimate $\hat{\theta}_{\text{ML}}(Z)$ (see Figure 7-4 on page 200 and Section 7.15). The estimate $\hat{\theta}_{\text{BTLS}}(Z)$ is calculated numerically in exactly the same way as the weighted generalized total least squares in Section 7.10.3.

The asymptotic ($F \to \infty$) properties of the first step of the iterative procedure (7-97) can be analyzed using Theorem 7.21 and Corollary 7.22. If the initial guess $\theta^{(0)}$ is deterministic, then Theorem 7.21 is valid and the bootstrapped total least squares estimate $\hat{\theta}_{\text{BTLS}}(Z) = \theta^{(1)}$ has the same properties as $\hat{\theta}_{\text{WGTLS}}(Z)$ (see Section 7.10.3). If the choice $\theta^{(0)} = \hat{\theta}(Z)$ is made, then it is obvious that (7-97) is no longer a quadratic function of the measurements Z. Assuming that the initial guess $\hat{\theta}(Z)$ satisfies the properties of Theorem 7.21, for example, $\hat{\theta}(Z) = \hat{\theta}_{\text{LS}}(Z)$ or $\hat{\theta}(Z) = \hat{\theta}_{\text{GTLS}}(Z)$, then Theorem 7.21 is still valid for $\hat{\theta}_{\text{BTLS}}(Z) = \theta^{(1)}$ with three minor modifications (see Corollary 7.22). The first step of (7-97) can be written as

$$V_{\text{BTLS}}(\theta, Z) = f_F(\theta, \theta^{(0)}, Z) \tag{7-98}$$

Taking the expected value of the cost function (7-98), where $\theta^{(0)} = \hat{\theta}(Z)$ has been replaced by its limit ($F \to \infty$) value θ_* gives (7-19) with $V_F(\theta) = \mathcal{E}\{f_F(\theta, \theta_*, Z)\}$ and

$$V_F(\theta) = \frac{\sum_{k=1}^{F} \frac{\mathcal{E}\{|e(\Omega_k, \theta, Z_0(k))|^2\}}{\sigma_e^{2r}(\Omega_k, \theta_*)}}{\sum_{k=1}^{F} \frac{\sigma_e^2(\Omega_k, \theta)}{\sigma_e^{2r}(\Omega_k, \theta_*)}} + 1 \tag{7-99}$$

Hence, the bootstrapped total least squares estimate $\hat{\theta}_{\text{BTLS}}(Z) = \theta^{(1)}$ is consistent, even if $\theta^{(0)} = \hat{\theta}(Z)$ is inconsistent (apply quick analysis tool number 2 of Section 7.5). If the limit value θ_* does not depend on the noise level ν, then $\tilde{\theta}_{\text{BTLS}}(Z_0)$, $\theta_{*\text{BTLS}}$ are the noiseless solutions when there are model errors (quick analysis tool number 3). This is the case for $\theta^{(0)} = \hat{\theta}_{\text{GTLS}}(Z)$ but not for $\theta^{(0)} = \hat{\theta}_{\text{LS}}(Z)$. From (7-97) it follows that $V_{\text{BTLS}}(\lambda\theta, Z) = V_{\text{BTLS}}(\theta, Z)$ so that $\hat{\theta}_{\text{BTLS}}(Z)$ is independent of the particular, chosen constraint $a_i = 1$, $b_i = 1$, or $\|\theta\|_2^2 = 1$ (quick tool number 4). Because $\theta^{(1)} = \hat{\theta}_{\text{BTLS}}(Z)$ satisfies Theorem 7.21, the same reasoning can be applied to $\theta^{(2)}$, and so on, showing that the estimates obtained in the successive iteration steps have exactly the same properties as $\theta^{(1)}$. We conclude that the BTLS algorithm (7-97) generates consistent estimates in each iteration step. Hence, the iterative algorithm can be stopped at any iteration number (four iterations are usually sufficient). Further iteration (hopefully) decreases the uncertainty in the nonasymptotic case ($F \neq \infty$). In the absence of model errors, $\tilde{\theta}(Z_0) = \theta_0$ or $\theta_* = \theta_0$ for model (7-8) with $\Omega = s$, it follows from Corollary 7.22 that the asymptotic ($F \to \infty$) uncertainty of $\hat{\theta}_{\text{BTLS}}(Z) = \theta^{(1)}$ with $\theta^{(0)} = \hat{\theta}(Z)$ equals that of $\hat{\theta}_{\text{BTLS}}(Z) = \theta^{(1)}$ with $\theta^{(0)} = \theta_*$ (see Appendix 7.N). See Table 7-5 on page 238 for an overview of the properties of the BTLS estimator.

7.12.4 Weighted (Total) Least Squares

The IQML and BTLS estimators need an initial guess of the model parameters to reconstruct the optimal ML weighting iteratively and, hence, are not self-starting. In this section, a noniterative approximation of the optimal ML weighting is given that does not require explicit knowledge of the model parameters θ.

The approximation is constructed as follows. Taking out the factor $|A(\Omega_k, \theta)||B(\Omega_k, \theta)|$ in the ML weighting (7-34) yields

$$\begin{aligned}\sigma_e^2(\Omega_k, \theta) = |A(\Omega_k, \theta)||B(\Omega_k, \theta)|(\sigma_{\hat{Y}}^2(k)/|G(\Omega_k, \theta)| + \sigma_{\hat{U}}^2(k)|G(\Omega_k, \theta)| \\ -2\mathrm{Re}(\sigma_{\hat{Y}\hat{U}}^2(k)\exp(-j\angle G(\Omega_k, \theta))))\end{aligned} \tag{7-100}$$

Replacing the unknown plant transfer function $G(\Omega_k, \theta)$ by the measured frequency response function $G(\Omega_k)$ or $Y(k)/U(k)$ and the factor $|A(\Omega_k, \theta)||B(\Omega_k, \theta)|$ by a θ-independent function $f(\Omega_k)$ in (7-100) gives the following approximation:

$$W^{-2}(\Omega_k) = f(\Omega_k)[\sigma_{\hat{Y}}^2(k)/|G(\Omega_k)| + \sigma_{\hat{U}}^2(k)|G(\Omega_k)| - 2\mathrm{Re}(\sigma_{\hat{Y}\hat{U}}^2(k)\exp(-j\angle G(\Omega_k)))] \tag{7-101}$$

The explicit form of the function $f(\Omega)$ depends on the particular domain Ω and is given below (see (7-103) and (7-104)). The reader is referred to Rolain and Pintelon (1999) for the rationale behind the construction of $f(\Omega)$. To avoid problems of division by zero in (7-101), regularization is applied in the frequency bands where $|G(\Omega_k)|$ is of the order of the magnitude of the noise standard deviation $\sigma_G(k)$:

$$W_{\text{reg}}^{-2}(\Omega_k) = \begin{cases} W^{-2}(\Omega_k) + \varepsilon W^{-2}(\Omega_{k+1}) & W^{-2}(\Omega_k) < \varepsilon W^{-2}(\Omega_{k+1}) \\ W^{-2}(\Omega_k) & \text{otherwise} \end{cases} \tag{7-102}$$

where ε is of the order of the numerical precision of the computer.

For *continuous-time systems*, $\Omega = s$, $\sqrt{s}$, or $\tanh(\tau_R s)$, the function $f(\Omega)$ has the form

$$\begin{aligned} f(\Omega_k) &= \frac{1}{2}(g_{n_a}(\Omega_k)|G(\Omega_k)| + g_{n_b}(\Omega_k)/|G(\Omega_k)|) \\ g_n(\Omega) &= (|\Omega|^{n+1} - 1)^2/(|\Omega| - 1)^2 \end{aligned} \tag{7-103}$$

Recall that the frequency axis is scaled by $\omega_{\text{scale}} = (\omega_{\min} + \omega_{\max})/2$ when identifying continuous-time systems (see Section 7.4), so that Ω in (7-103) represent the scaled frequency $(s \to s/\omega_{\text{scale}})$.

For *discrete-time systems*, $\Omega = z^{-1}$, the function $f(\Omega)$ has the form

$$\begin{aligned} f(z_k^{-1}) &= (g^n(f_k, f_L) + g^n(f_k, f_U))^2 \\ g(f_k, f) &= |\cos(\omega_k T_s) - \cos(\omega T_s)| + (||\cos(\omega_L T_s)| - 0.5| + ||\cos(\omega_U T_s)| - 0.5|)/2 \end{aligned} \tag{7-104}$$

with $n = \max(n_a, n_b) + 1$ and f_L, f_U the lower and upper frequencies of the "active" band of the plant. The active band $[f_L, f_U]$ is defined as the largest segment of continuous frequency points for which

$$\begin{aligned} & h(k) > \frac{1}{F}\sum\nolimits_{k=1}^{F} h(k) \\ & h(k) = |G(z_k^{-1})|/\sigma_T + \sigma_T/|G(z_k^{-1})| \end{aligned} \tag{7-105}$$

with σ_T^2 the mean (over the frequency) variance of the transfer function measurement $G(z_k^{-1})$ or $Y(k)/U(k)$,

$$\sigma_T^2 = \frac{1}{F}\sum\nolimits_{k=1}^{F} |G(z_k^{-1})|^2 \left(\frac{\sigma_Y^2(k)}{|Y(k)|^2} + \frac{\sigma_U^2(k)}{|U(k)|^2} - 2\text{Re}\left(\frac{\sigma_{YU}^2(k)}{Y(k)\bar{U}(k)}\right)\right) \tag{7-106}$$

The noise influence on $h(k)$ in (7-105) is reduced by a running sum filter with a window length equal to 1% of the number of available frequency points.

The weighting (7-101) can be used to construct optimally weighted linear least squares (WLS) (7-37) or weighted generalized total least squares (WGTLS) (7-75) estimators. Because the weighting is a strong nonlinear function of the measurements Z, it is very difficult, if not impossible, to make precise statements about the asymptotic behavior of the WLS and WGTLS estimates obtained. They are inconsistent but (hopefully) lie within the attraction basin of the global minimum of the ML cost function. Although (7-101) may be a rough approximation due to the lack of knowledge about θ, a sensible improvement of the estimates w.r.t. to the unweighted case is obtained, even if the approximated and exact weight differ by as much as two orders of magnitudes. This low sensitivity is the key to the success of the proposed method. The power of the weighting is illustrated in Figure 7-9 on page 231 for a sixth-order discrete time system.

7.13 INSTRUMENTAL VARIABLES

If two or more periods of the measured time signals are available, the measurements can be split up into two time records, each of them containing an integer number of signal periods. The DFT spectra calculated using the second time record can then be used as instrumental sequences for the linear least squares identification, based on the DFT spectra of the first time record (Van den Bos, 1991). The instrumental sequences obtained are almost ideal because they are strongly correlated with the true unknown DFT spectra and practically uncorrelated with the noise of the first time record (in case of colored noise a small but nonzero correlation may exist between the noise of the successive signal periods). The classical instrumental variable equations are asymmetric in the measurements and the instrumental sequences (see (1-58)). They can be made symmetric if the roles of the measurements and the instrumental sequences are interchanged and added to the original equations. Proceeding in this way, full use of the complete data set (measurements and instrumental sequences) is achieved. The equivalent cost function of the resulting enhanced instrumental variables estimator is

$$V_{\mathrm{IV}}(\theta, Z) = \sum_{k=1}^{F} \mathrm{Re}(e(\Omega_k, \theta, Z^{[1]}(k))\overline{e(\Omega_k, \theta, Z^{[2]}(k))}) \tag{7-107}$$

where [1] and [2] indicate that the spectra are calculated using, respectively, the first and the second experiment (time record). Note that the cost function (7-107) can take negative values. Likewise, for the LS (7-32), TLS (7-64), and GTLS (7-71) cost functions, the high-frequency errors are overemphasized in (7-107).

Although the cost function $V_{\mathrm{IV}}(\theta, Z)$ cannot be written under the quadratic form (7-11), Theorem 7.21, with $V_F(\theta, Z) = V_{\mathrm{IV}}(\theta, Z)/F$, is still valid for $\hat{\theta}_{\mathrm{IV}}(Z)$ (see Appendix 7.O). Assuming that the two experiments are independent, the expected value of (7-107) equals (7-19) with

$$\mathcal{E}\{v_F(\theta, N_Z)\} = 0 \tag{7-108}$$

Applying quick analysis tools number 2 and 3 of Section 7.5 shows that $\hat{\theta}_{\mathrm{IV}}(Z)$ is consistent and, when model errors are present $\tilde{\theta}_{\mathrm{IV}}(Z_0)$, $\theta_{*\mathrm{IV}}$ are the noiseless solutions. From (7-107) it follows that $V_{\mathrm{IV}}(\lambda\theta, Z) = \lambda^2 V_{\mathrm{IV}}(\theta, Z)$ so that $\hat{\theta}_{\mathrm{IV}}(Z)$ depends on the particular constraint $a_i = 1$ or $b_i = 1$ chosen (apply quick tool number 4). See Table 7-5 on page 238 for an overview of the properties of the IV estimator.

Note that the IV method lowers the bias of the corresponding LS estimates on the complete data set (DFT spectra of the first and second time records put together) at the price of a higher variance. The mean square error of the IV estimates tends asymptotically ($F \to \infty$) to zero, whereas that of the LS estimates tends asymptotically to the square of its bias. Hence, the IV method will perform better than the LS method for F sufficiently large. Compare, for example, the IV with the LS estimates in Figure 7-4 on page 200.

7.14 SUBSPACE ALGORITHMS

7.14.1 Model Equations

Subspace identification methods estimate the state space representation of (7-7), namely

$$G(\xi, \theta) = C(\xi I_{n_a} - A)^{-1} B + D \tag{7-109}$$

where $\xi = z$ for discrete-time systems and $\xi = s$ for continuous-time systems. The identification procedure starts from a transformed version of the state space equations (5-18) and (5-19). These are constructed as follows. Assume that the input is periodic and that an integer number of periods of the steady-state response is observed. The discrete Fourier transform (DFT) of (5-18) and (5-19) then becomes

$$\begin{aligned} \xi_k X(k) &= AX(k) + BU(k) \\ Y(k) &= CX(k) + DU(k) \end{aligned} \tag{7-110}$$

with $X(k)$ the DFT of the state vector $x(t)$. By recursive use of the second and the first equation of (7-110) we find that

$$\begin{aligned} \xi_k^p Y(k) &= \xi_k^{p-1}(C\xi_k X(k) + D\xi_k U(k)) \\ &= \xi_k^{p-1}(CAX(k) + CBU(k) + D\xi_k U(k)) \\ &= \dots \\ &= CA^p X(k) + (CA^{p-1}B + CA^{p-2}B\xi_k + \dots + CB\xi_k^{p-1} + D\xi_k^p)U(k) \end{aligned} \tag{7-111}$$

Writing the last equation of (7-111) for $p = 0, 1, \dots, r-1$ on top of each other gives

$$W_r(k)Y(k) = O_r X(k) + S_r W_r(k)U(k) \tag{7-112}$$

with

$$W_r(k) = \begin{bmatrix} 1 \\ \xi_k \\ \dots \\ \xi_k^{r-1} \end{bmatrix}, \; O_r = \begin{bmatrix} C \\ CA \\ \dots \\ CA^{r-1} \end{bmatrix} \text{ and } S_r = \begin{bmatrix} D & 0 & \dots & 0 & 0 \\ CB & D & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ CA^{r-2}B & CA^{r-3}B & \dots & CB & D \end{bmatrix} \tag{7-113}$$

Collecting (7-112) for $k = 1, 2, \dots, F$ gives

$$\mathbf{Y} = O_r\mathbf{X} + S_r\mathbf{U} \tag{7-114}$$

with

$$\begin{aligned} \mathbf{Y} &= \left[W_r(1)Y(1) \;\; W_r(2)Y(2) \;\; \dots \;\; W_r(F)Y(F)\right] \\ \mathbf{U} &= \left[W_r(1)U(1) \;\; W_r(2)U(2) \;\; \dots \;\; W_r(F)U(F)\right] \\ \mathbf{X} &= \left[X(1) \;\; X(2) \;\; \dots \;\; X(F)\right] \end{aligned} \tag{7-115}$$

The complex data matrices $\mathbf{Y}$ and $\mathbf{U}$ have r rows and F columns. $\mathbf{X}$ is a complex n_a by F matrix, and O_r and S_r are, respectively, real r by n_a and r by r matrices. Equation (7-114), with r larger than the model order n_a, is the basic model used in subspace identification.

The extended observability matrix O_r has the shift property

$$O_{r[1:r-1,:]}A = O_{r[2:r,:]} \tag{7-116}$$

which will be used in the identification procedure. O_r is not unique because it depends on the choice of the state variables. Indeed, replacing (A, B, C, D, X) by $(TAT^{-1}, TB, CT^{-1}, D, TX)$, with T an invertible matrix, in the state space equations (7-110) does not change the input-output transfer function (7-109) but does change O_r

$$O_r \to O_r T^{-1} \tag{7-117}$$

Note that $O_r\mathbf{X}$ and S_r in model equation (7-114) are invariant w.r.t. the invertible transformation T.

For identifiability purposes we will assume that the state space realization (7-110) is observable, $\text{rank}(O_r) = n_a$ for any $r \geq n_a$, and controllable,

$$\text{rank}([B \;\; AB \;\; \dots \;\; A^{q-1}B]) = n_a \tag{7-118}$$

for any $q \geq n_a$.

For noisy input-output DFT spectra, $N_U(k) \neq 0$ and $N_Y(k) \neq 0$, model (7-114) becomes

$$\mathbf{Y} = O_r\mathbf{X} + S_r\mathbf{U} + \mathbf{N_Y} - S_r\mathbf{N_U} \tag{7-119}$$

where $\mathbf{N_Y}$ and $\mathbf{N_U}$ have the same structure as $\mathbf{Y}$ and $\mathbf{U}$ in (7-115).

7.14.2 Subspace Identification Algorithms

Subspace identification algorithms are basically a three-step procedure. First, an estimate $\hat{O}_r$ of the extended observability matrix is obtained using model (7-119). This is the most difficult step and consists mainly of eliminating the term depending on the input and reducing the noise influence. Next, $\hat{A}$ and $\hat{C}$ are found as the least squares solution of the overdetermined set of equations (7-116) and as the first row of $\hat{O}_r$ (see (7-113)), respectively. Finally, $\hat{B}$ and $\hat{D}$ are found as the linear least squares solution of

$$V_{\text{SUB}}(C, D, \hat{A}, \hat{C}, Z) = \sum_{k=1}^{F} W^2(\xi_k)\left|Y(k) - (\hat{C}(\xi_k I_{n_a} - \hat{A})^{-1}B + D)U(k)\right|^2 \tag{7-120}$$

where $W(\xi_k)$ is a well-chosen real weighting function.

We present two algorithms, one for discrete-time systems ($\xi = z$), based on McKelvey et al. (1996), and one for continuous-time system ($\xi = s$), based on Van Overschee and De Moor (1996a). The numerically efficient implementation of these algorithms is due to Verhaegen (1994).

Algorithm 7.24 (Subspace Algorithm for Discrete-Time Systems)

1. Estimate O_r given the data $Y(k)$, $U(k)$ and the noise (co)variances $\sigma_Y^2(k)$, $\sigma_U^2(k)$, $\sigma_{YU}^2(k)$:

1a. Initialization:

(i) If $\sigma_U^2(k) \neq 0$, replace $Y(k)$, $U(k)$, and $\sigma_Y^2(k)$ by, respectively, $Y(k)/U(k)$, 1, and $\sigma_G^2(k)$ (7-53).

(ii) If the required transfer function model is improper, $n_a < n_b$, interchange the role of the input and output.

(iii) Choose a value of $r > n_a$ and form the matrices

$$Z = \begin{bmatrix} \mathrm{Re}(\mathbf{U}) & \mathrm{Im}(\mathbf{U}) \\ \mathrm{Re}(\mathbf{Y}) & \mathrm{Im}(\mathbf{Y}) \end{bmatrix} \text{ and } C_{\mathbf{Y}} = \mathrm{Re}(\mathbf{C}\mathbf{C}^H)$$

with $\mathbf{C} = [W_r(1)\sigma_Y(1) \; W_r(2)\sigma_Y(2) \; \ldots \; W_r(F)\sigma_Y(F)]$ and where $\mathbf{U}$ and $\mathbf{Y}$ are defined in (7-115), and $W_r(k)$ as in (7-113) with $\xi = z$.

1b. Elimination of the input term in (7-119): calculate the QR factorization of Z^T, $Z^T = QR$, or $Z = R^T Q^T$,

$$Z = \begin{bmatrix} R_{11}^T & 0 \\ R_{12}^T & R_{22}^T \end{bmatrix} \begin{bmatrix} Q_1^T \\ Q_2^T \end{bmatrix}$$

where R_{ij} are r by r blocks of the upper triangular matrix R.

1c. Reduction of the noise influence in (7-119): calculate the singular value decomposition of $C_{\mathbf{Y}}^{-1/2} R_{22}^T$,

$$C_{\mathbf{Y}}^{-1/2} R_{22}^T = U \Sigma V^T$$

where $C_{\mathbf{Y}}^{1/2}$ is a square root of $C_{\mathbf{Y}}$ (see Section 13.4.4), and estimate O_r as

$$\hat{O}_r = C_{\mathbf{Y}}^{1/2} U_{[:,\, 1:n_a]}$$

2. Estimate A and C, given the estimate $\hat{O}_r$: solve the shift property (7-116) in least squares sense and select the first row of $\hat{O}_r$

$$\hat{A} = \hat{O}^+_{r[1:r-1,\, :]} \hat{O}_{r[2:r,\, :]} \quad \text{and} \quad \hat{C} = \hat{O}_{r[1,\, :]}$$

with + the Moore-Penrose pseudoinverse (see Section 13.5).

3. Estimate B and D, given the estimates $\hat{A}$ and $\hat{C}$: minimize (7-120) w.r.t. B and D with $W(z_k) = 1/\sigma_Y(k)$.

Proof. See Appendix 7.R. □

One could use Algorithm 7.24 with $\xi = s$ for continuous-time systems. This works reasonably well for small values of r. However, for larger values, the matrix Z in Algorithm 7.24 becomes ill conditioned, resulting in poor estimates. This problem is solved by introducing two scalar orthogonal polynomial bases that orthogonalize, respectively, the first r rows of Z and the last r rows of Z. It can be shown that there are no other two scalar polynomial bases that result in a smaller condition number of Z (Rolain et al., 1995). The final algorithm is also a three-step procedure. First, a generalized extended observability matrix $\hat{O}_{r\perp}$ is estimated. This matrix has a generalized shift structure that is used to estimate A. Next, $\hat{A}$ and $\hat{C}$ are estimated using $\hat{O}_{r\perp}$. Finally, $\hat{B}$ and $\hat{D}$ are the linear least squares solution of (7-120).

Algorithm 7.25 (Subspace Algorithm for Continuous-Time Systems)

1. Estimate $O_{r\perp}$ given the data $Y(k)$, $U(k)$ and the noise (co)variances $\sigma_Y^2(k)$, $\sigma_U^2(k)$, $\sigma_{YU}^2(k)$:

1a. Initialization:

(i) If $\sigma_U^2(k) \neq 0$, replace $Y(k)$, $U(k)$, and $\sigma_Y^2(k)$ by, respectively, $Y(k)/U(k)$, 1, and $\sigma_G^2(k)$ (7-53).

(ii) If the required transfer function model is improper $n_a < n_b$, interchange the role of the input and output.

(iii) Choose a value of $r > n_a$ and normalize the frequencies s_k with $\omega_{\text{scale}} = (\omega_{\max} + \omega_{\min})/2$ $(s_k \to s_k/\omega_{\text{scale}})$.

1b. Orthogonalization of the output data: calculate the r by F matrix $\mathbf{Y}_\perp$ as follows: initialization:

$$\mathbf{Y}_{\perp[1,:]} = \mathbf{Y}_{[1,:]}/\alpha_1 \qquad \text{with} \quad \alpha_1 = \|\mathbf{Y}_{[1,:]}\|_2$$

$$\mathbf{Y}_{\perp[2,:]} = \mathbf{Y}_{\perp[1,:]}D_s/\alpha_2 \quad \text{with} \quad \alpha_2 = \|\mathbf{Y}_{\perp[1,:]}D_s\|_2$$

recursion: for $n = 3$ to r

$$\mathbf{Y}_{\perp[n,:]} = (\mathbf{Y}_{\perp[n-1,:]}D_s + \alpha_{n-1}\mathbf{Y}_{\perp[n-2,:]})/\alpha_n \quad \text{with}$$

$$\alpha_n = \|\mathbf{Y}_{\perp[n-1,:]}D_s + \alpha_{n-1}\mathbf{Y}_{\perp[n-2,:]}\|_2$$

where $Y_{[1,:]} = [Y(1)\ \ Y(2)\ \ \ldots\ \ Y(F)]$ and $D_s = \operatorname{diag}(s_1, s_2, \ldots, s_F)$.

1c. Orthogonalization of the input data: perform the same calculation as in step 1b, but starting from $\mathbf{U}_{[1,:]} = [U(1)\ \ U(2)\ \ \ldots\ \ U(F)]$. The result is an r by F matrix $\mathbf{U}_\perp$ and numbers β_n, $n = 1, 2, \ldots, r$.

1d. Form the following matrices:

$$Z_\perp = \begin{bmatrix} \operatorname{Re}(\mathbf{U}_\perp) & \operatorname{Im}(\mathbf{U}_\perp) \\ \operatorname{Re}(\mathbf{Y}_\perp) & \operatorname{Im}(\mathbf{Y}_\perp) \end{bmatrix} \quad \text{and} \quad C_{\mathbf{Y}_\perp} = \operatorname{Re}(\mathbf{C}_\perp \mathbf{C}_\perp^H)$$

where $\mathbf{C}_\perp$ is calculated by starting from $\mathbf{C}_{[1,:]} = [\sigma_Y(1)\ \ \sigma_Y(2)\ \ \ldots\ \ \sigma_Y(F)]$, initialization:

$$\mathbf{C}_{\perp[1,:]} = \mathbf{C}_{[1,:]}/\alpha_1 \text{ and } \mathbf{C}_{\perp[2,:]} = \mathbf{C}_{\perp[1,:]}D_s/\alpha_2$$

recursion: for $n = 3$ to r

$$\mathbf{C}_{\perp[n,:]} = \mathbf{C}_{\perp[n-1,:]}D_s/\alpha_n + \alpha_{n-1}/\alpha_n\mathbf{C}_{\perp[n-2,:]}$$

1e. Elimination of the input term: calculate the QR factorization of $Z_\perp^T$, $Z_\perp^T = QR$ or,

$$Z_\perp = \begin{bmatrix} R_{11}^T & 0 \\ R_{12}^T & R_{22}^T \end{bmatrix} \begin{bmatrix} Q_1^T \\ Q_2^T \end{bmatrix}$$

where R_{ij} are r by r blocks of the upper triangular matrix R.

1f. Reduction of the noise influence: calculate the singular value decomposition of $C_{\mathbf{Y}_\perp}^{-1/2}R_{22}^T$,

$$C_{\mathbf{Y}_\perp}^{-1/2}R_{22}^T = U\Sigma V^T$$

where $C_{\mathbf{Y}_\perp}^{1/2}$ is a square root of $C_{\mathbf{Y}_\perp}$ (see Section 13.4.4), and estimate $O_{r\perp}$ as

$$\hat{O}_{r\perp} = C_{\mathbf{Y}_\perp}^{1/2}U_{[:,\,1:n_a]}$$

2. Estimate A and C given the estimate $\hat{O}_{r\perp}$: solve the generalized shift property in least squares sense and select the first row of $\hat{O}_{r\perp}$

$$\hat{A} = [D_1\hat{O}_{r\perp[1:r-1,\,:]}]^+[\hat{O}_{r\perp[2:r,\,:]} - b] \text{ and } \hat{C} = \alpha_1\hat{O}_{r\perp[1,\,:]}$$

with $+$ the Moore-Penrose pseudoinverse (see Section 13.5) and

$$b = \begin{bmatrix} 0 \\ D_2\hat{O}_{r\perp[1:r-2,\,:]} \end{bmatrix}, \begin{aligned} D_1 &= \text{diag}(1/\alpha_2, 1/\alpha_3, \ldots, 1/\alpha_r) \\ D_2 &= \text{diag}(\alpha_2/\alpha_3, \alpha_3/\alpha_4, \ldots, \alpha_{r-1}/\alpha_r) \end{aligned}$$

3. Estimate B and D, given the estimates $\hat{A}$ and $\hat{C}$: minimize (7-120) w.r.t. B and D with $W(s_k) = 1/\sigma_Y(k)$.
4. Denormalization of the estimates: multiply $\hat{A}$ and $\hat{C}$ by ω_{scale}.

Proof. See Appendix 7.S. □

Algorithm 7.25 differs in three ways from that described in Van Overschee and De Moor (1996a). First, the recursions in steps 1b and 1c are performed on rows with unit 2-norm. Next, the orthogonal projection is calculated via a QR factorization (see step 1e of Algorithm 7.25). Finally, one additional equation is used to estimate A (see step 2 of Algorithm 7.25). While the first two modifications improve the numerical stability of the algorithm, the third modification decreases the estimation error.

7.14.3 Stochastic Properties

The persistence-of-excitation condition is somewhat different for subspace algorithms compared with algorithms minimizing a cost function of the form (7-11). Therefore, we must add the following assumptions to Section 7.6.1.

Assumption 7.26 (Persistence of Excitation): There exists an F_0 such that for any $F \geq F_0$, ∞ included, $\text{Re}(\mathbf{U}\mathbf{U}^H/F) \geq cI_r$ with $0 < c < \infty$ and c independent of F.

Assumption 7.27 (Identifiability Condition): There exists an F_0 such that for any $F \geq F_0$, ∞ included, $\text{rank}(\mathbf{Y}_0^{\text{re}}\Pi) \geq n_a$.

Note that under Assumptions 7.14 (distinct frequencies) and 7.16 (no model errors), Assumption 7.7 for (7-120) and Assumption 7.27 are fulfilled, if and only if (A, C) and (A, B) in (7-109) are, respectively, observable and controllable (see Appendix 7.R).

Theorem 7.28 (Asymptotic Properties $\hat{\theta}_{\text{SUB}}(Z)$): Consider model (7-109), parameterized in its state space representation, and assume that the input-output data stem from the steady-state response of a system to a periodic excitation, observed during an integer number of periods (time or frequency domain experiment). The estimate $\hat{\theta}_{\text{SUB}}(Z)$ obtained via Algorithm 7.24 or Algorithm 7.25 has the following asymptotic ($F \rightarrow \infty$) properties:

1. *Stochastic convergence:* $\hat{\theta}_{\text{SUB}}(Z)$ converges strongly to the noiseless solution $\theta_{*\text{SUB}}$ (assumptions of Sections 7.6.1 and 7.6.5 and Assumptions 7.26 and 7.27).
2. *Stochastic convergence rate:* $\hat{\theta}_{\text{SUB}}(Z)$ converges in probability at the rate $O_p(F^{-1/2})$ to $\theta_{*\text{SUB}}$ (assumptions of Sections 7.6.2 and 7.6.5 and Assumptions 7.26 and 7.27).
3. *Consistency:* $\hat{\theta}_{\text{SUB}}(Z)$ converges strongly to the true solution θ_0 (assumptions of Sections 7.6.5 and 7.6.6 and Assumptions 7.26 and 7.27).

Proof. See Appendix 7.R and Appendix 7.S. □

Although the subspace (SUB) estimates are strongly consistent ($F \rightarrow \infty$) for any $r \geq n_a + 1$, with r independent of F, the finite sample properties of $\hat{\theta}_{\text{SUB}}(Z)$ strongly depend on the choice of r. For example, values of r close to $n_a + 1$ usually result in poor estimates. An appropriate choice of r is therefore recommended. We propose to choose r such that

$$\sum_{k=1}^{F} \frac{\left|Y(k) - G(\Omega_k, \hat{\theta}_{\text{SUB}}(Z))U(k)\right|^2}{\sigma_Y^2(k)} \tag{7-121}$$

is minimal. This optimization requires an exhaustive search for all $r \geq n_a + 1$ values (the cost function (7-121) is a craggy function of r, with many peaks and dips). In practice, we limit the search to the interval $[1.5n_a, 6n_a]$. However, sometimes it may be necessary to go beyond the upper limit $6n_a$ to find the optimum (see Section 7.15.3, modeling of a synchronous motor). It also turns out that the optimal value of r strongly depends on the plant and the noise characteristics.

The results of the SUB estimates (Algorithm 7.25 with $r = 5$) on the second-order simulation example are shown in Figure 7-4 on page 200. Note that the SUB method estimates five free model parameters while the other methods estimate only three free model parameters. This is due to the fact that the subspace algorithms cannot impose the order of the numerator polynomial. See Table 7-5 on page 238 for an overview of the properties of the SUB estimator.

7.15 ILLUSTRATION AND OVERVIEW OF THE PROPERTIES

The NLS-FRF (7-46), LOG (7-54), GTLS (7-71), ML (7-79), BTLS (7-97), IV (7-107), and SUB (see Section 7.14) estimators perform equally well on the second-order example (see Figure 7-4 on page 200 and Figure 7-6 on page 206). This is due to the very simple nature (low order, low amplitude dynamics, low frequency range, no model errors) of the simulation example. The differences are more apparent in the two simulation examples of this section. Two real measurement examples are also shown to illustrate an aspect that is not shown by the simulation examples: sensitivity to (small) model errors like unmodeled dynamics and nonlinearities. For all the simulation and real measurement examples, the optimal value of r in the SUB Algorithms 7.24 and 7.25 has been selected by an exhaustive search in the interval $[1.5n, 6n]$, except for the modeling of the electrical machine, where the search has been done in the interval $[1.5n, 18n]$.

7.15.1 Simulation Example 1

The simulated plant is a fifth-order continuous-time Butterworth filter with an extra transmission zero at $\omega = 3$ rad/s. The coefficients of the transfer function are given in Table 7-2 and the amplitude and phase characteristics are shown in Figure 7-7. A data set of $F = 100$ equally distributed frequencies is generated in the band $[0.05\text{ Hz}, 5\text{ Hz}]$

TABLE 7-2 Coefficients of the Transfer Function of the Fifth-Order Butterworth Filter with a Transmission Zero

b_0	b_1	b_2			
1	0	1/9			
a_0	a_1	a_2	a_3	a_4	a_5
1	0.449941	0.101223	1.40740e-2	1.20939e-3	5.19623e-5

$$\begin{aligned} Y(k) &= G_0(s_k) + N_Y(k) \\ U(k) &= 1 + N_U(k) \end{aligned} \tag{7-122}$$

with $N_Y(k)$ and $N_U(k)$, $k = 1, 2, \ldots, F$, independent, zero mean, circular complex Gaussian-distributed random variables with variance 2×10^{-6}. One hundred data sets of the type (7-122) are generated. For each set the LS, "full" IWLS, NLS, LOG, GTLS, ML, "full" IQML, and "full" BTLS estimates of model (5-20) with $n_a = 5$ and $n_b = 2$, and the SUB estimate (Algorithm 7.25 with $r = 20$) of model (5-26) with $n_a = 5$ are calculated. Note that the SUB estimate of model (5-26) is equivalent to that of model (5-20) with $n_a = n_b = 5$. All estimators use the constraint $b_1 = 0$ (the zero is forced to lie on the $j\omega$ axis) and $\|\theta\|_2^2 = 1$, except the LS, "full" IWLS, and SUB estimators. The LS and "full" IWLS use $b_0 = 1$ and $b_1 = 0$, and the SUB estimator uses no constraint at all. To perform a bias test, the normalized squared residuals of the mean parameter estimates are calculated for each set of 100 estimates of the model parameters,

$$b = (\langle\hat{\theta}(Z)\rangle - \theta_0)^T(\hat{C}_\theta/R)^+(\langle\hat{\theta}(Z)\rangle - \theta_0) \tag{7-123}$$

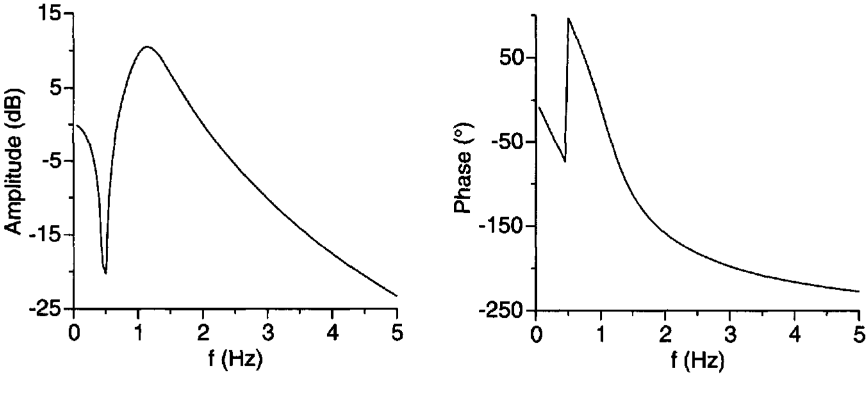


Figure 7-7. Fifth-order Butterworth filter with transmission zero (see Table 7-2): true transfer function.

with θ_0 the true model parameters, $\langle \hat{\theta}(Z) \rangle$ and $\hat{C}_\theta$ the sample mean and sample covariance matrix of the data set,

$$\begin{aligned}\langle \hat{\theta}(Z) \rangle &= \frac{1}{R}\sum_{r=1}^{R} \hat{\theta}^{[r]}(Z)\\ \hat{C}_\theta &= \frac{1}{R-1}\sum_{r=1}^{R} (\hat{\theta}^{[r]}(Z) - \langle \hat{\theta}(Z) \rangle)(\hat{\theta}^{[r]}(Z) - \langle \hat{\theta}(Z) \rangle)^T\end{aligned} \tag{7-124}$$

and R the number of elements in the data set ($R = 100$). If $\hat{\theta}(Z)$ is an unbiased Gaussian estimate, then b is a Hotelling T^2-statistic that is

$$n_\theta \frac{(R-1)}{(R-n_\theta)} F(n_\theta, R-n_\theta) \tag{7-125}$$

distributed with $n_\theta = 7$, the number of free model parameters, and $R = 100$ (see Section 14.3). Because $\hat{\theta}(Z)$ is asymptotically ($F \rightarrow \infty$) normally distributed (see Theorem 7.21, property 4), it is possible to perform a bias test on the estimates with a given confidence level. For example, the 95% percentile of (7-125) equals 15.7 for $n_\theta = 7$ (all estimates except SUB) and $R = 100$ and 23.2 for $n_\theta = 11$ (SUB) and $R = 100$. Hence, with 95% confidence, the estimates are unbiased if $b \le 15.7$ ($b \le 23.2$ for SUB), otherwise they are biased. According to Table 7-3, all the estimates, except the LS, are unbiased.

Using each set of 100 estimates of the model parameters, we can also calculate the relative mean square error of the transfer function estimate

$$\text{RMSE}(\underline{G}(s_k, \hat{\theta}(Z))) \approx \frac{1}{R}\sum_{r=1}^{R} \left|(G(s_k, \hat{\theta}^{[r]}(Z)) - G_0(s_k))/G_0(s_k)\right|^2 \tag{7-126}$$

within an error of 1 dB and compare it with the Cramér-Rao lower bound on the relative transfer function error $(G(s_k, \hat{\theta}(Z)) - G_0(s_k))/G_0(s_k)$. The results are shown in Figure 7-8. It follows that BTLS has ML efficiency and that both estimators reach the Cramér-Rao lower bound. Both LS and GTLS estimators perform equally well; however, the mean square error (MSE) of the LS estimates is due to the bias (see Table 7-3) whereas that of the GTLS estimates is due to the variance (see Table 7-3). The bad performance of the LS and GTLS estimates is due to their inappropriate frequency weighting. The LOG, NLS, and SUB estimators deteriorate somewhat in efficiency w.r.t. the ML and BTLS estimates, but their efficiency is

TABLE 7-3 Bias Test on the Parameter Estimates: Unbiased if $b \le 15.7$ ($b \le 23.2$ for SUB)

Estimator	b (7-123)	Result Bias Test
LS	2.6E4	biased
IWLS	3.9	unbiased
NLS	2.3	unbiased
LOG	5.8	unbiased
GTLS	6.4	unbiased
ML	1.7	unbiased
IQML	1.6	unbiased
BTLS	1.7	unbiased
SUB	12.5	unbiased

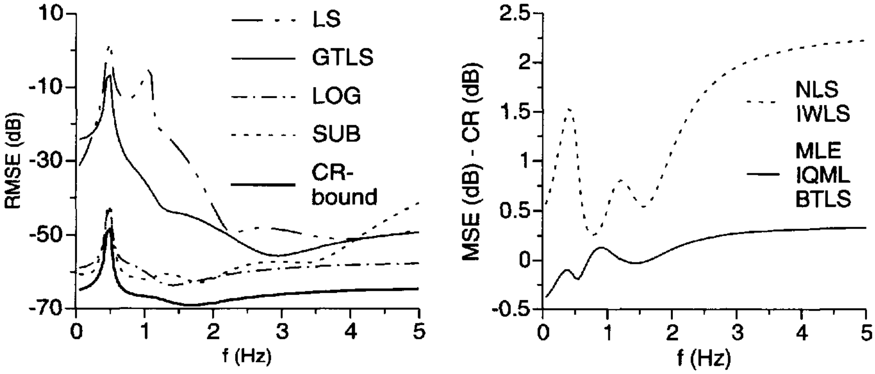


Figure 7-8. Fifth-order simulation example (see Figure 7-7 and Table 7-2): comparison of the (relative) mean square error (R)MSE of the transfer function estimate with the corresponding Cramér-Rao lower bound.

still much better than that of the GTLS method. Because of the high signal-to-noise ratio and the absence of model errors, the IWLS estimator performs as well as the NLS estimator, and the IQML and BTLS estimates coincide with the ML estimates.

7.15.2 Simulation Example 2

The goal of this simulation example is to compare different candidate starting value algorithms: LS (7-32), GTLS (7-71), WLS (7-37), and WGTLS (7-75) with weighting (7-101) and SUB (Algorithm 7.24 with $r = 32$). A sixth-order inverse Chebyshev discrete-time filter with a stopband attenuation of 40 dB and a cutoff frequency of 0.05 is selected as test example (see Figure 7-9a). The discrete time system is excited at $F = 300$ equally spaced frequencies in the band $[0, 0.5]f_s$, with unit amplitude. Next, independent (over the frequency), zero mean, circular complex uniformly distributed noise is added to both the input and output spectra with (co)variances

$$\begin{aligned} &\text{var}(N_Y(k)) = 1\times10^{-6} + 9\times10^{-4}|G_0|^2 \qquad \text{var}(N_U(k)) = 0.161 \\ &\text{covar}(N_Y(k), N_U(k)) = 0 \end{aligned} \tag{7-127}$$

The noisy frequency response function $G(z_k^{-1}) = Y(k)/U(k)$ is shown in Figure 7-9(b). The GTLS, WGTLS, and ML estimates are calculated using the constraint $\|\theta\|_2^2 = 1$, while the LS and WLS estimates use the constraint $a_0 = 1$. No constraint is used in the SUB estimate. Figure 7-9(c) and (d) show the estimated transfer functions in the band $[0, 0.25]f_s$, and Table 7-4 gives the value of the maximum likelihood cost function for the different solutions. Starting from the LS and GTLS solutions (see Figure 7-9(c)) the ML estimate gets stuck in a local minimum (see Figure 7-9(d) and Table 7-4). This is due to the fact that the LS and GTLS solutions place a transmission zero, completely out of the frequency band of interest. The two ML solutions are almost indistinguishable in the band $[0, 0.25]f_s$ but differ somewhat outside that band. Starting from the WLS, WGTLS, and SUB solutions (see Figure 7-9(c)), we find the global minimum of the ML cost function (see Figure 7-9(d) and Table 7-4). Although the WLS solution has a higher ML cost than the GTLS solution, it lies within the attraction basin of the global minimum of the ML estimator. This shows that it may be unsafe to select starting values based on the value of the ML cost.

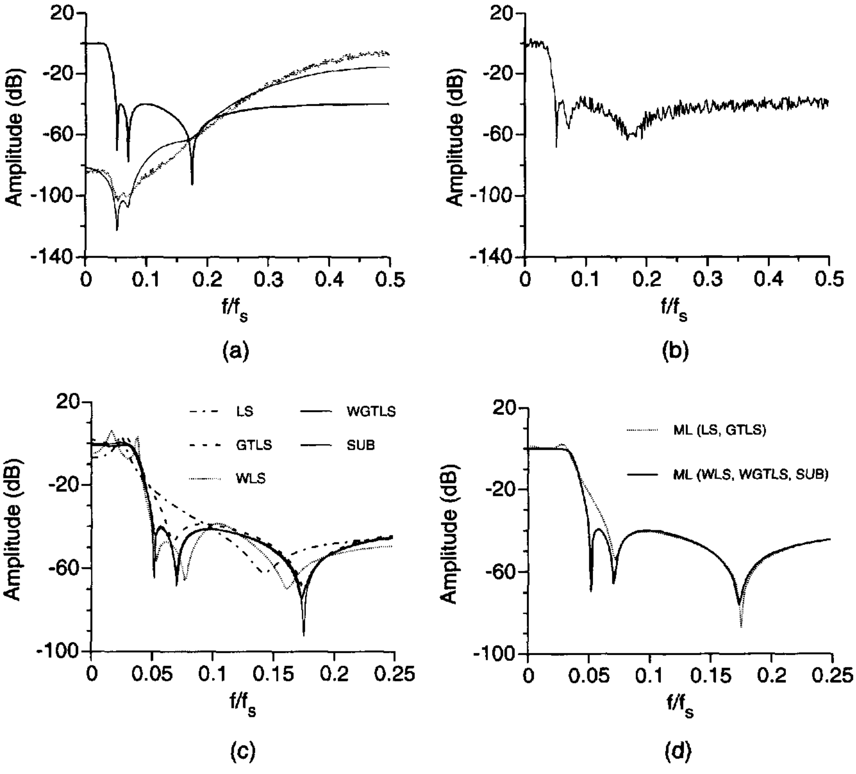


Figure 7-9. Second simulation example. (a) True frequency response function (bold line), maximum likelihood weighting (7-34) evaluated in θ_0 (solid line), and weighting (7-101) (dots); (b) noisy frequency response function; (c) LS, GTLS, WLS, WGTLS, and SUB solutions; (d) ML estimates starting from the solutions shown in (c).

7.15.3 Real Measurement Examples

Two measurement examples that illustrate the properties of the estimators particularly well are shown here. The norm constraint $\|\theta\|_2^2 = 1$ has been used in both examples for the NLS-FRF, LOG, GTLS, ML, IQML, and "full" BTLS estimators, and for the LS and IV estimators $b_0 = 1$ in the q-axis impedance model and $a_0 = 1$ in the flight flutter data model. No constraint is used in the SUB estimates. Because in both examples an improper model $(n_b > n_a)$ is selected, the SUB estimates of model (5-26) are calculated using $1/G(s_k)$ in-

TABLE 7-4 Maximum Likelihood (ML) Cost Function of the Starting Value Algorithms and the Corresponding ML Solution. Least squares: LS and ML (LS); generalized total least squares: GTLS and ML (GTLS); weighted least squares: WLS and ML (WLS); and weighted generalized total least squares: WGTLS and ML (WGTLS).

Estimator	LS	GTLS	WLS	WGTLS	SUB
ML cost function	2140	497	1770	354	333
Estimator (starting value)	ML (LS)	ML (GTLS)	ML (WLS)	ML (WGTLS)	ML (SUB)
ML cost function	436	431	317	317	317

stead of $G(s_k)$. The optimal value of r in the SUB Algorithm 7.25 is 61 and 35 for, respectively, the first and second examples. In the first measurement example the "full" IQML method was used, while in the second example it was necessary to relax the weighting of the IQML method ($r = 0.5$). For each measurement example, two sets of measured input and output spectra were available.

In the first measurement example (see Figures 7-10 and 7-11), the q-axis impedance of a 3.4 MW synchronous motor is modeled with a rational form in s of order $n_b = 4$ over

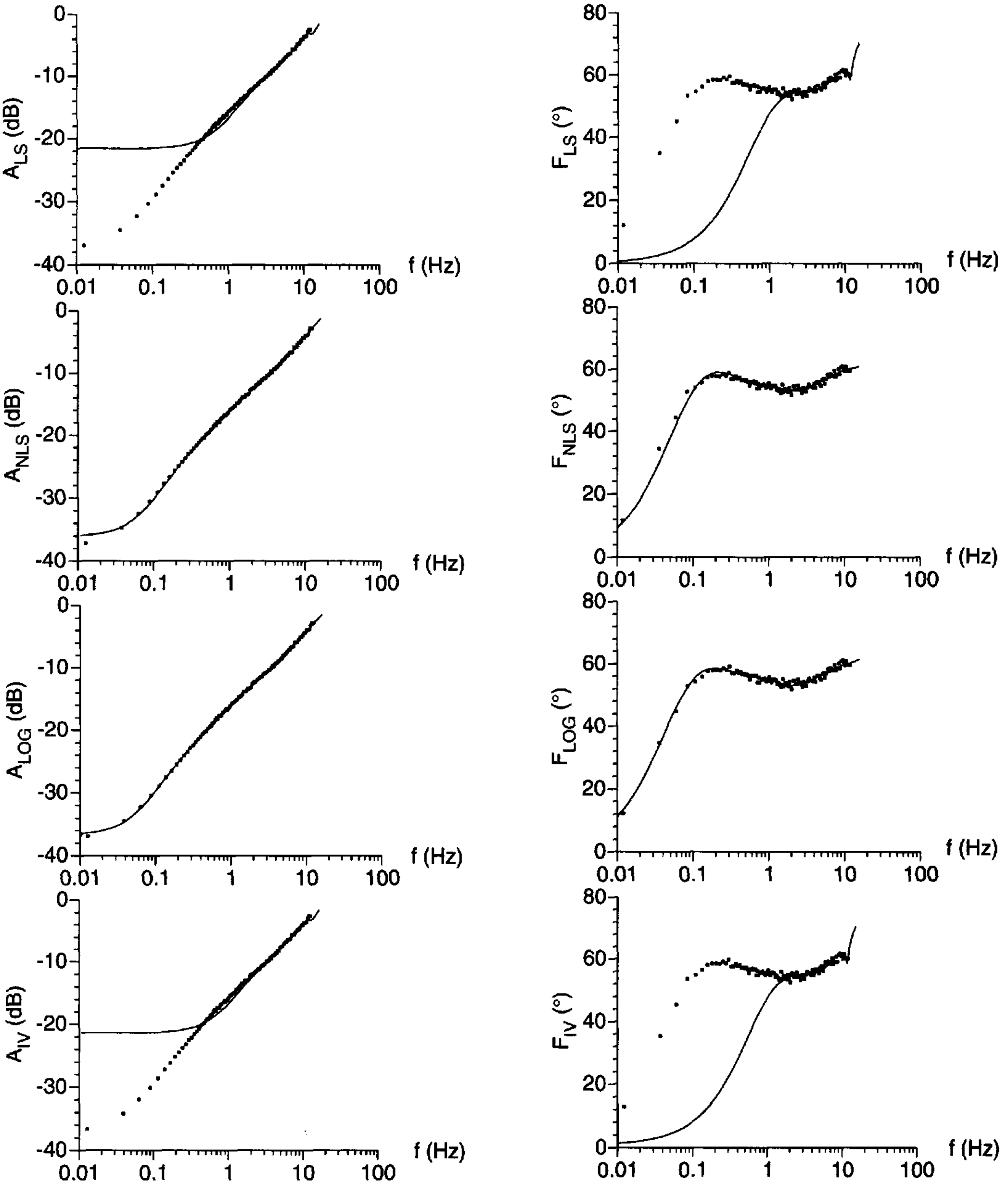


Figure 7-10. Comparison of the measurements (dots) and the estimates requiring no noise information (solid line) of the q-axis impedance of a synchronous machine (model $n_a = 3$, $n_b = 4$). From left to right, amplitude and phase.

$n_a = 3$. The measurements were carried out using a multisine excitation of 1000 A consisting of $F = 100$ frequencies logarithmically spaced in the band $[12\text{ mHz}, 12\text{ Hz}]$. The nonparametric noise model was obtained by analyzing $M = 30$ periods of the input and output signals. Note the particularly large dynamic range in both the amplitude and frequency band. All estimators use the averaged input-output spectra, $X(k) = M^{-1}\sum_{m=1}^{M} X^{[m]}(k)$ with $X = U$ and Y, except the IV estimator, which uses the two sets, $X_1(k) = 2/M\sum_{m=1}^{M/2} X^{[m]}(k)$ and $X_2(k) = 2/M\sum_{m=M/2+1}^{M} X^{[m]}(k)$ with $X = U$ and Y. As expected, the LS, GTLS, and IV estimates are poor in the low-frequency range. The difference between the IQML (norm constraint), NLS-FRF, LOG, and ML estimates is almost indistinguishable. Referring to the large amplitude dynamics, the performance of the NLS-FRF estimator is remarkable. Figure 7-11 also shows the IQML solution using the constraint $b_0 = 1$. It illustrates, again, the influence of the parameter constraint on the estimates for cost functions that are NOT scale invariant.

In the second measurement example (see Figures 7-12 and 7-13), the vibrations of the wings of an airplane are modeled with a rational form in s of order $n_b = 11$ over $n_a = 10$. LMS International (Belgium) have provided us with the experimental data. The measurements were carried out using a burst swept-sine excitation. Three sets of input-output signals of equal length are available. It is impossible to average the three measurements because they are not synchronized. 144 frequencies lie in the frequency band of interest $[4\text{ Hz}, 11\text{ Hz}]$, giving three sets of 144 input-output DFT lines: $\{Y^{[m]}(k), U^{[m]}(k), k = 1, 2, \ldots, 144\}$, $m = 1, 2, 3$. These $F = 3 \times 144$ input-output DFT lines are used for all the estimators except the IV and SUB estimators. The IV method uses one set as instrumental sequence, while the SUB algorithm uses the FRF measurement, averaged over the three sets $\sum_{m=1}^{3} Y^{[m]}(k)/U^{[m]}(k)$, $k = 1, 2, \ldots, 144$. The nonparametric noise model was obtained by analyzing the disturbing noise during the dead time in between consecutive bursts. Although the NLS, LOG, ML, and SUB estimates explain the measurements very well, a careful analysis of the ML cost reveals the presence of small plant model errors (a few tenths of a dB on the amplitude of the transfer function). These small modeling errors account for the better performance of the LS estimates w.r.t. the GTLS estimates. The poor quality of the LS and IV fits is due to the bad weighting of the residuals in their cost functions. Because of its more appropriate weighting of the residuals, the IQML estimator performs better than the GTLS in both measurement examples.

7.15.4 Overview of the Properties

Even if an identification method is based on sound theoretical principles, it can be put into practice only if the normal equations (7-16) or (7-18) are numerically stable and the corresponding cost function can easily be minimized. A global minimization property of the procedure or easy generation of reliable starting values is highly desirable. As constraint independence of the estimates allows the use of overparameterized models (see Chapter 18), it is important that the (equivalent) cost function of the identification method is scale invariant. Consistency and efficiency are important properties to assure that small stochastic deviations in the data do not result in, respectively, large systematic and large stochastic errors on the parameter estimates. Because in practice the true plant model does not often belong to the model set, it is desirable that the estimates are not sensitive to (small) plant modeling errors

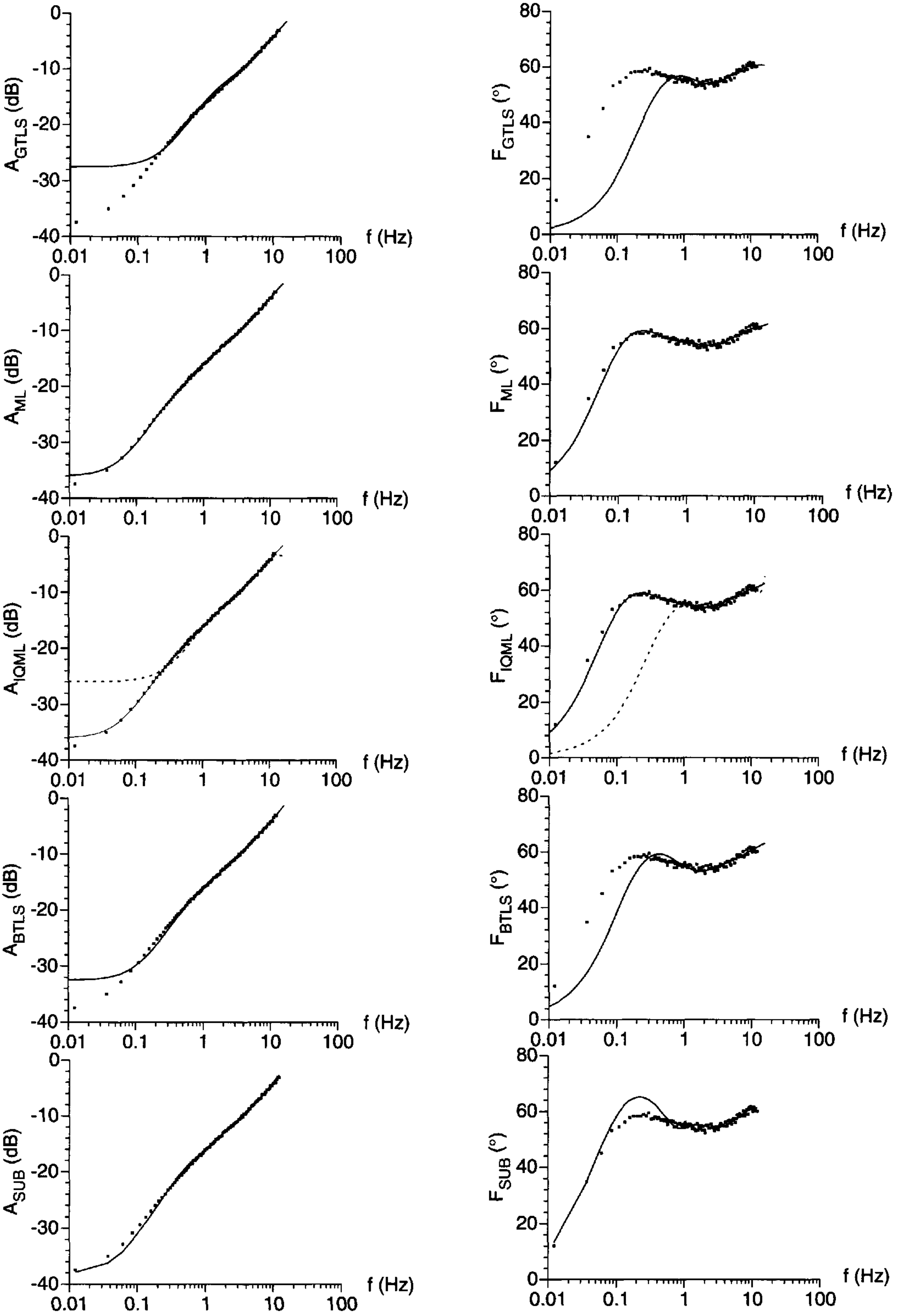


Figure 7-11. Comparison of the measurements (dots) and the estimates requiring noise information (solid line) of the q-axis impedance of a synchronous machine (model $n_a = 3$, $n_b = 4$). From left to right, amplitude and phase. For IQML, the estimates used the constraint $\|\theta\|_2^2 = 1$ (solid line) and $b_0 = 1$ (dashed line).

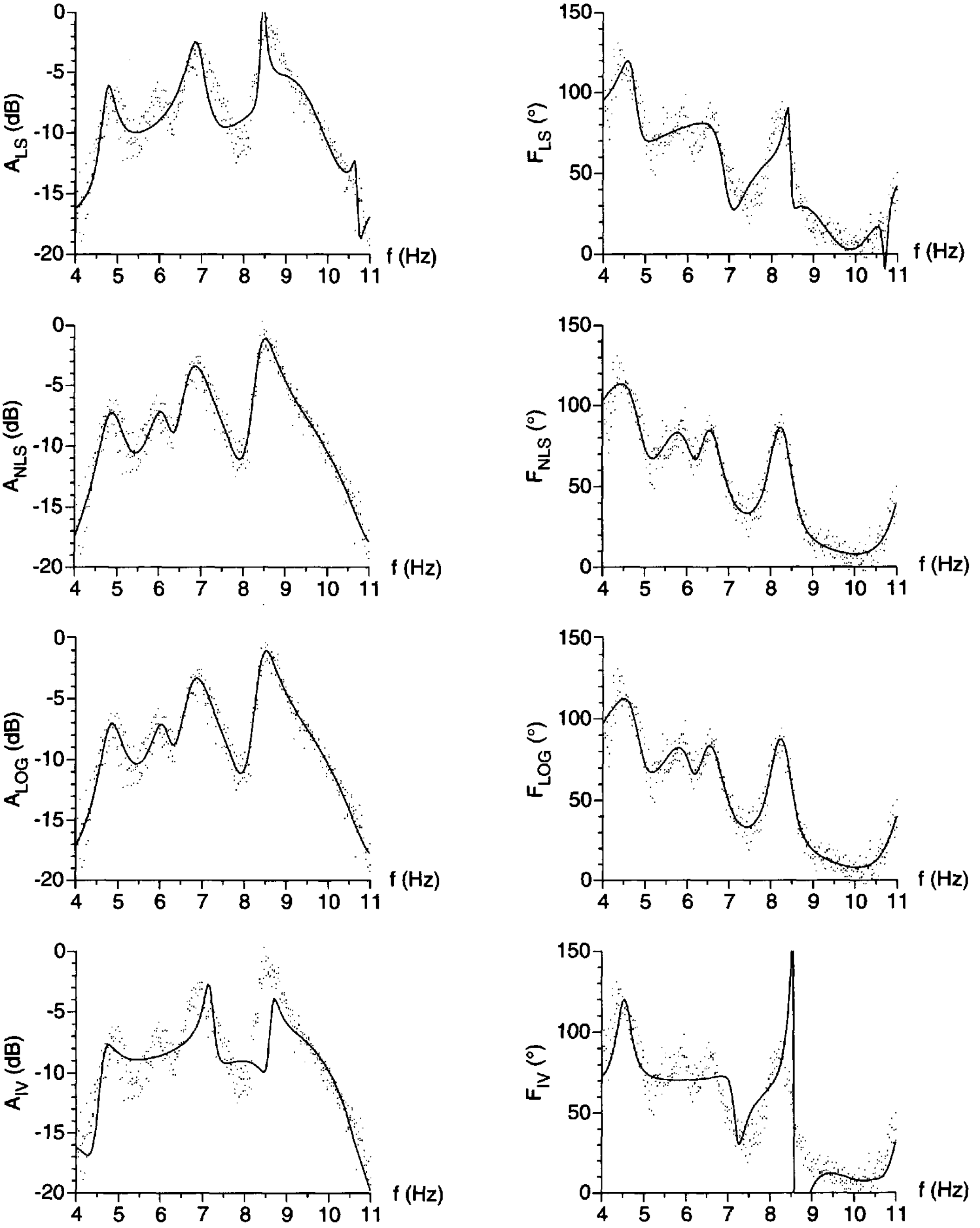


Figure 7-12. Comparison of the measurements (dots) and the estimates requiring no noise information (solid line) of the flight flutter data (model $n_a = 10$, $n_b = 11$). From left to right, amplitude and phase.

and that they converge to the noiseless solution. It is also important that the estimates are not sensitive to noise model errors, for example, wrong noise (co)variances or noncircular complex noise. Table 7-5 on page 238 gives an overview of these properties for some of the estimators discussed in the previous sections.

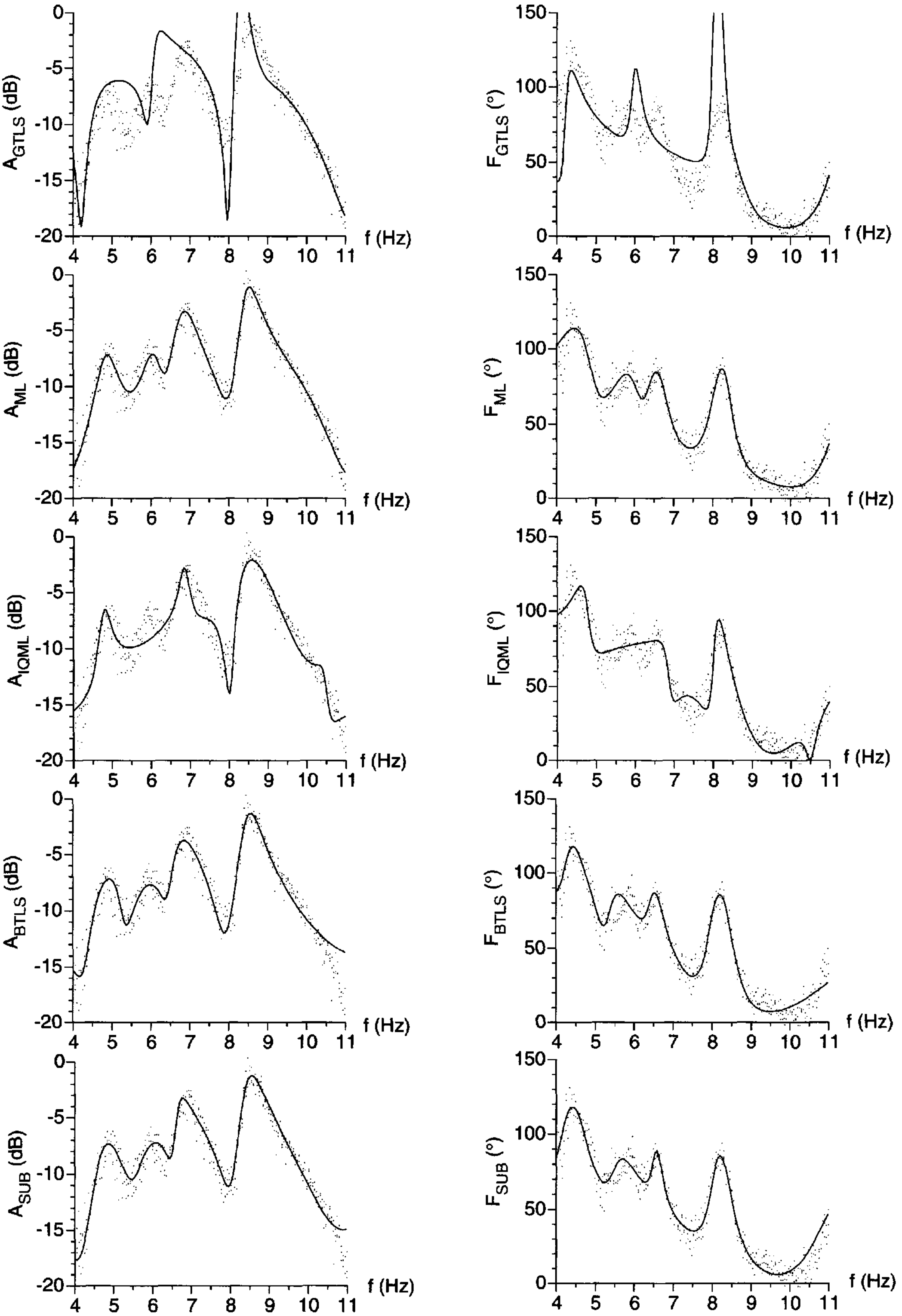


Figure 7-13. Comparison of the measurements (dots) and the estimates requiring noise information (solid line) of the flight flutter data (model $n_a = 10$, $n_b = 11$). From left to right, amplitude and phase.

1. If the noise is circular complex and if the worst case input and output signal-to-noise ratios are larger than 10 dB (see Section 7.9), then the nonlinear least squares estimator, based on frequency response function measurements (NLS-FRF), as well as the logarithmic least squares (LOG) estimator and the subspace algorithms (SUB) are "practically consistent," and when there are model errors they converge to the "practically noiseless solution." For circular complex noise $N_Z(k)$ with even pdf, the biases of the NLS-FRF and LOG estimates are a function of the fourth-order moments of the noise (co)covariances. However, if the noise is not circular complex, then the bias is a function of the second-order moments of the noise (co)variances (see Appendix 7.T). If the input is exactly known, then the NLS-FRF and SUB estimators are consistent or converge to the noiseless solution without any approximation.

2. The maximum likelihood (ML), generalized total least squares (GTLS), bootstrapped total least squares (BTLS), and subspace (SUB) estimates cannot be consistent if the wrong noise (co)variances are used. The resulting bias of the ML, GTLS, and BTLS estimates is proportional to the difference between the true and the actual noise (co)variances (see Appendix 7.T). To have a consistent ML estimate, it is sufficient that the actual noise covariance matrix $\hat{C}_{N_Z(k)}$ equals the true noise covariance matrix $C_{N_Z(k)} = \mathrm{Cov}(N_Z(k))$ within a frequency-dependent scaling factor

$$\hat{C}_{N_Z(k)} = f(k)C_{N_Z(k)} \tag{7-128}$$

(see Appendix 7.T). Note that the consistency proofs of the ML, GTLS, BTLS, and SUB estimates do not require that the noise is circular complex (see Sections 7.10.3, 7.11, 7.12.3, and 7.14.3). Hence, the consistency property of the ML, GTLS, BTLS, and SUB estimates is robust w.r.t. to the circular complex noise assumption.

3. The efficiency of the iterative quadratic maximum likelihood (IQML) estimator strongly depends on the signal-to-noise ratio and on the presence of model errors. Its sensitivity to plant model errors is good if the parameter constraint $\|\theta\|_2^2 = 1$ is used.

4. BTLS converges to the noiseless solution if the limit value θ_* of the starting value $\theta^{(0)}$ is independent of the noise level ν.

5. If the disturbing noise on the instrumental sequences is independent of the disturbing noise on the measurements, then the instrumental variables (IV) estimator is consistent and, when there are plant model errors, converges to the noiseless solution, irrespective of the true noise model. Otherwise, the estimate depends on the correlation between the disturbing noise on the instrumental sequences and the disturbing noise on the measurements.

7.16 HIGH-ORDER SYSTEMS

When identifying higher order transfer function models (7-7), (7-8) (typical $n_a, n_b > 30$) with the rational forms (5-20), (5-37), the condition number of the normal equations (7-18) can become so large that it is impossible to calculate a reliable solution within the available

TABLE 7-5 Overview of the Properties of Some Estimators in the General Case of Input-Output Errors

Estimator	Consistent	Conv. to Noiseless Solution	Efficiency	Prior Noise Knowledge	Global Minim. Procedure	Constraint Dependent	Sense. to Plant Model Errors	Sense. to Noise Model Errors
LS	No	No	Poor	No	Yes	Yes	Medium	—
IWLS	No	No	Poor	No	Yes	Yes	Medium	—
NLS-I/O	No	No	Medium	No	No	No	Very good	—
NLS-FRF	Yes$^{(1)}$	Yes$^{(1)}$	Medium	No	No	No	Very good	Very good$^{(1)}$
LOG	Yes$^{(1)}$	Yes	Good	No	No	No	Very good	Very good$^{(1)}$
GTLS	Yes	Yes	Medium	Yes	Yes	No	Poor	Good$^{(2)}$
ML	Yes	Yes	Excellent	Yes	No	No	Very good	Good$^{(2)}$
IQML	No	No	(3)	Yes	Yes	Yes	Medium/Good$^{(3)}$	?
BTLS	Yes	Yes$^{(4)}$	Very good	Yes	Yes	No	Good	Good$^{(2)}$
IV	Yes	Yes	Poor	No	Yes	Yes	Medium	Very good$^{(5)}$
SUB	Yes$^{(1)}$	Yes	Very good	Yes	Yes	—	Good	Good$^{(2)}$

arithmetic precision. Therefore, to tackle high-order systems the numerator and denominator polynomials of the plant and transient models are expanded in scalar or vector orthogonal polynomials (5-25), (5-40), which are chosen such that they minimize (improve) the condition number. The polynomials are orthogonal w.r.t. to some inner product defined by the cost function and, hence, dependent on the estimator used. The whole process will be explained for the IWLS estimator (7-37) in Sections 7.16.1 and 7.16.2 and afterward generalized to the other estimators in Section 7.16.3. In what follows, we assume that the parameter constraint $a_{n_p} = 1$ is used. In order to simplify the notations, we will limit the discussion to transfer function model (7-7). Generalization of the results to model (7-8) is straightforward.

7.16.1 Scalar Orthogonal Polynomials

The IWLS cost function (7-37) can be written as the sum of three terms

$$\begin{aligned}&\sum_{k=1}^{F} W^2(\Omega_k, \theta^{(i-1)})|Y(k)|^2|A(\Omega_k, \theta^{(i)})|^2 + \sum_{k=1}^{F} W^2(\Omega_k, \theta^{(i-1)})|U(k)|^2|B(\Omega_k, \theta^{(i)})|^2\\&-2\mathrm{Re}(\sum_{k=1}^{F} W^2(\Omega_k, \theta^{(i-1)})Y(k)\bar{U}(k)A(\Omega_k, \theta^{(i)})\bar{B}(\Omega_k, \theta^{(i)}))\end{aligned} \tag{7-129}$$

Under the identifiability conditions of Theorem 5.9 the Jacobian matrix corresponding to the IWLS cost function (7-37),

$$J_{[k,r]}(\theta^{(i)}, Z) = W(\Omega_k, \theta^{(i-1)})\partial e(\Omega_k, \theta, Z(k))/\partial\theta^{(i)}_{[r]} \tag{7-130}$$

where $e(\Omega_k, \theta, Z(k))$ is given by (7-9) and $\theta^T = [a_0 a_1 \ldots a_{n_a-1} b_0 b_1 \ldots b_{n_b}]$ with $a_{n_a} = 1$, has full rank: $\mathrm{rank}(J_{\mathrm{re}}) = n_a + n_b + 1$. Hence, each of the first two terms in (7-129) defines an inner product of scalar polynomials

$$\begin{aligned}\langle x(\Omega), y(\Omega)\rangle_a &= \mathrm{Re}(\textstyle\sum_{k=1}^{F} W^2(\Omega_k, \theta^{(i-1)})|Y(k)|^2 x(\Omega)\bar{y}(\Omega))\\ \langle t(\Omega), z(\Omega)\rangle_b &= \mathrm{Re}(\textstyle\sum_{k=1}^{F} W^2(\Omega_k, \theta^{(i-1)})|U(k)|^2 t(\Omega)\bar{z}(\Omega))\end{aligned} \tag{7-131}$$

with $x(\Omega)$, $y(\Omega)$ polynomials of order smaller than or equal to n_a and $t(\Omega)$, $z(\Omega)$ polynomials of order smaller than or equal to n_b (proof: see Lemma 13.6). Using definitions (7-131) and

$$A(\Omega, \theta) = \textstyle\sum_{r=0}^{n_a} a_r p_r(\Omega),\ B(\Omega, \theta) = \sum_{r=0}^{n_b} b_r q_r(\Omega) \tag{7-132}$$

the matrix $M = \mathrm{Re}(J^H(\theta^{(i)}, Z)J(\theta^{(i)}, Z))$ of the normal equation (7-17) becomes

$$\begin{aligned}M_{[r+1, s+1]} &= \langle p_s(\Omega), p_r(\Omega)\rangle_a \qquad & r, s = 0, \ldots, n_a\\ M_{[r+n_a+2, s+n_a+2]} &= \langle q_s(\Omega), q_r(\Omega)\rangle_b \qquad & r, s = 0, \ldots, n_b\end{aligned} \tag{7-133}$$

The polynomials $p_r(\Omega)$, $r = 0, 1, \ldots, n_a$, and $q_r(\Omega)$, $r = 0, 1, \ldots, n_b$, are calculated via a Gram-Schmidt orthogonalization (see Section 13.11) using inner products $\langle \, , \, \rangle_a$, and $\langle \, , \, \rangle_b$ respectively. Hence, $\langle p_s(\Omega), p_r(\Omega)\rangle_a = \delta_{sr}$, $\langle q_s(\Omega), q_r(\Omega)\rangle_b = \delta_{sr}$ and

$$\mathrm{Re}(J^H(\theta^{(i)}, Z)J(\theta^{(i)}, Z)) = \begin{bmatrix} I_{n_a} & C_1 \\ C_1^T & I_{n_b+1} \end{bmatrix} \tag{7-134}$$

It can be shown that (7-134) is best conditioned: no other scalar polynomial bases for the numerator and denominator of the rational transfer function model can be found resulting in a better conditioned form $\mathrm{Re}(J^H(\theta^{(i)}, Z)J(\theta^{(i)}, Z))$ (Forsythe and Strauss, 1955; Rolain et al., 1995). The IWLS solution is calculated by not using the special structure (7-134) but by solving the overdetermined set of equations (7-18). Proceeding in this way, the solution is insensitive to a loss of orthogonality among the computed basis polynomials. In Richardson and Formenti (1982), the scalar orthogonal polynomials were applied for the first time to improve the numerical conditioning of the linear least squares method ((7-129) with $W(\Omega_k, \theta^{(i-1)}) = 1$) in modal analysis problems ((7-32) with equation error (7-9) and $\Omega = s$).

7.16.2 Vector Orthogonal Polynomials

It is easy to verify that the IWLS cost function (7-37) can also be written as

$$\sum_{k=1}^{F} W^2(\Omega_k, \theta^{(i-1)}) \begin{bmatrix} A(\Omega_k, \theta^{(i)}) \\ B(\Omega_k, \theta^{(i)}) \end{bmatrix}^H \begin{bmatrix} |Y(k)|^2 & -\bar{Y}(k)U(k) \\ -Y(k)\bar{U}(k) & |U(k)|^2 \end{bmatrix} \begin{bmatrix} A(\Omega_k, \theta^{(i)}) \\ B(\Omega_k, \theta^{(i)}) \end{bmatrix} \tag{7-135}$$

Under the identifiability conditions of Theorem 5.9, the Jacobian matrix corresponding to the IWLS cost function (7-37) is

$$J_{[k,r]}(\theta^{(i)}, Z) = W(\Omega_k, \theta^{(i-1)})\partial e(\Omega_k, \theta, Z(k))/\partial\theta^{(i)}_{[r]} \tag{7-136}$$

where $e(\Omega_k, \theta, Z(k))$ is given by (7-9) and $\theta^T = [a_0 a_1 \ldots a_{n_a+n_b}]$ with $a_{n_a+n_b+1} = 1$ has full rank: $\mathrm{rank}(J_{\mathrm{re}}) = n_a + n_b + 1$. Hence, the cost function (7-135) defines an inner product of vector polynomials

$$\langle x(\Omega), y(\Omega)\rangle = \mathrm{Re}\Big(\sum_{k=1}^{F} W^2(\Omega_k, \theta^{(i-1)})y^H(\Omega_k) \begin{bmatrix} |Y(k)|^2 & -\bar{Y}(k)U(k) \\ -Y(k)\bar{U}(k) & |U(k)|^2 \end{bmatrix} x(\Omega)\Big) \tag{7-137}$$

where $x(\Omega)$ and $y(\Omega)$ are 2 by 1 vector polynomials of order smaller than or equal to $n_a + n_b + 1$ (proof: see Lemma 13.7). The vector polynomials $x_r^T(\Omega) = [p_r(\Omega) \; q_r(\Omega)]$, $r = 0, 1, \ldots, n_a + n_b + 1$, are calculated via a Gram-Schmidt orthogonalization (see Section 13.11) using inner product (7-137). Hence we have

$$\langle x_s(\Omega), x_r(\Omega)\rangle = \delta_{sr} \tag{7-138}$$

and

$$\mathrm{Re}(J^H(\theta^{(i)}, Z)J(\theta^{(i)}, Z)) = I_{n_a+n_b+1} \tag{7-139}$$

Clearly, (7-139) has the smallest possible condition number: $\kappa(\mathrm{Re}(J^H J)) = 1$. The IWLS solution is given by

$$G(\Omega, \hat{\theta}_{\mathrm{IWLS}}(Z)) = q_{n_a+n_b+1}(\Omega)/p_{n_a+n_b+1}(\Omega) \tag{7-140}$$

(see Appendix 7.U). Note that solution (7-140) explicitly makes use of the orthogonality of the polynomial basis and, hence, is sensitive to a loss of orthogonality among the computed basis polynomials. A numerically stable and time-efficient implementation of the orthogonalization procedure can be found in Van Barel and Bultheel (1994) for discrete time models $(\Omega = z^{-1})$.

7.16.3 Application to the Estimators

Because the LS (7-32), IWLS (7-36), IQML (7-96), and WLS (7-101) estimators are special cases of the general IWLS estimator (7-37), the calculation of the orthogonal polynomials follows the same lines as in Sections 7.16.1 and 7.16.2. They are chosen such that they minimize the condition number of the normal equation (7-18).

For all the estimators whose (equivalent) cost function is a nonquadratic function of the model parameters, it is impossible to generate, in each iteration step of the Newton-Gauss procedure, a set of orthogonal polynomials that minimize the condition number of the normal equation (7-18). Indeed, the big difference between the IWLS solution and the nonlinear minimization scheme is that the former generates a solution in each iteration step, eventually based on an initial guess, and the latter generates an increment w.r.t. the initial guess. Because the initial guess and the increment should be calculated in the same polynomial basis, it is impossible to minimize the condition number of (7-18). However, it is still possible to make the solution well conditioned. This is done in the following way.

As already emphasized in Section 7.10, the solution of the total least squares estimators, GTLS (7-71), WGTLS (7-75), and BTLS (7-97), is not calculated via the nonlinear minimization scheme (7-18), but via the GSVD of the matrix pair $(W_{\mathrm{Re}}J_{\mathrm{re}}(Z), C)$. Compared with the IWLS cost function (7-37), $J(Z)$ is the Jacobian of the error vector $J(Z) = \partial e(\theta, Z)/\partial\theta^{(i)}$, W is a diagonal matrix with $W_{[k,k]} = W(\Omega_k, \theta^{(i-1)})$, and C is a square root of the column covariance matrix of $W_{\mathrm{Re}}j_{\mathrm{re}}(N_Z)$, with $j(N_Z) = J(Z) - J(Z_0)$. The orthogonal polynomial basis minimizing the condition number of the IWLS estimator is used for the corresponding TLS estimator with $W = I_F$ for GTLS, $W_{[k,k]} = W(\Omega_k, \theta^{(i-1)})$ for WGTLS, and $W_{[k,k]} = \sigma_e^{-r}(\Omega_k, \theta^{(i-1)})$ for BTLS. This choice minimizes the condition number of $W_{\mathrm{Re}}J_{\mathrm{re}}$.

For the NLS-IO (7-42), NLS-FRF (7-46), LOG (7-54), and ML (7-79) estimators we use the orthogonal basis of the starting value algorithm. This choice leads to well, but not best, conditioned normal equations.

7.16.4 Notes

The IWLS solution calculated in the vector orthogonal polynomial basis is given by the highest order vector polynomial (7-140): $a_0 = a_1 = \cdots = a_{n_a+n_b} = 0$ and $a_{n_a+n_b+1} = 1$. If this solution is already of high quality, then the other estimators, (W)GTLS, BTLS, NLS-

IO, NLS-FRF, LOG, and ML, calculated in the same basis, will only marginally perturb the solution: $|a_0|, |a_1|, \ldots, |a_{n_a+n_b}| \ll 1$ and $a_{n_a+n_b+1} = 1$.

Because the inner products (7-131) and (7-137) depend on the measurements, the orthogonal basis depends on the disturbing noise. Therefore, the estimated numerator and denominator orthogonal polynomial coefficients of different experiments cannot be compared. Also, the properties of Theorem 7.21 cannot be applied to the estimated numerator and denominator coefficients. However, because the proof of Theorem 7.21 is independent of the parametrization, its properties are still valid for the invariants of the model, for example, the poles and the zeros.

Evaluating orthogonal polynomials at a particular frequency, or calculating the roots, should always be done through the recursion formula used to construct the orthogonal basis AND NOT via the expansion in powers of Ω, which is numerically ill conditioned for high orders (see Sections 13.11 and 13.12 and Exercise 1.13). To preserve the numerical stability, the calculations for continuous-time systems ($\Omega = s$ or $\sqrt{s}$) should be performed using the normalized frequencies (see Section 7.4).

7.17 SYSTEMS WITH TIME DELAY

The main difficulty of estimating systems with an unknown time delay (plant model (5-29) or (5-30)) is that the corresponding NLS-IO (7-42), NLS-FRF (7-46), LOG (7-54), and ML (7-79) cost functions teem with local minima. A "sufficiently high" quality starting value for the delay is necessary to avoid the local minima. In time domain reflectometry, the time difference between the edges of the excitation pulse and the reflected (transmitted) pulse is a good initial guess of the delay (Pintelon and Van Biesen, 1990). This approach is no longer possible for overlapping pulses, periodic and random excitations. In these cases, a starting value can be obtained via the sample cross-correlation $\hat{R}_{yu}(\tau)$ between the output and the input signals,

$$\hat{\tau} = \arg\max_{\tau}\left|\hat{R}_{yu}(\tau)\right| = \arg\max_{\tau}\left|\frac{1}{N-\tau}\sum_{t=\tau}^{N-1}\tilde{y}(t)\tilde{u}(t-\tau)\right| \tag{7-141}$$

where $\tilde{x}(t) = x(t) - N^{-1}\sum_{t=0}^{N-1}x(t)$ with $x = y, u$, or via the mean slope of the unwrapped phase of the measured frequency response function

$$\hat{\tau} = -\frac{1}{k_2-k_1}\sum_{k=k_1}^{k_2-1}\frac{\angle G(\Omega_{k+1}) - \angle G(\Omega_k)}{\omega_{k+1}-\omega_k} \tag{7-142}$$

where $[\omega_{k_1}, \omega_{k_2}]$ defines the passband of the system. In both cases, the delay is an estimate of the sum of the true delay of the plant minus the slope of the linearized phase of the rational part of the plant. Using the initial guess (7-141) or (7-142) as fixed value in the plant model (5-29) or (5-30), we can calculate starting values for the numerator and denominator coefficients in (5-29) or (5-30), for example, through the IWLS, WGTLS, or SUB estimates (see Sections 7.12.4 and 7.14).

7.18 IDENTIFICATION IN FEEDBACK

Figure 7-14 shows a block diagram of a basic linear feedback experiment. According to the nature of the reference signal $r(t)$, there is a subtle difference between what is considered as the true excitation of the plant and the disturbing noise. If the reference signal is *periodic*,

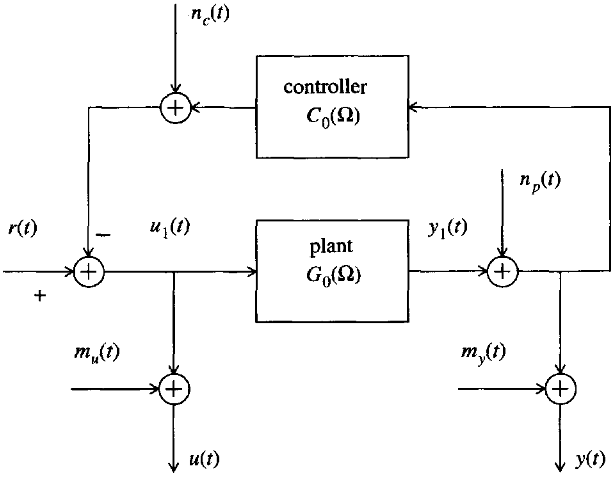


Figure 7-14. Feedback experiment with $r(t)$ the reference signal, $m_u(t)$, $m_y(t)$ the measurement noise sources, $n_p(t)$ the process noise, $n_c(t)$ the controller noise, and $u_1(t)$, $y_1(t)$ the input and output of the plant.

then any deviation from the periodic behavior is considered as noise. The true input-output DFT spectra in (7-1) are then given by

$$\begin{aligned} Y_0(k) &= \frac{G_0(\Omega_k)}{1+G_0(\Omega_k)C_0(\Omega_k)}R(k) \\ U_0(k) &= \frac{1}{1+G_0(\Omega_k)C_0(\Omega_k)}R(k) \end{aligned} \tag{7-143}$$

with $R(k)$ the DFT spectrum of the reference signal, where $G_0(\Omega)$, $C_0(\Omega)$ are the true plant and controller transfer functions, respectively. The frequency domain errors $N_U(k)$ and $N_Y(k)$ in (7-1) are related to the DFT spectra of the disturbing noise sources in Figure 7-14 as

$$\begin{aligned} N_Y(k) &= M_Y(k) + \frac{N_P(k) - G_0(\Omega_k)N_C(k)}{1+G_0(\Omega_k)C_0(\Omega_k)} \\ N_U(k) &= M_U(k) - \frac{N_C(k) + C_0(\Omega_k)N_P(k)}{1+G_0(\Omega_k)C_0(\Omega_k)} \end{aligned} \tag{7-144}$$

Clearly, the disturbances $N_U(k)$ and $N_Y(k)$ are mutually correlated and are independent of the true input $U_0(k)$. Assumption 7.3 or 7.4 is fulfilled and, hence, Theorem 7.21 is valid for periodic excitations and systems in feedback. If the reference signal is *arbitrary*, then the controller noise $n_c(t)$ and the feedback part of the process noise $n_p(t)$ are indistinguishable from the contribution of the reference signal $r(t)$ to the excitation $u_1(t)$. Hence, the true input-output signals are $u_0(t) = u_1(t)$ and $y_0(t) = y_1(t)$. The technical difficulty arising, especially if the noise model is unknown (see Chapter 8), is that the true input signal $u_0(t)$ is correlated with the process noise $n_p(t)$ and, hence, also with the disturbing error at the output.

7.19 MODELING IN THE PRESENCE OF NONLINEAR DISTORTIONS

The goal of a linear identification experiment in the presence of nonlinear distortions can be the identification of the true underlying linear system, or the best linear approximation of the overall system, including the nonlinearities. The first case is useful for physical modeling and, if the system behaves linearly for small inputs, then crest factor optimized excitation signals are most suited for the identification experiment (see Chapter 4). The second case is useful if a linear input-output description is required for a certain class of excitation signals. In this section, we handle the second case. The validity (utility) of the linear model is application dependent and should be established in practice.

The identification starts from measured input-output DFT spectra of a time domain experiment with a random phase multisine (see Figure 7-15). Assuming that an integer number of periods of the steady-state response are observed, we have

$$\begin{aligned} Y(k) &= G_R(s_k)U_0(k) + N_Y(k) \\ U(k) &= U_0(k) + N_U(k) \end{aligned} \qquad (7\text{-}145)$$

with $G_R(s)$ the related linear dynamic system (see Section 5.8), $N_U(k) = M_U(k)$, and $N_Y(k) = N_P(k) + Y_S(k) + M_Y(k)$. The properties of the stochastic nonlinear contributions $Y_S(k)$ are quite similar to those of the measurement and process noise in a time domain experiment (see Sections 5.8 and 7.6). Therefore, Theorem 7.21, where $G_0(s)$ is replaced by $G_R(s)$, remains valid (proof: see Appendix 7.V).

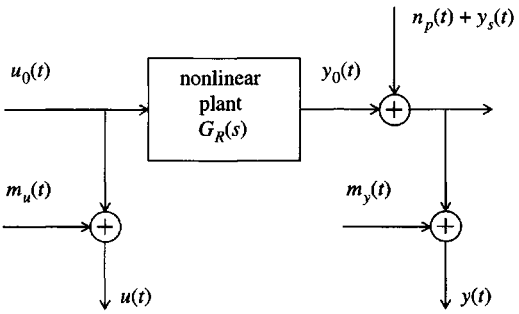


Figure 7-15. Time domain experiment: a random phase multisine is applied to a nonlinear plant $y(t) = G[u(t)]$. The DFT spectra of N observed input-output samples are calculated. $F = O(N)$ DFT frequencies of the input-output spectra are retained. $m_u(t)$ and $m_y(t)$ are the input and output measurement errors, $n_p(t)$ is the process noise, and $y_s(t)$ is the stochastic nonlinear contribution having the same periodicity as the excitation $u_0(t)$.

7.20 MISSING DATA

The form of the output $Y(\Omega, \theta)$, predicted by the model, basically changes if input and/or output samples are missing. Instead of (7-7) and (7-8), we get the following from transfer function models (5-49) and (5-50):

$$Y^{\mathrm{m}}(\Omega_k, \Theta) = G(\Omega_k, \theta)U^{\mathrm{m}}(k) + T(\Omega_k, \theta) + z_k^{-K_u}G(\Omega_k, \theta)I_u(z_k^{-1}, \psi) - z_k^{-K_y}I_y(z_k^{-1}, \psi) \qquad (7\text{-}146)$$

with $\Theta^T = [\theta^T \psi^T]$, ψ the vector containing the M_u missing input samples and M_y missing output samples, $Y^{\mathrm{m}}(\Omega_k, \Theta)$ the output predicted by the model, $U^{\mathrm{m}}(k)$ the DFT spectrum of the missing input data set, and $\Omega = z^{-1}$ or s (see Section 5.3.3). Inspired by the maximum likelihood solution (7-81), we can construct the following weighted nonlinear least squares (WNLS) estimator:

$$V_{\text{WNLS}}(\Theta, Z^{\text{m}}) = \sum_{k=0}^{N-1} \frac{|Y^{\text{m}}(k) - Y^{\text{m}}(\Omega_k, \Theta)|^2}{\sigma_{\hat{Y}}^2(\Omega_k, \theta)} \tag{7-147}$$

with Z^{m} the missing data set and $Y^{\text{m}}(k)$ the DFT spectrum of the missing output data set. $\sigma_{\hat{Y}}^2(\Omega_k, \theta)$ is the variance of the output error (7-44) calculated by using the (co)variances of the complete disturbing noise sequence (no missing samples), for example, $\sigma_{\hat{U}}^2(k) = \text{var}(N_U(k))$ and $\sigma_{\hat{U}}^2(k) \neq \text{var}(N_U^{\text{m}}(k))$. Minimizing (7-147) w.r.t. Θ gives the WNLS estimate $\hat{\Theta}_{\text{WNLS}}(Z^{\text{m}})$ of the plant model parameters θ and the $M_u + M_y$ missing input and/or output samples ψ. To obtain starting values for the plant model parameters θ, the missing data are put equal to zero ($\psi = 0$ in (7-146)). If the number of consecutive missing samples is small, then better starting values for ψ can be obtained via linear interpolation of the known samples. This reduces the risk of being trapped in local minima (cost function (7-147) has more local minima than the problem without missing data).

The properties of $\hat{\theta}_{\text{WNLS}}(Z^{\text{m}})$ can be studied, assuming that the fraction of the missing samples does not increase with the amount of data

$$\frac{M_u + M_y}{2N} = O(N^0) \tag{7-148}$$

To show the consistency of $\hat{\theta}_{\text{WNLS}}(Z^{\text{m}})$, more restrictive assumptions are required than for the problem without missing data. In addition to the assumptions of Section 7.6.6 it is necessary that Assumptions 7.18, 7.19 and condition (7-148) are fulfilled (see Appendix 7.W). Note that the consistency proof relies entirely on the knowledge of the noise (co)variances. If the noise model is unknown and a parametric noise model is identified, then the estimates are no longer consistent. Hence, getting a consistent noise model is the key to the solution of the missing data problem.

More information about the missing output data problem in discrete-time modeling can be found in the literature on time series analysis (see, for example, Little and Rubin, 1987) and system identification (see, for example, Isaksson, 1993; Goodwin and Adams, 1994; Albertos et al., 1999). By considering the missing inputs as unknown parameters, the missing input data problem in discrete-time modeling can be solved by classical prediction error methods (Ljung, 1999).

7.21 MULTIVARIABLE SYSTEMS

Plant models (7-7) and (7-8) remain valid for multivariable systems. $Y(\Omega_k, \theta)$ is then the modeled n_y by 1 output vector, $G(\Omega_k, \theta)$ the n_y by n_u transfer function matrix, $U(k)$ the n_u by 1 vector of the input DFT spectra, and $T(\Omega_k, \theta)$ the n_y by 1 vector of the plant transients.

Following the lines of the scalar case, the multivariable versions of the NLS-IO (7-42), NLS-FRF (7-46), LOG (7-54), and ML (7-79) estimators can be constructed for any of the multivariable parameterizations of $G(\Omega_k, \theta)$ and $T(\Omega_k, \theta)$ described in Section 5.6 (see, for example, Guillaume et al., 1992a, 1996b; Peeters et al., 2000). The numerical minimization of the cost function using the Newton-Gauss scheme (7-18) is somewhat more subtle for the multivariable estimators than for the scalar case (see Guillaume and Pintelon, 1996 for more details).

The IWLS (7-37), WGTLS (7-75), IQML (7-96), BTLS (7-97), and IV (7-107) estimators need a parameterization leading to an equation error that is linear in the model parameters. If the identification starts from measured input-output DFT spectra, then the common denominator model (5-55) and left matrix fraction description (5-56) are suitable. The equation errors (7-9) and (7-10) remain valid with $A(\Omega_k, \theta)$ the denominator polynomial (common denominator model (5-55)) or the n_y by n_y denominator matrix polynomial (left matrix fraction description (5-56)), $Y(k)$ the n_y by 1 vector of the output DFT spectra, $B(\Omega_k, \theta)$ the n_y by n_u numerator matrix polynomial, and $I(\Omega_k, \theta)$ the n_y by 1 vector of the plant equivalent initial conditions. If the identification starts from the measured frequency response matrix (see Section 2.7) then, besides the common denominator model and left matrix fraction description, we can also use the right matrix fraction description (5-57). The corresponding n_y by 1 equation error vectors are

$$e(\Omega_k, \theta, Z(k)) = A(\Omega_k, \theta)G(\Omega_k) - B(\Omega_k, \theta) \tag{7-149}$$

for the common denominator model and left matrix fraction description and

$$e(\Omega_k, \theta, Z(k)) = G(\Omega_k)A(\Omega_k, \theta) - B(\Omega_k, \theta) \tag{7-150}$$

for the right matrix fraction description. Note that constructing an appropriate frequency weighting of the equation errors is somewhat more subtle for the multivariable WGTLS and BTLS estimators than for the scalar case (see Pintelon et al., 1998 for more details).

The subspace algorithms (see Section 7.14) require a multivariable version of model equation (7-119). It is easy to verify that (7-119) remains valid if $\mathbf{Y}$ and $\mathbf{U}$ in (7-115) are replaced by

$$\begin{aligned} \mathbf{Y} &= \left[W_r(1) \otimes Y(1) \;\; W_r(2) \otimes Y(2) \;\; \ldots \;\; W_r(F) \otimes Y(F)\right] \\ \mathbf{U} &= \left[W_r(1) \otimes U(1) \;\; W_r(2) \otimes U(2) \;\; \ldots \;\; W_r(F) \otimes U(F)\right] \end{aligned} \tag{7-151}$$

with $\otimes$ the Kronecker product (see Section 13.7), and similarly for $\mathbf{N_Y}$ and $\mathbf{N_U}$. The multivariable versions of Algorithms 7.24 and 7.25 can be found in McKelvey et al. (1996) and Van Overschee and De Moor (1996a).

7.22 TRANSFER FUNCTION MODELS WITH COMPLEX COEFFICIENTS

Typical applications of transfer function modeling with complex coefficients can be found in nuclear magnetic resonance modeling (see Kumaresan et al., 1990 and Section 5.4) and the identification of rotor bearing systems (see Lee, 1993; Peeters et al., 2000). Because the cost functions of all the estimators for rational transfer function models have been developed without using the fact that θ is real, they remain valid for complex parameters θ. Also, the properties of the estimators remain the same. Indeed, to see this, it is sufficient to replace $\theta \in \mathbb{C}^{n_\theta}$ by $\theta_{\text{re}} \in \mathbb{R}^{2n_\theta}$ and to note that Theorem 7.21 is valid, independent of the particular parameterization chosen.

If $\theta \in \mathbb{C}^{n_\theta}$ is replaced by $\theta_{\text{re}} \in \mathbb{R}^{2n_\theta}$, then all the formulas for the real case apply to the complex case, except that the real part in the definition of the inner products (7-131), (7-137), and (7-248) should be removed. The modification of the inner product changes only the recursion formula used to calculate the orthogonal polynomial basis (see Section 13.11). For example, the normal equation (7-18) becomes

$$J_{\text{re}}(\theta_{\text{re}}^{(i-1)}, Z)\Delta\theta_{\text{re}}^{(i)} = -\varepsilon_{\text{re}}(\theta_{\text{re}}^{(i-1)}, Z) \qquad \text{(7-152)}$$

with $J_{\text{re}}(\theta_{\text{re}}, Z) = \partial\varepsilon_{\text{re}}(\theta_{\text{re}}, Z)/\partial\theta_{\text{re}}$. If the weighted residual $\varepsilon(\theta, Z)$ is an analytic function of θ, then (7-152) is equivalent to

$$J(\theta^{(i-1)}, Z)\Delta\theta^{(i)} = -\varepsilon(\theta^{(i-1)}, Z) \qquad \text{(7-153)}$$

with $J(\theta, Z) = \partial\varepsilon(\theta, Z)/\partial\theta$ (see Appendix 7.X). This is the case for IWLS (7-37), NLS-IO (7-42), NLS-FRF (7-46), LOG (7-54), IQML (7-96), and IV (7-107) estimators. This is not true for the ML estimator because $\sigma_e(\Omega, \theta)$ in (7-83) and $\sigma_Y(\Omega, \theta)$ in (7-84) are not analytic functions of θ. Hence, the ML normal equation (7-152) cannot be simplified to (7-153).

The solution of the WGTLS (7-75) and BTLS (7-97) estimators is calculated as the right generalized singular vector corresponding to the smallest generalized singular value of the matrix pair $(W_{\text{Re}}J_{\text{re}}(Z), C)$, with W the diagonal weighting matrix (7-73), $J(Z) = \partial e(\theta, Z)/\partial\theta$, and C a square root of the column covariance matrix of $W_{\text{Re}}j_{\text{re}}(N_Z)$, with $j(N_Z) = J(Z) - J(Z_0)$. Because $e(\theta, Z)$ is an analytic function of θ, the solution can also be calculated as the right generalized singular vector corresponding to the smallest generalized singular value of the matrix pair $(WJ(Z), C_c)$ with C_c a square root of the column covariance matrix of $Wj(N_Z)$ (see Appendix 7.Y).

7.23 EXERCISES

7.1. Show that the signals defined in Assumption 7.11 are quasi-stationary (7-21) (hint: assume that an integer number of periods is observed for periodic signals and use $u(t) = \text{IDFT}(U(k))$ with $U(N-k) = \overline{U}(k)$).

7.2. Consider the setup shown in Figure 7-14 with $r(t)$ a periodic signal and $m_u(t) = 0$, $m_y(t) = 0$ (no measurement errors). Show that Assumption 7.20(iii) is fulfilled (hint: use Eq. (7-144)).

7.3. Show that the contribution of the disturbing noise to the expected value of the linear least squares cost function is given by (7-33) (hint: use (7-12) with $\Delta(\Omega_k, \theta, N_Z(k)) = A(\Omega_k, \theta)N_Y(k) - B(\Omega_k, \theta)N_U(k)$).

7.4. Show that the difference between the Jacobians of the nonlinear least squares cost (7-31) and the iterative weighted least squares cost (7-36) is given by

$$(J_{\text{NLS}}(\theta^{(i-1)}, Z) - J_{\text{IWLS}}(\theta^{(i-1)}, Z))_{[k,l]} = -\frac{e(\Omega_k, \theta^{(i-1)}, Z(k))}{|A(\Omega_k, \theta^{(i-1)})|^2}\frac{\partial|A(\Omega_k, \theta^{(i-1)})|}{\partial\theta_{[l]}^{(i-1)}}$$

(hint: compare the ith Newton-Gauss step (7-17) applied to the nonlinear least squares cost (7-31) with the ith normal equation of the IWLS cost function (7-36)).

7.5. Show that the contribution of the disturbing noise to the expected value of the nonlinear least squares cost function is given by (7-43) (hint: use (7-12) with $\Delta(\Omega_k, \theta, N_Z(k)) = N_Y(k) - G(\Omega_k, \theta)N_U(k)$).

7.6. Consider the weighted generalized total least squares estimator (7-75). Show that property 5 of Theorem 7.21 is still valid to

$$V_{*\mathrm{WGTLS}}(\theta) = \frac{\int_{f_{\min}}^{f_{\max}} W^2(\Omega(f))\,\mathcal{E}\{\,|e(\Omega(f), \theta, Z_0(f))|^2\}n(f)df}{\int_{f_{\min}}^{f_{\max}} W^2(\Omega(f))\sigma_e^2(\Omega(f), \theta)n(f)df} + 1$$

(hint: divide the numerator and denominator of (7-75) by F and follow the lines of Appendix 7.E, Section 7.E.3).

7.7. Assume that the errors $N_Y(k)$, $N_U(k)$ on the measured input and output spectra $Y(k)$, $U(k)$ are independent (over k) random variables that are not circular complex distributed ($\mathcal{E}\{N_X^2(k)\} \neq 0$, $X = U, Y$). Show that the Markov estimator of model (5-32) minimizes

$$\frac{1}{2}\sum_{k=1}^{F} \frac{e_{\mathrm{R}}^2(k, \theta)\sigma_{\mathrm{I}}^2(k, \theta) + e_{\mathrm{I}}^2(k, \theta)\sigma_{\mathrm{R}}^2(k, \theta) - 2e_{\mathrm{R}}(k, \theta)e_{\mathrm{I}}(k, \theta)\sigma_{\mathrm{RI}}^2(k, \theta)}{\sigma_{\mathrm{R}}^2(k, \theta)\sigma_{\mathrm{I}}^2(k, \theta) - \sigma_{\mathrm{RI}}^4(k, \theta)}$$

with $e_{\mathrm{R}}(k, \theta) = \mathrm{Re}(e(\Omega_k, \theta, Z))$, $e_{\mathrm{I}}(k, \theta) = \mathrm{Im}(e(\Omega_k, \theta, Z))$, $\sigma_{\mathrm{R}}^2(k, \theta) = \mathrm{var}(\mathrm{Re}(e(\Omega_k, \theta, N_Z)))$, $\sigma_{\mathrm{I}}^2(k, \theta) = \mathrm{var}(\mathrm{Im}(e(\Omega_k, \theta, N_Z)))$, $\sigma_{\mathrm{RI}}^2(k, \theta) = \mathrm{covar}(\mathrm{Re}(e(\Omega_k, \theta, N_Z)), \mathrm{Im}(e(\Omega_k, \theta, N_Z)))$, and $e(\Omega_k, \theta, Z)$ given by (7-9) (hint: use (17-12) with $e_k(\theta, z_k) = (e(\Omega_k, \theta, Z))_{\mathrm{re}}$ and $(\ )_{\mathrm{re}}$ defined in (13-48)).

7.8. Show that the difference between the Jacobians of the ML cost (7-79) and the IQML cost (7-96) is given by

$$(J_{\mathrm{ML}}(\theta^{(i-1)}, Z) - J_{\mathrm{IQML}}(\theta^{(i-1)}, Z))_{[k, l]} = -\frac{e(\Omega_k, \theta^{(i-1)}, Z(k))}{\sigma_e^2(\Omega_k, \theta^{(i-1)})}\frac{\partial\sigma_e(\Omega_k, \theta)}{\partial\theta_{[l]}^{(i-1)}}$$

(hint: compare the ith Newton-Gauss step (7-17) applied to the ML cost (7-79) with the ith normal equation of the IQML cost function (7-96)).

7.9. Assume that the wrong noise (co)variances are used in the GTLS estimator (7-71). Show under Assumptions 7.16 and 7.17 that the bias $\tilde{\theta}(Z_0) - \theta_0$ is given by (7-277) with

$$V_F{}'(\theta_0) = 2\frac{\sum_{k,l=1}^{F} |Y_0(k)|^2|Y_0(l)|^2\mathrm{Re}(\frac{\partial \ln(G(\Omega_k, \theta))}{\partial\theta_0}(V_U(k)(\hat{V}_U(l) + \hat{V}_Y(l)) - \hat{V}_U(k)(V_U(l) + V_Y(l))))}{\sum_{k=1}^{F}|Y_0(k)|^2(\hat{V}_U(k) + \hat{V}_Y(k))}$$

where $V_U(k)$, $V_Y(k)$, $\hat{V}_U(k)$, and $\hat{V}_Y(k)$ are defined in Appendix 7.T. Show that the bias is not zero if the noise covariance matrix used, $\hat{C}_{N_Z(k)}$, satisfies (7-128). Note that the bias expression for the BTLS estimator (7-97) is similar to that of the GTLS estimator. (hint: follow the lines of Appendix 7.T; use $\hat{V}_Y(k) = f(k)V_Y(k)$, $\hat{V}_U(k) = f(k)V_U(k)$ to show that the bias is not zero under condition (7-128)).

7.24 APPENDIXES

Appendix 7.A: A Second-Order Simulation Example

The second-order system $G(s, \theta) = 1/(1 + s + s^2)$ is excited at $F = 100$ frequencies, equally distributed in the band $[0.1, 10]/(2\pi)^2$ Hz. The true input $U_0(k) = 1$ and output $Y_0(k)$ spectra are disturbed by independent, zero mean, circular complex Gaussian noise with variance 0.04: $N_U(k), N_Y(k) \in N^c(0, 0.04)$ (see Section 14.1). Two sets of noisy simulated data $\{U^{[1]}(k),\ Y^{[2]}(k),\ k = 1, 2, \ldots, 100\}$ and $\{U^{[2]}(k),\ Y^{[2]}(k),\ k = 1, 2, \ldots, 100\}$ are generated. The noisy frequency response function $G(s_k)$, shown in Figure 7-1, is the ratio of the averaged output $Y(k) = (Y^{[1]}(k) + Y^{[2]}(k))/2$ to the averaged input $U(k) = (U^{[1]}(k) + U^{[2]}(k))/2$ spectra and is used as simulation data for the least squares

(LS), iterative weighted least squares (IWLS), nonlinear least squares based on frequency response function (NLS-FRF), total least squares (TLS), logarithmic least squares (LOG), and subspace (SUB) estimators. The nonlinear least squares based on input-output data (NLS, NLS-IO), generalized total least squares (GTLS), maximum likelihood (ML), iterative quadratic maximum likelihood (IQML), and bootstrapped total least squares (BTLS) estimators use the averaged input $U(k)$ and output $Y(k)$ spectra as simulation data, and the instrumental variables (IV) method uses the two original noisy data sets separately. The constraint $\|\theta\|_2^2 = 1$ is used to calculate all the estimates except for the SUB algorithm, which uses no constraint, and for the LS, IWLS, IQML, and IV methods, which use $a_0 = 1$.

Appendix 7.B: Signal-to-Noise Ratio of DFT Spectra Measurements of Random Excitations

Consider a random excitation $x(t)$, which is mixing of order 2 (the mixing condition limits the span of dependence of $x(t)$, see Section 14.4) and with $\text{var}(x(t)) > 0$ for any t, infinity included. The variance of its DFT spectrum $X(k) = N^{-1/2}\sum_{t=0}^{N-1} x(t) z_k^{-t}$ equals

$$\text{var}(X(k)) = \frac{1}{N}\sum_{t_1, t_2 = 0}^{N-1} \text{covar}(x(t_1), x(t_2)) z_k^{-t_1 + t_2} \tag{7-154}$$

Because $x(t)$ is mixing of order 2, we have that (see (13-38) with $\text{cum}(x, y) = \text{covar}(x, y)$)

$$\sum_{t_1, t_2 = 0}^{N-1} |\text{covar}(x(t_1), x(t_2))| = O(N)$$

and, hence, (7-154) can be bounded above by

$$\text{var}(X(k)) \leq \frac{1}{N}\sum_{t_1, t_2 = 0}^{N-1} |\text{covar}(x(t_1), x(t_2))| = O(N^0) \tag{7-155}$$

The same reasoning holds for disturbing noise $v(t)$ satisfying the same conditions as $x(t)$, so that the signal-to-noise ratio $[\text{var}(X(k))/\text{var}(V(k))]^{1/2}$ is an $O(N^0)$.

Appendix 7.C: Signal-to-Noise Ratio of DFT Spectra Measurements of Periodic Excitations

Consider a multisine $x(t)$ with finite power, $N^{-1}\sum_{t=0}^{N-1} x^2(tT_s) = O(N^0)$, and consisting of F harmonically related frequencies

$$x(t) = \sum_{r=1}^{F} A_r \sin(2\pi m_r f_0 t + \varphi_k) \tag{7-156}$$

where $m_r \in \mathbb{N}$, $r = 1, 2, \ldots, F$ and $m_1 < m_2 < \cdots < m_F$. Assume that we observe the multisine during an integer number of periods, $NT_s/T_0 = Nf_0/f_s \in \mathbb{N}$, and that we respect the Nyquist condition $m_F f_0 < f_s/2$. Assume, furthermore, that the disturbing noise $v(t)$ is mixing of order 2 with $\text{var}(v(t)) > 0$ for any t, infinity included. Here, we handle two cases: F is independent of N, $F = O(N^0)$, and F increases with N, $F = O(N)$.

7.C.1 F Is Independent of N. Because F is independent of N and $x(t)$ has finite power, we have $A_r = O(N^0)$, $r = 1, 2, \ldots, F$. Using $\sin(x) = (e^{jx} - e^{-jx})/(2j)$,

$\sum_{t=0}^{N-1} x^t = (1-x^N)/(1-x)$, and $z_k^N = 1$, the DFT spectrum $X(k) = N^{-1/2}\sum_{t=0}^{N-1} x(tT_s)z_k^{-t}$ equals

$$X(k) = \begin{cases} \dfrac{\sqrt{N}}{2j}A_r e^{j\varphi_r} & k\dfrac{f_s}{N} = m_r f_0 \\ 0 & k\dfrac{f_s}{N} \neq m_r f_0 \end{cases} \tag{7-157}$$

for $k = 0, 1, \ldots, N/2$ with $X(N-k) = \bar{X}(k)$ for $k = N/2+1, \ldots, N-1$. Because $A_r = O(N^0)$, we have $X(k) = O(N^{1/2})$ at the excited DFT frequencies. For disturbing noise $v(t)$, which is mixing of order 2, we have $\operatorname{var}(V(k)) = O(N^0)$ (see Appendix 7.B) so that the signal-to-noise ratio $|X(k)|/\sqrt{\operatorname{var}(V(k))}$ of the multisine at the excited DFT frequencies increases as $O(N^{1/2})$.

7.C.2 F Increases as O(N). Because $F = O(N)$ and $x(t)$ has finite power, we have $A_r = O(N^{-1/2})$, $r = 1, 2, \ldots, F$, and, hence, $|X(k)| = O(N^0)$ at the excited frequencies (see (7-157)). Combining this result with $\operatorname{var}(V(k)) = O(N^0)$ (see Appendix 7.B) gives $|X(k)|/\sqrt{\operatorname{var}(V(k))} = O(N^0)$.

Appendix 7.D: Asymptotic Behavior Cost Function for a Time Domain Experiment

The cost functions that can handle time domain experiments can be written as

$$V_F(\theta, Z) = \frac{1}{F}\sum_{k \in \mathbb{F}} W^2(\Omega_k, \theta)|Y(k) - G(\Omega_k, \theta)U(k) - T(\Omega_k, \theta)|^2 \tag{7-158}$$

with $\mathbb{F}$ a subset of the DFT frequencies $\{0, 1, \ldots, N-1\}$ and $W(\Omega, \theta)$ the absolute value of a rational function of Ω. We will show that

$$V_F(\theta, Z) = \frac{1}{F}\sum_{k \in \mathbb{F}} W^2(\Omega_k, \theta)|Y(k) - G(\Omega_k, \theta)U(k)|^2 + R(\theta) \tag{7-159}$$

with $R(\theta) = O_p(F^{-1})$ uniformly in Θ_r.

Elaborating (7-158) gives (7-159) with

$$\begin{aligned} R(\theta) = {} & \frac{1}{F}\sum_{k \in \mathbb{F}} W^2(\Omega_k, \theta)|T(\Omega_k, \theta)|^2 \\ & - 2\operatorname{Re}\Big(\frac{1}{F}\sum_{k \in \mathbb{F}} W^2(\Omega_k, \theta)(Y(k) - G(\Omega_k, \theta)U(k))\bar{T}(\Omega_k, \theta)\Big) \end{aligned} \tag{7-160}$$

The first term in (7-160) is an $O_p(F^{-1})$ because the numerator coefficients of $T(\Omega, \theta)$ tend to zero as $O_p(F^{-1/2})$ (Lemma 5.7). The second term in (7-160) can be written as the sum of two terms of the form

$$\mathrm{Re}\Big(\frac{1}{F^{3/2}}\sum_{k\in\mathbb{F}} X(k)F(\Omega_k, \theta)\Big) \tag{7-161}$$

with $X = Y$ or U and $F(\Omega, \theta)$ a noncausal, rational filter of finite order (independent of F). The additional factor $F^{-1/2}$ stems from the numerator coefficients of $T(\Omega, \theta)$. We now extend the sum in (7-161) to all the DFT frequencies

$$2\mathrm{Re}\Big(\frac{1}{F^{3/2}}\sum_{k\in\mathbb{F}} X(k)F(\Omega_k, \theta)\Big) = \frac{1}{F^{3/2}}\sum\nolimits_{k=0}^{N-1} X_1(k)F(\Omega_k, \theta) \tag{7-162}$$

with

$$X_1(k) = \begin{cases} \bar{X}_1(N-k) = X(k) & k \in \mathbb{F} \\ 0 & \text{elsewhere} \end{cases} \tag{7-163}$$

Using $X_f(k) = X_1(k)F(\Omega_k, \theta)$ and $F = O(N)$, (7-162) becomes

$$\frac{1}{F^{3/2}}\sum\nolimits_{k=0}^{N-1} X_f(k) = \frac{\sqrt{N}}{F^{3/2}}x_f(0) = O(F^{-1})x_f(0) \tag{7-164}$$

where $x_f(t) = \mathrm{IDFT}(X_f(k))$ is, within some transient effects, the response of $x_1(t)$ to the noncausal rational filter $F(\Omega, \theta)$. The original noisy signal $x(t)$ consists of the sum of a signal term $x_0(t)$ and a noise term $n_x(t)$ that satisfy respectively Assumptions 7.1 and 7.3. Therefore, the second-order moments of $x_0(t)$ and $n_x(t)$ are uniformly bounded. This is also valid for $x_1(t) = \mathrm{IDFT}(X_1(k))$ because it is obtained by replacing the original DFT spectrum $X(k)$ by zeros at some DFT frequencies (see (7-163) with $\mathbb{F} \subset \{0, 1, \ldots, N-1\}$). Finally, the second-order moments of $x_f(t)$ are uniformly bounded because it is, within some transient effects, the response of $x_1(t)$ to the noncausal rational filter $F(\Omega, \theta)$. Hence, $x_f(0) = O_p(F^0)$ so that $R(\theta) = O_p(F^{-1})$ uniformly in Θ_r, which concludes the proof. Note that using Lemma 14.23 exactly the same reasoning can be followed for the noise transient terms in $N_Y(k)$ and $N_U(k)$. □

Appendix 7.E: Asymptotic Properties of Frequency Domain Estimators with Deterministic Weighting (Theorem 7.21)

If the frequency domain errors of the time and frequency domain experiments were mixing of order four (infinity), then Theorem 7.21 (except property 5) would follow immediately from the results of Chapters 15 and 17 (then the assumptions of Section 7.6 fulfill all the necessary conditions). For the frequency domain experiment the frequency domain errors are mixing of order four (Assumption 7.4) but not of order infinity (moments of order higher than $4+\varepsilon$ do not necessarily exist, see Assumption 7.13). Hence, properties 1, 2, 3, 6, and 7 are valid but properties 4 and 8 still remain to be proved. Because after a DFT the noise is not mixing of order four (infinity) (see Section 14.16), all the properties of Theorem 7.21 remain to be proved for the time domain experiment. Fortunately, the resulting technical difficulties in the proofs can easily be solved using the results of Section 14.16. To understand fully the proofs of this appendix, we advise reading Chapter 15 first.

7.E.1 Stochastic Convergence (Properties 1, 2, 3, 6, and 7). For the frequency domain experiment, properties 1, 2, 3, 6, and 7 follow directly from Theorems 15.6, 15.19, 15.21, 15.11, and 15.28, respectively. To show that the properties are valid for the time domain experiment, it is sufficient to show that the cost function and its higher order derivatives w.r.t. θ still converge strongly (weakly) and uniformly in Θ_r to their expected values when the mixing assumption of order four (Assumption 15.1 with $P = 4$) is replaced by Assumption 7.3. We will prove this for the cost function; the proof for its higher order derivatives w.r.t. θ follows exactly the same lines.

Because $\Delta(\Omega_k, \theta, N_Z(k))$ is linear in $N_Z(k)$ with $\Delta(\Omega_k, \theta, 0) = 0$ (see (7-12)) we have $\Delta(\Omega_k, \theta, N_Z(k)) = M_1(\Omega_k, \theta)N_Z(k)$ with $M_1(\Omega_k, \theta) \in \mathbb{C}^{1\times 2}$. It facilitates rewriting (7-13) as

$$\begin{aligned} V_F(\theta, Z) = {} & V_F(\theta, Z_0) + \frac{1}{F}\sum\nolimits_{k=1}^{F}\left|M_1(\Omega_k, \theta)N_Z(k)\right|^2 \\ & + 2\mathrm{Re}\left(\frac{1}{F}\sum\nolimits_{k=1}^{F} M_2(\Omega_k, \theta)N_Z(k)\right) \end{aligned} \tag{7-165}$$

with $M_2(\Omega_k, \theta) = \bar{\varepsilon}(\Omega_k, \theta, Z_0(k))M_1(\Omega_k, \theta)$. $V_F(\theta, Z_0)$, $M_1(\Omega_k, \theta)$, and $M_2(\Omega_k, \theta)$ are continuous in Θ_r (Assumption 7.6) and, therefore, also uniformly bounded in Θ_r. If the input is random, then $V_F(\theta, Z_0)$ and $M_2(\Omega_k, \theta)$ have uniformly bounded second-order moments (see Assumption 7.1). Hence, under Assumptions 7.1, 7.3, and 7.6, the sums in (7-165),

$$V_F(\theta, Z_0),\ \frac{1}{F}\sum\nolimits_{k=1}^{F}\left|M_1(\Omega_k, \theta)N_Z(k)\right|^2 \text{ and } 2\mathrm{Re}\left(\frac{1}{F}\sum\nolimits_{k=1}^{F} M_2(\Omega_k, \theta)N_Z(k)\right) \tag{7-166}$$

satisfy the conditions of Theorems 14.28 and 14.32 (strong laws of large numbers) and, therefore, converge uniformly w.p. 1 to their expected value at the rate $O_p(F^{-1/2})$ in the compact set Θ_r. We conclude that $V_F(\theta, Z)$ converges uniformly w.p. 1 to its expected value $V_F(\theta)$ at the rate $O_p(F^{-1/2})$ in Θ_r.

For the consistency and the bias (properties 6 and 7), we have to make a distinction between correct models ((7-7) with $\Omega = z^{-1}$, s, $\sqrt{s}$, or $\tanh(\tau_R s)$ and (7-8) with $\Omega = z^{-1}$) and asymptotically correct models ((7-8) with $\Omega = s$). For the correct models we have $\tilde{\theta}(Z_0) = \theta_0$ and for the asymptotically correct model, $\theta_* = \lim_{F\to\infty}\tilde{\theta}(Z_0) = \theta_0$ (Assumption 7.17). Note that for the signals defined in the time domain experiment (Assumption 7.1), the model error $\delta(s_k)$ of (7-8) with $\Omega = s$ converges weakly to zero at the rate $O_p(F^{-1/2})$ (Lemma 5.6, with $F = O(N)$ for a time domain experiment). Hence, $\hat{\theta}(Z)$ is weakly consistent with bias $O(F^{-1/2})$ for model (7-8) with $\Omega = s$.

7.E.2 Asymptotic Normality (Properties 4 and 6). $\sqrt{F}(\hat{\theta}(Z) - \tilde{\theta}(Z_0))$ is asymptotically normally distributed if and only if $\sqrt{F}\delta_\theta(Z)$, and, hence, the vector $\sqrt{F}V_F'^T(\tilde{\theta}(Z_0), Z)$ is asymptotically normally distributed (see (7-25)). Taking the derivative of (7-11) w.r.t. θ gives for $\sqrt{F}V_F'^T(\tilde{\theta}(Z_0), Z)$

$$\begin{aligned} \sqrt{F}V_F'^T(\tilde{\theta}(Z_0), Z) & = \frac{1}{\sqrt{F}}\sum\nolimits_{k=1}^{F} 2\mathrm{Re}\left(\left(\frac{\partial\varepsilon(\Omega_k, \theta, Z(k))}{\partial\tilde{\theta}(Z_0)}\right)^H \varepsilon(\Omega_k, \tilde{\theta}(Z_0), Z(k))\right) \\ & = \frac{1}{\sqrt{F}}\sum\nolimits_{k=1}^{F} x(k) \end{aligned} \tag{7-167}$$

where $x(k)$ depends on zero, first, and second order powers of $N_Z(k)$. We will show here that (7-167) is asymptotically normally distributed for a frequency domain experiment (Assumption 7.13) and for a time domain experiment (Assumption 7.12).

Under Assumption 7.13 (frequency domain experiment) $x(k)$ is independently distributed (over the frequency k) with bounded absolute moments of order $2 + \varepsilon$ and 3 and with $\sum_{k=1}^{F} \text{var}(x(k)) = O(F)$. Hence, $F^{-1/2}\sum_{k=1}^{F} x(k)$ converges in law to a normal distribution at the rate $O(F^{-1/2})$ (see Section 14.10, version 2 of the central limit theorem).

Each entry of $x(k)$ in (7-167) can be written as the sum of terms of the form $X(k)\bar{V}(k)$ where $V(k)$ and $X(k)$ depend either on the disturbing noise or on the true input-output DFT spectra. We now study the sum

$$F^{-1/2}\sum_{k=1}^{F} X(k)\bar{V}(k) \tag{7-168}$$

under Assumptions 7.1, 7.11, and 7.12 for each combination of $X(k)$, $V(k)$ giving a random term $X(k)\bar{V}(k)$. If $V(k)$ and $X(k)$ depend on the DFT spectrum of one of the following signals, filtered iid noise, a normalized nonrandom periodic excitation (see Definition 3.4), a normalized random multisine (see Definition 3.2), or normalized periodic noise (see Definition 3.4), then (7-168) converges in law to a normal distribution at the rate $O(F^{-1/2})$ (proof: apply Theorems 14.29 and 14.33). We conclude that $F^{-1/2}\sum_{k=1}^{F} x(k)$ converges in law to a normal distribution at the rate $O(F^{-1/2})$.

7.E.3 Deterministic Convergence (Property 5). Theorem 15.24 is valid if Assumptions 15.4, 15.22, and 15.23 are fulfilled. We will show that the expected value of the cost function $V_F(\theta)$ converges uniformly in Θ_r to $V_*(\theta)$ at the rate $O(F^{-2})$. The proof for $V_F'(\theta)$ and $V_F''(\theta)$ follows the same lines. The expected value of the cost function equals (see (7-11))

$$V_F(\theta) = \frac{1}{F}\sum_{k=1}^{F} \mathcal{E}\{|e(\Omega_k, \theta, Z(k))|^2\} \tag{7-169}$$

Note that for a time domain experiment the influence of the transient term $T(\Omega_k, \theta)$ in the cost function $V_F(\theta, Z)$ can be neglected in the convergence rate analysis (see Appendix 7.D). Let $\mathcal{E}\{|e(\Omega(f), \theta, Z(f))|^2\}$ be the limit value $(F \to \infty)$ of $\mathcal{E}\{|e(\Omega_k, \theta, Z(k))|^2\}$. We have

$$\mathcal{E}\{|e(\Omega_k, \theta, Z(k))|^2\} = \mathcal{E}\{|e(\Omega(f), \theta, Z(f))|^2\}\big|_{f = f_k} \tag{7-170}$$

for a frequency domain experiment, whereas due to the leakage errors in the DFT spectra of the true input-output signals Z_0 and/or the disturbing noise $N_Z(k)$

$$\mathcal{E}\{|e(\Omega_k, \theta, Z(k))|^2\} = \mathcal{E}\{|e(\Omega(f), \theta, Z(f))|^2\}\big|_{f = f_k} + O(F^{-1}) \tag{7-171}$$

for a time domain experiment (see Section 2.2 and Appendix 5.F with $F = O(N)$). Under Assumptions 7.14 and 7.15 the Riemann sum

$$\frac{1}{F}\sum_{k=1}^{F} \mathcal{E}\{|e(\Omega(f), \theta, Z(f))|^2\}\big|_{f = f_k} \tag{7-172}$$

converges to

$$V_*(\theta) = \int_{f_{\min}}^{f_{\max}} \mathcal{E}\{\,|\varepsilon(\Omega(f), \theta, Z(f))|^2\}n(f)df \tag{7-173}$$

at the rate $O(F^{-2})$ (see Ralston and Rabinowitz, 1984; midpoint rule (4.10-10)). Hence, the convergence rate of $V_F(\theta)$ to $V_*(\theta)$ is an $O(F^{-2})$ for a frequency domain experiment and an $O(F^{-1})$ for a time domain experiment. Under Assumption 7.6 $\mathcal{E}\{\,|\varepsilon(\Omega_k, \theta, Z(k))|^2\}$ is a continuous function of $\theta \in \Theta_r$, and, hence, also uniformly bounded in Θ_r. Therefore, the convergence of $V_F(\theta)$ to $V_*(\theta)$ is uniform in Θ_r. Note that the integrand in (7-173) may be zero in some subintervals of $[f_{\min}, f_{\max}]$. In that case, the integral can be written as the sum of integrals. We conclude that Theorem 15.24 is valid with $K = 2$ for a frequency domain experiment and $K = 1$ for a time domain experiment.

7.E.4 Asymptotic Efficiency (Property 8). In the efficiency study, the covariance matrix of the limiting random variable $\delta_\theta(Z)$ or the truncated estimate $\underline{\hat{\theta}}(Z)$ is compared with the Cramér-Rao lower bound (14-88). As the bias b_θ and the derivative of the bias w.r.t. θ_0 of the truncated estimate $\underline{\hat{\theta}}(Z)$ tend to zero as $O(F^{-1})$ (property 7) and $\text{Cov}(\delta_\theta(Z))$ $(\text{Cov}(\underline{\hat{\theta}}(Z)))$ tends to zero as $O(F^{-1})$ (Assumption 7.9), it is sufficient to compare $\text{Cov}(\delta_\theta(Z))$ or $\text{Cov}(\underline{\hat{\theta}}(Z))$ with $Fi^{-1}(\theta_0)$ in (14-88).

Under Assumptions 7.18 and 7.19 the Fisher information matrix of the model parameters is given by

$$Fi(\theta_0) = FV_F''(\theta_0) = 2\text{Re}\left(\left(\frac{\partial\varepsilon(\theta, Z_0)}{\partial\theta_0}\right)^H\left(\frac{\partial\varepsilon(\theta, Z_0)}{\partial\theta_0}\right)\right) \tag{7-174}$$

(see Section 17.3, formula (17-22)).

Under Assumption 7.18, (7-82) is a Markov estimator. For such an estimator the expression of the covariance matrix (7-26) can be elaborated. The cost function $V(\theta, Z)$ can be written as

$$V(\theta, Z) = \varepsilon^H(\theta, Z)\varepsilon(\theta, Z) = \frac{1}{2}(\sqrt{2}\varepsilon_{\text{re}}(\theta, Z))^T(\sqrt{2}\varepsilon_{\text{re}}(\theta, Z))$$

and similarly for $v(\theta, Z)$ in (7-13). Therefore, Theorem 17.3 is still valid when $\varepsilon(\theta, z)$ and $\Delta(\theta, n_z)$ are replaced by $\sqrt{2}\varepsilon_{\text{re}}(\theta, Z)$ and $\sqrt{2}\Delta_{\text{re}}(\theta, N_Z)$, respectively (compare (7-11) and (7-13) with (17-8)). Applying Lemmas 13.3 and 13.4 to expression (17-32) of Theorem 17.3, we get

$$\begin{aligned}
&\text{Cov}(\sqrt{F}\delta_\theta(Z)) = V_F''^{-1}(\theta_0) + V_F''^{-1}(\theta_0)q_F(\theta_0)V_F''^{-1}(\theta_0)\\
&q_F(\theta_0) = F\mathcal{E}\{\,v_F'^T(\theta_0, N_Z)v_F'(\theta_0, N_Z)\} +\\
&\qquad 2\text{Re}\left(2\text{herm}\left(\mathcal{E}\left\{\left(\frac{\partial\varepsilon(\theta, Z_0)}{\partial\theta_0}\right)^H\right\}\mathcal{E}\{\,\Delta(\theta_0, N_Z)v_F'(\theta_0, N_Z)\}\right)\right)
\end{aligned} \tag{7-175}$$

where $v_F(\theta, N_Z) = \Delta^H(\theta, N_Z)\Delta(\theta, N_Z)/F$ is defined in (7-13) and where the expected values are taken w.r.t. to the disturbing noise N_Z and the observations Z_0.

Under Assumptions 7.18 and 7.19, (7-82) is the maximum likelihood solution. Comparing (7-175) with (7-174) for deterministic Z_0 and Gaussian errors N_Z (Assumptions 7.18

and 7.19) shows that in general the maximum likelihood solution is asymptotically inefficient, $q_F(\theta_0) \neq 0$. Under Assumption 7.20 the rank of the $2F$ by $2F$ matrix C_{N_Z} equals F. Applying Theorem 17.4 with $r = 1$ and $t = 2$ shows, then, that $v_F(\theta, N_Z)$ is independent of θ, so that $q_F(\theta_0) = 0$. We conclude that the maximum likelihood estimate is asymptotically efficient under Assumption 7.20.

To analyze the influence of the noise level v on the inefficiency term in (7-175), we apply quick tool number 6 of Section 7.5. It follows that $V_F''^{-1}(\theta_0) = O(\mathrm{v}^2)$ and $q_F(\theta_0) = O(\mathrm{v}^{-1})$, which gives (7-90). If the pdf of N_Z is even, then the second term in the expression of $q_F(\theta_0)$ is zero, so that $q_F(\theta_0) = O(\mathrm{v}^0)$. □

Appendix 7.F: Asymptotic Properties of Frequency Domain Estimators with Stochastic Weighting (Corollary 7.22)

To understand fully the proof of this appendix, we advise reading Chapters 15 and 16 first. To prove the corollary, it is sufficient to verify that all the conditions of the theorems in Chapter 16 are fulfilled. The cost function $V(\theta, Z)/F$ (7-14), where the stochastic vector $\eta(Z)$ has been replaced by the deterministic vector η, is denoted by $f_F(\theta, \eta, Z)$. Clearly, $f_F(\theta, \eta, Z)$ satisfies the assumptions of Section 7.6. Therefore the cost function $f_F(\theta, \eta, Z)$ and its higher order derivatives w.r.t. θ converge w.p. 1 to their expected values (proof: follow the same lines as in Appendix 7.E). By assumption, the stochastic vector $\eta(Z)$ satisfies all the properties of Theorem 7.21, and the cost function $f_F(\theta, \eta, Z)$ has continuous third-order derivatives w.r.t. $x = [\theta^T \eta^T]^T$. We conclude that all the assumptions of Chapter 16 are fulfilled. From Theorems 16.5 and 16.6, it follows that $\tilde{\theta}(Z_0)$ and θ_* are the minimizers of, respectively, $V_F(\theta) = \mathcal{E}\{f_F(\theta, \eta_*, Z)\}$ and $V_*(\theta) = \lim_{F \to \infty} \mathcal{E}\{f_F(\theta, \eta_*, Z)\}$. The expected value of $\delta_\theta(Z)$ may not exist because the moments of $\eta(Z)$ in $V_F'(\tilde{\theta}(Z_0), Z) = f_F'(\theta, \eta(Z), Z)$ do not necessarily exist. Moreover, if it exists, we will have, in general, $\mathcal{E}\{\delta_\theta(Z)\} \neq 0$. Equation (7-30) follows from Theorem 16.25. Because $\eta(Z)$ satisfies, by assumption, Theorem 7.21, it follows directly that $\tilde{\eta}(Z) - \eta_*$ is given by $\delta_\eta(Z)$ (7-25) in Theorem 16.25. □

Appendix 7.G: Expected Value of an Analytic Function

Consider an analytic function $f(z)$ that has the property $f(0) = 0$. Its Taylor series expansion at the origin is then given by

$$f(z) = \sum_{r=1}^{\infty} \frac{f^{(r)}(0)}{r!} z^r \text{ for any } |z| < R \tag{7-176}$$

with R the convergence radius. For zero mean circular complex errors z (see Assumption 7.18), we have $\mathcal{E}\{z\} = 0$ and $\mathcal{E}\{z^2\} = 0$. If, in addition, the errors have an even pdf, then $\mathcal{E}\{z^{2r+1}\} = 0$. Hence, for uniformly bounded random variables $|z| < R$ the expected value of (7-176) becomes

$$\mathcal{E}\{f(z)\} = \sum_{r=2}^{\infty} \frac{f^{(2r)}(0)}{(2r)!} \mathcal{E}\{z^{2r}\} = O(\mathcal{E}\{z^4\}) \tag{7-177}$$

For circular complex normally distributed z we also have $\mathcal{E}\{z^r\} = 0$ (see Exercise 14.8) and, hence, $\mathcal{E}\{f(z)\} = 0$ if $R = \infty$.

The two functions of interest are $f(z) = 1/(1+z) - 1$ and $f(z) = \ln(1+z)$. Both functions have the property $f(0) = 0$ and the convergence radius of their Taylor series expansion at the origin is $R = 1$. Hence, for circular complex uniformly bounded noise $|z| < 1$ with even pdf we have

$$\mathcal{E}\{1/(1+z) - 1\} = O(\mathcal{E}\{z^4\}) \quad \text{and} \quad \mathcal{E}\{\ln(1+z)\} = O(\mathcal{E}\{z^4\}) \tag{7-178}$$

For unbounded noise, the Taylor series expansion (7-176) diverges for all realizations $|z| > 1$ and (7-178) is no longer valid. However, for sufficiently large signal-to-noise ratios $\mathcal{E}\{|z|^2\} \ll 1$, the probability to "hit" a value $|z| \geq 1$ is small and (7-177) is a very good approximation. For example, for Gaussian noise the right-hand sides of (7-178) would be zero, while the expected value is a very small number, given by

$$\mathcal{E}\{\frac{1}{1+z} - 1\} = -\exp(-1/\sigma_z^2) \text{ and } \mathcal{E}\{\ln(1+z)\} = -\frac{1}{2}\text{Ei}(-1/\sigma_z^2) \tag{7-179}$$

with $\sigma_z^2 = \mathcal{E}\{|z|^2\}$ and Ei( ) the exponential integral function (Guillaume et al., 1992b). Applying to (7-179) to (7-55) with $z = N_Y(k)/Y_0(k)$ and $z = N_U(k)/U_0(k)$ gives (7-56). Using (7-48), the expected value of $N_G(k)/G_0(\Omega_k)$ (7-47) can be written as

$$\mathcal{E}\{N_G(k)/G_0(\Omega_k)\} = \mathcal{E}\{\frac{1}{1+z} - 1\} + \rho(k)\frac{\sigma_U(k)\sigma_Y(k)}{\sigma_V^2(k)}\frac{U_0(k)}{Y_0(k)}\mathcal{E}\{\frac{v}{1+z}\} \tag{7-180}$$

where $z = N_U(k)/U_0(k)$ and $v = N_V(k)/U_0(k)$. Using $z = m + v$ with $m = M(k)/U_0(k)$ we find

$$\frac{v}{1+z} = -\frac{1}{1+z} + \frac{1+v}{1+m+v} = \left(-\frac{1}{1+z} + 1\right) + \left(\frac{1}{1+m/(1+v)} - 1\right) \tag{7-181}$$

The expected value of the second term in (7-181) is further elaborated. Because m and v are mutually independent, we have

$$\begin{aligned} \mathcal{E}\{\frac{1}{1+m/(1+v)} - 1\} &= \mathcal{E}\{\mathcal{E}\{\frac{1}{1+m/(1+v)} - 1|v\}\} \\ &= \mathcal{E}\{-\exp(-|1+v|^2/\sigma_m^2)\} \\ &= -\exp(-1/\sigma_z^2)\sigma_m^2/\sigma_z^2 \end{aligned} \tag{7-182}$$

with $\sigma_z^2 = \sigma_m^2 + \sigma_v^2$. The second equality uses (7-179) and the third equality uses the circular complex normality of v. Collecting (7-179), (7-181), and (7-182) gives

$$\mathcal{E}\{\frac{v}{1+z}\} = \exp(-1/\sigma_z^2)\sigma_v^2/\sigma_z^2 \tag{7-183}$$

Putting (7-183), with $\sigma_z^2 = \sigma_U^2(k)/|U_0(k)|^2$ and $\sigma_v^2 = \sigma_V^2(k)/|U_0(k)|^2$, into (7-180) gives (7-49). □

Appendix 7.H: Total Least Squares Solution—Equivalences (Lemma 7.23)

To simplify the notations, we put $A = W_{\text{Re}}J_{\text{re}}(Z)C^{-1}$ and $x = C\theta$. This facilitates writing (7-61) and (7-63) as $Ax \approx 0$ and

$$\arg\min_{\tilde{A},\,x} \|A - \tilde{A}\|_F^2 \text{ subject to } \tilde{A}x = 0 \text{ and } \|x\|_2^2 = 1 \tag{7-184}$$

respectively. Using the method of the Lagrange multipliers, the constrained minimization problem (7-184) can be reformulated as follows:

$$\arg\min_{\tilde{A},\,x,\,\mu} \text{trace}((A - \tilde{A})(A - \tilde{A})^T) + \mu^T \tilde{A}x \text{ subject to } \|x\|_2^2 = 1 \tag{7-185}$$

where $\mu \in \mathbb{R}^{2F}$ is a Lagrange multiplier vector. Expressing the stationarity of the preceding cost function w.r.t. $\tilde{A}$ yields

$$-2(A - \tilde{A}) + \mu x^T = 0 \text{ or } 2(A - \tilde{A}) = \mu x^T \tag{7-186}$$

(use derivative rule (13-62) of Section 13.9.2 and $\mu^T \tilde{A}x = \text{trace}(\mu^T \tilde{A}x)$). Right multiplication of (7-186) by x, taking into account that $\tilde{A}x = 0$ (stationary cost function w.r.t. μ), gives $\mu = 2Ax/\|x\|_2^2$. Elimination of μ in (7-186) gives the following expression: $A - \tilde{A} = Axx^T/\|x\|_2^2$. Replacing $A - \tilde{A}$ in (7-185) by this expression and taking into account the constraint $\tilde{A}x = 0$ results in

$$\arg\min_{x} \|Ax\|_2^2/\|x\|_2^2 \text{ subject to } \|x\|_2^2 = 1 \tag{7-187}$$

We will show that the constrained minimization problem (7-187) is equivalent to

1. $\arg\min_{x} \|Ax\|_2^2/\|x\|_2^2$
2. $\arg\min_{x} \|Ax\|_2^2$ subject to $\|x\|_2^2 = 1$
3. Finding the eigenvector x corresponding to the smallest eigenvalue λ of the eigenvalue problem $A^T A x = \lambda x$.

In equivalent form number 1 the norm constraint is already included in the cost function, therefore the constraint $\|x\|_2^2 = 1$ in (7-187) can be removed. Equivalence 2 follows directly from equivalence 1. To prove equivalent form number 3, we reformulate equivalence 2 using a Lagrange multiplier λ

$$\arg\min_{x,\,\lambda} \|Ax\|_2^2 - \lambda(\|x\|_2^2 - 1) \tag{7-188}$$

Expressing the stationarity of the preceding cost function w.r.t. x yields

$$x^T A^T A - \lambda x^T = 0 \text{ or } A^T A x = \lambda x \tag{7-189}$$

subject to $\|x\|_2^2 = 1$ (stationarity cost function w.r.t. λ), which is an eigenvalue problem. Putting the solutions (x_k, λ_k), $k = 1, 2, \ldots, n_\theta$ of (7-189) in (7-188) taking into account the constraint $\|x\|_2^2 = 1$ gives

$$\arg\min_{x_k, k} \lambda_k \qquad k = 1, 2, \ldots, n_\theta, \tag{7-190}$$

It shows that the eigenvector x_k corresponding to the smallest eigenvalue λ_k of A minimizes (7-188).

Substituting $A = W_{\text{Re}} J_{\text{re}}(Z) C^{-1}$ and $x = C\theta$ into equivalence 3 of (7-187) gives

$$C^{-T}(W_{\text{Re}} J_{\text{re}}(Z))^T (W_{\text{Re}} J_{\text{re}}(Z))\theta = \lambda C\theta \tag{7-191}$$

Left multiplication of (7-191) by C^T gives equivalence 3 of the lemma. Making the same substitution in equivalences 1 and 2 of (7-187), and taking into account that

$$\|Ax\|_2^2 = \|W_{\text{Re}} J_{\text{re}}(Z)\theta\|_2^2 = \|(WJ(Z)\theta)_{\text{re}}\|_2^2 = \|WJ(Z)\theta\|_2^2 \tag{7-192}$$

(see Lemma 13.4), proves equivalences 1 and 2 of the lemma. □

Appendix 7.I: Expected Value Total Least Squares Cost Function

Because $\|j(N_Z)\theta\|_2^2$ is a real number, we have

$$\mathcal{E}\{\|j(N_Z)\theta\|_2^2\} = \text{Re}(\mathcal{E}\{\theta^T j^H(N_Z) j(N_Z)\theta\}) = \theta^T \mathcal{E}\{\text{Re}(j^H(N_Z) j(N_Z))\}\theta \tag{7-193}$$

with $\text{Re}(j^H(N_Z) j(N_Z)) = j_{\text{re}}^T(N_Z) j_{\text{re}}(N_Z)$ (Lemma 13.4), which proves the first equality in (7-66). Using $j(N_Z)\theta = (J(Z) - J(Z_0))\theta = e(\theta, Z) - e(\theta, Z_0)$ we find

$$\mathcal{E}\{\|j(N_Z)\theta\|_2^2\} = \sum_{k=1}^F \sigma_e^2(\Omega_k, \theta) \tag{7-194}$$

which proves the second equality in (7-66). □

Appendix 7.J: Explicit Form of the Total Least Squares Cost Function

7.J.1 Total Least Squares. Using (7-59), the cost function appearing in equivalence number 2 of Lemma 7.23, with $C = I_{n_\theta}$ and $W = I_F$, can be written as

$$\|J(Z)\theta\|_2^2 = \|e(\theta, Z)\|_2^2 = \sum_{k=1}^F |e(\Omega_k, \theta, Z(k))|^2 \tag{7-195}$$

which is exactly (7-64). For the TLS with weight (7-68) we have

$$\|WJ(Z)\theta^{(i)}\|_2^2 = \|We(\theta^{(i)}, Z)\|_2 = \textstyle\sum_{k=1}^{F} W^2(\Omega_k, \theta^{(i-1)})|e(\Omega_k, \theta^{(i)}, Z(k))|^2 \tag{7-196}$$

which is exactly (7-37).

7.J.2 Generalized Total Least Squares. According to equivalence number 1 of Lemma 7.23, with $W = I_F$, the GTLS cost function equals

$$V_{\text{GTLS}}(\theta, Z) = \|J(Z)\theta\|_2^2 / \|C\theta\|_2^2 \tag{7-197}$$

Using $C^TC = \mathcal{E}\{\text{Re}(j^H(N_Z)j(N_Z))\}$ and (7-194) we can rewrite $\|C\theta\|_2^2$ as

$$\|C\theta\|_2^2 = \mathcal{E}\{\theta^T\text{Re}(j^H(N_Z)j(N_Z))\theta\} = \mathcal{E}\{\|j(N_Z)\theta\|_2^2\} = \textstyle\sum_{k=1}^{F} \sigma_e^2(\Omega_k, \theta) \tag{7-198}$$

Division of (7-195) by (7-198) gives (7-71).

For the WGTLS estimator with left and right weighting (7-73) and (7-74), we have

$$V_{\text{WGTLS}}(\theta, Z) = \|WJ(Z)\theta\|_2^2 / \|C\theta\|_2^2 \tag{7-199}$$

(see Lemma 7.23, equivalence number 1). Following the same lines as for (7-198), we find $\|C\theta\|_2^2 = \mathcal{E}\{\|Wj(N_Z)\theta\|_2^2\}$. Applying (7-196) to $\|WJ(Z)\theta\|_2^2$ and (7-194) to $\mathcal{E}\{\|Wj(N_Z)\theta\|_2^2\}$ gives (7-75) after division. □

Appendix 7.K: Rank of the Column Covariance Matrix

We will show that the rank of the column covariance matrix C_{WJ} (7-74) is rank deficient under Assumption 7.20(i) or 7.20(ii). For the diagonal weighting (7-73), the kth row of $Wj(N_Z)$, with $j(N_Z) = J(Z) - J(Z_0)$, can be written as

$$(Wj(N_Z))_{[k,:]} = W(\Omega_k)N_Z^T(k)S^T(k) \tag{7-200}$$

with $N_Z^T(k) = [N_Y(k)\ N_U(k)]$,

$$S^T(k) = \begin{bmatrix} P_k^T(n_a) & 0 \\ 0 & -P_k^T(n_b) \end{bmatrix} \tag{7-201}$$

and $P_k^T(n) = \begin{bmatrix} 1 & \Omega_k & \dots & \Omega_k^n \end{bmatrix}$. The column covariance matrix C_{WJ} (7-74) then becomes

$$C_{WJ} = \mathcal{E}\{\text{Re}((Wj(N_Z))^H(Wj(N_Z)))\} = \text{Re}(\textstyle\sum_{k=1}^{F} W^2(\Omega_k)S(k)\text{Cov}(N_Z(k))S^H(k)) \tag{7-202}$$

Under Assumption 7.20, $\text{Cov}(N_Z(k))$ has rank one for $k = 1, 2, \dots, F$, so that

$$\text{Cov}(N_Z(k)) = c(k)c^H(k) \tag{7-203}$$

with $c(k) \in \mathbb{C}^2$ (see the SVD expansion (13-18)). Using (7-203), (7-202) can be written as $C_{WJ} = \mathrm{Re}(B^H B) = B_{\mathrm{re}}^T B_{\mathrm{re}}$ where the k th row of B is given by

$$B_{[k,:]} = W(\Omega_k) c^H(k) S^H(k) = W(\Omega_k) \left[\bar{c}_{[1]}(k) P_k^H(n_a) \; -\bar{c}_{[2]}(k) P_k^H(n_b) \right] \tag{7-204}$$

and where the rank of B_{re} determines the rank of C_{WJ}. According to Assumption 7.20, we can distinguish three cases. Under Assumption 7.20(i), there is no input noise and $c_{[1]}(k) \neq 0$, $c_{[2]}(k) = 0$ for any k, so that $\mathrm{rank}(B_{\mathrm{re}}) = n_a + 1$. Under Assumption 7.20(ii), there is no output noise, and $c_{[1]}(k) = 0$, $c_{[2]}(k) \neq 0$ for any k, so that $\mathrm{rank}(B_{\mathrm{re}}) = n_b + 1$. Under Assumption 7.20(iii), the input-output errors are totally correlated and $c_{[1]}(k) \neq 0$, $c_{[2]}(k) \neq 0$ for any k, so that in general B_{re} is of full rank. It is rank deficient only if, in addition, $c_{[1]}(k)/c_{[2]}(k)$ is real and independent of k. This is, for example, the case for totally correlated white noise errors $n_y(t)$, $n_u(t)$ $(\mathrm{Cov}(N_Z(k))$ is then independent of k). □

Appendix 7.L: Calculation of the Gaussian Maximum Likelihood Estimate

7.L.1 Gaussian Log-Likelihood Function. Under Assumptions 7.5, 7.18, and 7.19, the pdf of N_Z is given by

$$f_{N_Z}(N_Z) = \frac{1}{\pi^F \det(C_{N_Z})} \exp(-N_Z^H C_{N_Z}^{-1} N_Z) \tag{7-205}$$

(see (14-14)) with $C_{N_Z} = \mathrm{Cov}(N_Z)$,

$$C_{N_Z} = \mathrm{diag}(\mathrm{Cov}(N_Z(1)), \mathrm{Cov}(N_Z(2)), \ldots, \mathrm{Cov}(N_Z(F)))$$

$$\mathrm{Cov}(N_Z(k)) = \begin{bmatrix} \sigma_Y^2(k) & \sigma_{YU}^2(k) \\ \bar{\sigma}_{YU}^2(k) & \sigma_U^2(k) \end{bmatrix} \tag{7-206}$$

If C_{N_Z} is singular, then $C_{N_Z}^{-1}$ and $\det(C_{N_Z})$ are replaced by $C_{N_Z}^{+}$ and the product of the nonzero eigenvalues of C_{N_Z}, respectively. Replacing N_Z by $Z - Z_{\mathrm{p}}$ in (7-205) and taking the negative of the natural logarithm gives the negative log-likelihood function

$$-\ln f_{N_Z}(Z, Z_{\mathrm{p}}, \theta) = (Z - Z_{\mathrm{p}})^H C_{N_Z}^{+} (Z - Z_{\mathrm{p}}) + c \tag{7-207}$$

with $c = F\ln(\pi) + \ln(\det(C_{N_Z}))$.

7.L.2 Elimination of the Unknown Input-Output DFT Spectra in the Cost Function. Using Lemmas 13.3 and 13.4, Example 13.5 and Exercise 14.4, (7-78) can be written as

$$\frac{1}{2}(Z_{\mathrm{re}} - Z_{\mathrm{pre}})^T C_{N_{Z\mathrm{re}}}^{+} (Z_{\mathrm{re}} - Z_{\mathrm{pre}}) + \lambda_{\mathrm{re}}^T e_{\mathrm{re}}(\theta, Z_{\mathrm{p}}) \tag{7-208}$$

where $C_{N_{Z\mathrm{re}}} = \mathrm{Cov}(N_{Z\mathrm{re}}) = 0.5(C_{N_Z})_{\mathrm{Re}}$ with $C_{N_Z} = \mathrm{Cov}(N_Z)$. Because $e_{\mathrm{re}}(\theta, Z_{\mathrm{p}})$ is linear in Z_{pre}, minimization problem (7-208) is exactly equivalent to (17-6), and, hence, all the results of Section 17.2 are valid. Elimination of Z_{pre} in (7-208) gives (17-8)

$$V_{\mathrm{ML}}(\theta, Z) = \frac{1}{2}e_{\mathrm{re}}^T(\theta, Z)C_{e_{\mathrm{re}}}^{-1}(\theta)e_{\mathrm{re}}(\theta, Z) \tag{7-209}$$

with $C_{e_{\mathrm{re}}}(\theta) = \mathrm{Cov}(e_{\mathrm{re}}(\theta, N_Z))$. Because the noise residual $e(\theta, N_Z)$ is linear in N_Z, it is complex circular

$$\mathcal{E}\{(e(\theta, N_Z) - \mathcal{E}\{e(\theta, N_Z)\})(e(\theta, N_Z) - \mathcal{E}\{e(\theta, N_Z)\})^T\} = 0 \tag{7-210}$$

and $C_{e_{\mathrm{re}}}(\theta) = 0.5(C_e(\theta))_{\mathrm{Re}}$ (see Exercise 14.4). Note that $\mathcal{E}\{e(\theta, N_Z)\} \neq 0$ for model (7-10). Applying the result of 13.5 to (7-209) under Assumption 7.18 gives

$$\begin{aligned} V_{\mathrm{ML}}(\theta, Z) &= e^H(\theta, Z)C_e^{-1}(\theta)e(\theta, Z) \\ C_e(\theta) &= \mathrm{Cov}(e(\theta, N_Z)) = \mathrm{diag}(\sigma_e^2(\Omega_1, \theta), \sigma_e^2(\Omega_2, \theta), \ldots, \sigma_e^2(\Omega_F, \theta)) \end{aligned} \tag{7-211}$$

which is exactly (7-79).

In the derivation of (7-211), we implicitly assumed that no DC (Ω_0) and no Nyquist ($\Omega_{N/2}$) components were present in the data. Indeed, expressions (7-77) and (7-78) are valid only if all the elements of Z are complex, which is not the case for the DC and Nyquist components (real numbers). If DC and Nyquist are present, then under Assumption 7.18 the terms

$$\begin{aligned} &\frac{1}{2}N_Z^T(0)(\mathrm{Cov}(N_Z(0)))^{-1}N_Z^T(0) + \frac{1}{2}N_Z^T(N/2)(\mathrm{Cov}(N_Z(N/2)))^{-1}N_Z^T(N/2) \\ &+ \lambda_0 e(\Omega_0, \theta, Z_{\mathrm{p}}(0)) + \lambda_{N/2}e(\Omega_{N/2}, \theta, Z_{\mathrm{p}}(N/2)) \end{aligned} \tag{7-212}$$

where $N_Z(0)$, $N_Z(N/2)$, λ_0, $\lambda_{N/2}$, $e(\Omega_0, \theta, Z_{\mathrm{p}}(0))$, and $e(\Omega_{N/2}, \theta, Z_{\mathrm{p}}(N/2))$ are real numbers, should be added to (7-78) and (7-208). Their contribution to (7-209) equals

$$\frac{1}{2}\frac{e^2(\Omega_0, \theta, Z(0))}{\sigma_e^2(\Omega_0, \theta)} + \frac{1}{2}\frac{e^2(\Omega_{N/2}, \theta, Z(N/2))}{\sigma_e^2(\Omega_{N/2}, \theta)} \tag{7-213}$$

Because $e(\Omega_0, \theta, Z(0))$ and $e(\Omega_{N/2}, \theta, Z(N/2))$ are real numbers, the terms (7-213) remain unchanged in the transformation from (7-209) to (7-211).

7.L.3 Maximum Likelihood Estimate of the Input and Output DFT Spectra. In the previous section, it was shown that all the results of Section 17.2 are valid for Z_{re}. Hence, the ML estimate $\hat{Z}_{\mathrm{re}}$ of the true DFT spectra $Z_{0\mathrm{re}}$ equals (17-11)

$$C_{N_{Z\mathrm{re}}}C_{N_{Z\mathrm{re}}}^{+}\hat{Z}_{\mathrm{re}} = C_{N_{Z\mathrm{re}}}C_{N_{Z\mathrm{re}}}^{+}Z_{\mathrm{re}} - C_{N_{Z\mathrm{re}}}M_1^T(\hat{\theta})C_{e_{\mathrm{re}}}^{-1}(\hat{\theta})e_{\mathrm{re}}(\hat{\theta}, Z) \tag{7-214}$$

with $\hat{\theta} = \hat{\theta}_{\mathrm{ML}}(Z)$ and

$$M_1(\theta) = \frac{\partial e_{\mathrm{re}}(\theta, Z)}{\partial Z_{\mathrm{re}}} = \begin{bmatrix} \dfrac{\partial \mathrm{Re}(e(\theta, Z))}{\partial \mathrm{Re}(Z)} & \dfrac{\partial \mathrm{Re}(e(\theta, Z))}{\partial \mathrm{Im}(Z)} \\ \dfrac{\partial \mathrm{Im}(e(\theta, Z))}{\partial \mathrm{Re}(Z)} & \dfrac{\partial \mathrm{Im}(e(\theta, Z))}{\partial \mathrm{Im}(Z)} \end{bmatrix} \tag{7-215}$$

Because $e(\theta, Z)$ is an analytic function of Z, it satisfies the Cauchy-Riemann equations

$$\begin{aligned} \frac{\partial \mathrm{Re}(e(\theta, Z))}{\partial \mathrm{Re}(Z)} &= \frac{\partial \mathrm{Im}(e(\theta, Z))}{\partial \mathrm{Im}(Z)} \\ \frac{\partial \mathrm{Re}(e(\theta, Z))}{\partial \mathrm{Im}(Z)} &= -\frac{\partial \mathrm{Im}(e(\theta, Z))}{\partial \mathrm{Re}(Z)} \end{aligned} \tag{7-216}$$

(Henrici, 1974). Applying (7-216) to (7-215) gives

$$M_1(\theta) = \begin{bmatrix} \dfrac{\partial \mathrm{Re}(e(\theta, Z))}{\partial \mathrm{Re}(Z)} & -\dfrac{\partial \mathrm{Im}(e(\theta, Z))}{\partial \mathrm{Re}(Z)} \\ \dfrac{\partial \mathrm{Im}(e(\theta, Z))}{\partial \mathrm{Re}(Z)} & \dfrac{\partial \mathrm{Re}(e(\theta, Z))}{\partial \mathrm{Re}(Z)} \end{bmatrix} = \left(\frac{\partial e(\theta, Z)}{\partial Z}\right)_{\mathrm{Re}} \tag{7-217}$$

Using (7-217), $C_{e_{\mathrm{re}}}(\theta) = 0.5(C_e(\theta))_{\mathrm{Re}}$, $C_{N_{Z\mathrm{re}}} = 0.5(C_{N_Z})_{\mathrm{Re}}$, and Lemmas 13.3 and 13.4, (7-214) becomes

$$\begin{aligned} C_{N_Z} C_{N_Z}^{+} \hat{Z} &= C_{N_Z} C_{N_Z}^{+} Z - C_{N_Z} M^H(\hat{\theta}) C_e^{-1}(\hat{\theta}) e(\hat{\theta}, Z) \\ M(\theta) &= \frac{\partial e(\theta, Z)}{\partial Z} = \mathrm{diag}([A(\Omega_1, \theta) \ \ -B(\Omega_1, \theta)], \ldots, [A(\Omega_F, \theta) \ \ -B(\Omega_F, \theta)]) \end{aligned} \tag{7-218}$$

Putting (7-206) and (7-211) in (7-218) assuming that C_{N_Z} is regular $(C_{N_Z} C_{N_Z}^{+} = I_F)$ gives (7-85). Note that the solution (7-85) remains well defined if C_{N_Z} is singular.

7.L.4 Minimization of the Maximum Likelihood Cost Function. The maximum likelihood cost function (7-82) is minimized using the Newton-Gauss algorithm (7-18). It requires the calculation of the Jacobian matrix $J(\theta, Z) = \partial \varepsilon(\theta, Z)/\partial \theta$, where $\varepsilon(\theta, Z)$ is given by (7-83) or (7-84). In this appendix an explicit expression of the Jacobian matrix $J(\theta, Z)$ is given for rational transfer function models (cost function (7-82) with weighted residual (7-83)). The calculation of $J(\theta, Z)$ for the other transfer function models (partial fraction expansion and state space representation) follows exactly the same lines (cost function (7-82) with weighted residual (7-84)).

Using (7-10), (7-34), and (7-83) we find for $k = 1, 2, \ldots, F$,

$$\begin{aligned} J_{[k, r+1]}(\theta, Z) &= \frac{\partial \varepsilon(\Omega_k, \theta, Z(k))}{\partial a_r} \\ &= \frac{\Omega_k^r Y(k)}{\sigma_e(\Omega_k, \theta)} - \frac{\varepsilon(\Omega_k, \theta, Z(k))}{\sigma_e^2(\Omega_k, \theta)} \mathrm{Re}(\Omega_k^r [\sigma_Y^2(k) \bar{A}(\Omega_k, \theta) - \sigma_{YU}^2(k) \bar{B}(\Omega_k, \theta)]) \end{aligned}$$

with $r = 0, 1, \ldots, n_a$,

$$
\begin{aligned}
J_{[k,n_a+r+2]}(\theta,Z) &= \frac{\partial\varepsilon(\Omega_k,\theta,Z(k))}{\partial b_r}\\
&= \frac{-\Omega_k^r U(k)}{\sigma_e(\Omega_k,\theta)} + \frac{\varepsilon(\Omega_k,\theta,Z(k))}{\sigma_e^2(\Omega_k,\theta)}\mathrm{Re}(\Omega_k^r[\sigma_U^2(k)\bar{B}(\Omega_k,\theta) - \bar{\sigma}_{YU}^2(k)\bar{A}(\Omega_k,\theta)])
\end{aligned}
$$

with $r = 0, 1, \ldots, n_b$, and

$$
J_{[k,n_a+n_b+r+3]}(\theta,Z) = \frac{\partial\varepsilon(\Omega_k,\theta,Z(k))}{\partial i_r} = \frac{-\Omega_k^r}{\sigma_e(\Omega_k,\theta)}
$$

with $r = 0, 1, \ldots, n_i$. If a constraint of the form $\theta_{[j]} = 1$ is used, then the corresponding column in $J(\theta, Z)$ must be eliminated. If the constraint $\|\theta\|_2^2 = 1$ is used, then (7-18) is solved using the pseudoinverse (see Section 13.5) and $\theta^{(i)} = \theta^{(i-1)} + \Delta\theta^{(i)}$ is normalized $(\theta^{(i)} \rightarrow \theta^{(i)}/\|\theta^{(i)}\|_2)$ before making a new iteration step $(i \rightarrow i+1)$.

The Levenberg-Marquardt version of (7-16) with constraint $\theta_{[j]} = 1$ is

$$
(J_{\mathrm{re}}^T(\theta^{(i-1)},Z)J_{\mathrm{re}}(\theta^{(i-1)},Z) + \lambda^2 I_{n_\theta})\Delta\theta^{(i)} = -J_{\mathrm{re}}^T(\theta^{(i-1)},Z)\varepsilon_{\mathrm{re}}(\theta^{(i-1)},Z) \tag{7-219}
$$

(see Fletcher, 1991). The numerical stability of (7-219) is improved by solving the overdetermined set of equations

$$
\begin{bmatrix} J_{\mathrm{re}}(\theta^{(i-1)},Z) \\ \lambda I_{n_\theta} \end{bmatrix}\Delta\theta^{(i)} = -\begin{bmatrix} e_{\mathrm{re}}(\theta^{(i-1)},Z) \\ 0 \end{bmatrix} \tag{7-220}
$$

using a QR factorization (see Section 13.4.3). If the constraint $\|\theta\|_2 = 1$ is used, then the Levenberg-Marquardt version of (7-18) is calculated as

$$
\Delta\theta^{(i)} = -V\Lambda U^T e_{\mathrm{re}}(\theta^{(i-1)},Z) \tag{7-221}
$$

with

$$
\begin{aligned}
J_{\mathrm{re}}(\theta^{(i-1)},Z) &= U\mathrm{diag}(\sigma_1,\sigma_2,\ldots,\sigma_{n_\theta-1},0)V^T\\
\Lambda &= \mathrm{diag}(\frac{\sigma_1}{\sigma_1^2+\lambda^2},\frac{\sigma_2}{\sigma_2^2+\lambda^2},\ldots,\frac{\sigma_{n_\theta-1}}{\sigma_{n_\theta-1}^2+\lambda^2},0)
\end{aligned}
$$

and $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_{n_\theta-1}$. The initial value of λ in (7-220) and (7-221) is chosen proportional to the largest singular value of $J_{\mathrm{re}}(\theta^{(0)}, Z)$, for example, $\lambda = \sigma_1/100$. If the iteration step $\theta^{(i)} = \theta^{(i-1)} + \Delta\theta^{(i)}$ is successful (the ML cost function decreases), then λ is decreased as $\lambda \rightarrow 0.4\lambda$; otherwise (the ML cost function increases) λ is increased as $\lambda \rightarrow 10\lambda$ and the iteration is restarted from $\theta^{(i-1)}$.

Appendix 7.M: Number of Free Parameters in an Errors-in-Variables Problem

The total number of free parameters in an errors-in-variables problem (7-1) equals the sum of the number of free parameters in $\tilde{Z}_{0\text{re}}$ and the number of free model parameters in θ. According to Section 17.3, the number of free parameters in $\tilde{Z}_{0\text{re}}$ equals $\text{rank}(C_{N_{Z\text{re}}}) - 2F$. Using $C_{N_{Z\text{re}}} = 0.5(C_{N_Z})_{\text{Re}}$ and $\text{rank}((C_{N_Z})_{\text{Re}}) = 2\text{rank}(C_{N_Z})$ (Lemma 13.3) the total number of free parameters becomes

$$2\text{rank}(C_{N_Z}) - 2F + n_\theta \tag{7-222}$$

Under Assumption 7.20 we have $\text{rank}(C_{N_Z}) = F$ and (7-222) reduces to n_θ. We conclude that under Assumption 7.20 the total number of free parameters is independent of F. □

Appendix 7.N: Uncertainty of the BTLS Estimator in the Absence of Model Errors

Cost function (7-99), where θ_* is replaced by η, equals $f_F(\theta, \eta, Z)$. The expected value of the derivative of $f_F(\theta, \eta, Z)$ w.r.t θ is denoted as $g_F(\theta, \eta) = \mathcal{E}\{ f_F'(\theta, \eta, Z) \}$. In the absence of model errors, we have $\tilde{\theta}_{\text{BTLS}}(Z_0) = \theta_0$ ($\theta_{*\text{BTLS}} = \theta_0$ for model (7-8) with $\Omega = s$), so that $e(\Omega_k, \theta_0, Z_0(k)) = 0$ ($e(\Omega_k, \theta_0, Z_0(k)) = O(F^{-1/2})$, see Lemma 5.7), independently of η. Hence, from (7-99) and the definition of $g_F(\theta, \eta)$ it follows that $g_F(\theta_0, \eta) = 0$ $(g_F(\theta_0, \eta) = O(F^{-1}))$ for any η, so that $\partial g_F(\theta_0, \eta)/\partial \eta = 0$ $(\partial g_F(\theta_0, \eta)/\partial \eta = O(F^{-1}))$. We conclude that the second term in the right-hand side of (7-30) is zero (vanishes asymptotically w.r.t. the first term). It shows that the stochastic weighting does not increase the asymptotic uncertainty of the consistent BTLS estimate. □

Appendix 7.O: Asymptotic Properties of the Instrumental Variables Method

The basic step in the proof of properties 1, 2, 3, 6, and 7 of Theorem 7.21 (see Appendix 7.E) is the strong convergence of the cost function $V_{\text{IV}}(\theta, Z)$ and its (higher order) derivatives w.r.t. θ to their expected values. It is easy to verify that the strong laws of large numbers used to prove the strong convergence can also be applied to cost functions that are bilinear in the measurements $Z^{[1]}$ and $Z^{[2]}$. To prove the asymptotic normality (properties 4 and 6) it is sufficient to note that the central limit theorems used in Appendix 7.E also apply to cost functions that are bilinear in the measurements $Z^{[1]}$ and $Z^{[2]}$. The deterministic convergence (property 5) follows the same lines of Appendix 7.E exactly. □

Appendix 7.P: Equivalences between Range Spaces

In this appendix we show the following equivalences between the range spaces: (i) $\text{range}(O_r\mathbf{X}) = \text{range}(O_r)$, and (ii) $\text{range}(O_r\mathbf{X}^{\text{re}}\Pi) = \text{range}(O_r)$.

7.P.1 Proof of the First Equivalence. To prove the first equivalence, it is sufficient to show that $\mathbf{X}$ has full rank. The n_a by F matrix $\mathbf{X}$ in (7-115) is rank deficient if and only if there exists a (complex) row vector $C \neq 0$ such that $C\mathbf{X} = 0$. From (7-110) it follows that $X(k) = (\xi_k I_{n_a} - A)^{-1}BU(k)$. Assuming that $U(k) \neq 0$ it is possible to rewrite $C\mathbf{X} = 0$ as

$$C(\xi_k I_{n_a} - A)^{-1}B = 0 \text{ for } k = 1, 2, \ldots, F \tag{7-223}$$

Because, by assumption, at least $n_a + 1$ frequencies are distinct (Assumption 7.14) and the strictly proper system $G(\xi) = C(\xi I_{n_a} - A)^{-1}B$ can, at most, have $n_a - 1$ zeros, (7-223) can only be true if and only if $G(\xi) \equiv 0$. Assuming that the state space realization (7-110) is controllable, $G(\xi) \equiv 0$ can only be true if and only if $C = 0$ (Kailath, 1980). □

7.P.2 Proof of the Second Equivalence. The range space of $O_r \mathbf{X}^{\text{re}} \Pi$ equals the range space of O_r, unless rank cancellation occurs in $\mathbf{X}^{\text{re}} \Pi$. The rank cancellation does not occur if the intersection between the row spaces of $\mathbf{X}^{\text{re}}$ and $\mathbf{U}^{\text{re}}$ is empty. This is true if the $r + n_a$ by F matrix

$$\mathbf{Z} = \begin{bmatrix} \mathbf{U} \\ \mathbf{X} \end{bmatrix} \tag{7-224}$$

has rank $r + n_a$. $\mathbf{Z}$ is rank deficient if and only if there exists a row vector $L \neq 0$ such that $L\mathbf{Z} = 0$. Putting $L = [Dl_1 l_2 \ldots l_{r-1} C]$, and using $X(k) = (\xi_k I_{n_a} - A)^{-1}BU(k)$ (see (7-110)), $L\mathbf{Z} = 0$ can be written as

$$G(\xi_k) = 0 \text{ for } k = 1, 2, \ldots, F \tag{7-225}$$

with $G(\xi) = (D + C(\xi I_{n_a} - A)^{-1}B) + \sum_{m=1}^{r-1} l_m \xi^m$. Because, by assumption, at least $r + n_a + 1$ frequencies are distinct (Assumption 7.14) and $G(\xi)$ has, at most, $n_a + r - 1$ zeros, (7-225) can only be true if and only if $G(\xi) \equiv 0$. $G(\xi) \equiv 0$ is true if and only if $D = 0$, $l_m = 0$ for $m = 1, 2, \ldots, r-1$, and $C(\xi I_{n_a} - A)^{-1}B \equiv 0$. Assuming that the state space realization (7-110) is controllable, $C(\xi I_{n_a} - A)^{-1}B \equiv 0$ is true if and only if $C = 0$ (Kailath, 1980). □

Appendix 7.Q: Estimation of the Range Space

The range space of a matrix equals the span of its left singular vectors corresponding to the nonzero singular values. In the first part of this appendix we study the estimation of the left singular vectors of a noisy r by F matrix A for $F \to \infty$. In the second part these results are applied to estimation of the range space of O_r.

7.Q.1 Asymptotic Properties of the Left Singular Vectors. Consider the real r by F matrix $A = A_0 + N$, where A_0 is the deterministic part and N is the zero mean noise contribution. The left singular vectors of A are equal to the eigenvectors $\hat{x}$ of AA^T (see Exercise 13.16)

$$(AA^T/F)\hat{x} = \hat{\lambda}\hat{x} \tag{7-226}$$

If NN^T/F and $A_0 N^T/F$ converge w.p. 1 to, respectively, C_N and 0, then AA^T/F converges w.p. 1 to

$$\mathbf{A}_0\mathbf{A}_0^T + C_N \text{ with } \mathbf{A}_0\mathbf{A}_0^T = \lim_{F \to \infty} A_0 A_0^T/F \tag{7-227}$$

Hence, (7-226) converges w.p. 1 to

$$(\mathbf{A}_0\mathbf{A}_0^T + C_N)x_* = \lambda_* x_* \tag{7-228}$$

where x_* and λ_* are the limit values of respectively $\hat{x}$ and $\hat{\lambda}$.

If $C_N = \sigma^2 I_r$ then (7-228) becomes $\mathbf{A}_0\mathbf{A}_0^T x_* = (\lambda_* - \sigma^2)x_*$. Clearly, $x_* = x_0$, with x_0 the left singular vector of $\mathbf{A}_0$. This proves the strong convergence of $\hat{x}$ to x_0.

If $C_N \neq \sigma^2 I_r$ then we form the matrix $B = C_N^{-1/2}A$, where $C_N^{1/2}$ is a square root of C_N, $C_N^{1/2}C_N^{T/2} = C_N$. Because $C_N^{-1/2}(NN^T/F)C_N^{-T/2}$ converges w.p. 1 to I_r, the equation $(BB^T/F)\hat{y} = \hat{\mu}\hat{y}$ converges w.p. 1 to

$$\mathbf{B}_0\mathbf{B}_0^T y_* = (\mu_* - 1)y_* \text{ with } \mathbf{B}_0 = C_N^{-1/2}\mathbf{A}_0 \tag{7-229}$$

Clearly $y_* = y_0$, with y_0 the left singular vector of $\mathbf{B}_0 = C_N^{-1/2}\mathbf{A}_0$. It follows that $y_0 = C_N^{-1/2}x_0$, which proves the strong convergence of $C_N^{1/2}\hat{y}$ to x_0.

Note that we have shown the strong convergence of the estimate $\hat{x}$ to the solution x_0 of the noiseless problem $\mathbf{A}_0$, without requiring the existence of a true model. If a true model exists and if it belongs to the considered model set, then x_0 equals the true value.

7.Q.2 Estimation of the Range Space of O_r. The results of the first part of this appendix are applicable to $\mathbf{Y}^{\text{re}}\Pi$ if $\mathbf{N}\mathbf{N}^T/F$ and $\mathbf{X}^{\text{re}}\Pi\mathbf{N}^T/F$ converge w.p. 1 to, respectively, some C_N and 0. We will show that this is true under the assumptions of Section 7.6.1 and assuming that $N_U(k) = 0$.

Using (7-237) with $\mathbf{N}_\mathrm{U}^{\text{re}} = 0$ we find

$$\begin{aligned}
\mathbf{N}\mathbf{N}^T &= \mathbf{N}_\mathrm{Y}^{\text{re}}\mathbf{N}_\mathrm{Y}^{\text{re}T} - \mathbf{N}_\mathrm{Y}^{\text{re}}\mathbf{U}^{\text{re}T}(\mathbf{U}^{\text{re}}\mathbf{U}^{\text{re}T})^{-1}\mathbf{U}^{\text{re}}\mathbf{N}_\mathrm{Y}^{\text{re}T} \\
&= \mathrm{Re}(\mathbf{N}_\mathrm{Y}\mathbf{N}_\mathrm{Y}^H) - \mathrm{Re}(\mathbf{N}_\mathrm{Y}\mathbf{U}^H)(\mathrm{Re}(\mathbf{U}\mathbf{U}^H))^{-1}\mathrm{Re}(\mathbf{U}\mathbf{N}_\mathrm{Y}^H) \\
\mathbf{X}^{\text{re}}\Pi\mathbf{N}^T &= \mathrm{Re}(\mathbf{X}\mathbf{N}_\mathrm{Y}^H) - \mathrm{Re}(\mathbf{X}\mathbf{U}^H)(\mathrm{Re}(\mathbf{U}\mathbf{U}^H))^{-1}\mathrm{Re}(\mathbf{U}\mathbf{N}_\mathrm{Y}^H)
\end{aligned} \tag{7-230}$$

with

$$\begin{aligned}
\mathbf{U}\mathbf{U}^H &= \textstyle\sum_{k=1}^F |U_0(k)|^2 W_r(k) W_r^H(k) \\
\mathbf{X}\mathbf{U}^H &= \textstyle\sum_{k=1}^F X(k)\bar{U}_0(k) W_r(k) W_r^H(k)
\end{aligned}$$

and

$$\begin{aligned}
\mathbf{N}_\mathrm{Y}\mathbf{N}_\mathrm{Y}^H/F &= \frac{1}{F}\textstyle\sum_{k=1}^F |N_Y(k)|^2 W_r(k) W_r^H(k) \\
\mathbf{N}_\mathrm{Y}\mathbf{U}^H/F &= \frac{1}{F}\textstyle\sum_{k=1}^F N_Y(k)\bar{U}_0(k) W_r(k) W_r^H(k) \\
\mathbf{X}\mathbf{N}_\mathrm{Y}^H/F &= \frac{1}{F}\textstyle\sum_{k=1}^F X(k)\bar{N}_Y(k) W_r(k) W_r^H(k)
\end{aligned} \tag{7-231}$$

By assumption $\mathrm{Re}(\mathbf{U}\mathbf{U}^H)/F > cI_r$, with $0 < c < \infty$ and c independent of F, for any F, ∞ included, and, hence, $(\mathrm{Re}(\mathbf{U}\mathbf{U}^H))^{-1} = O(F^{-1})$. For a frequency domain experiment, $N_Y(k)$

is independently distributed over k, while for a time domain experiment, $|N_Y(k)|^2$ and $N_Y(k)$ converge w.p. 1 to random variables which are mixing of order 2 (see Section 14.16). Hence, the sums in (7-231) converge w.p. 1 to their expected values (see Section 14.9, versions 2 and 3 of the law of large numbers)

$$\begin{aligned} \mathcal{E}\{\mathbf{N}_{\mathbf{Y}}\mathbf{N}_{\mathbf{Y}}^{H}/F\} &= \frac{1}{F}\sum\nolimits_{k=1}^{F}\sigma_Y^2(k)W_r(k)W_r^H(k) \\ \mathcal{E}\{\mathbf{N}_{\mathbf{Y}}\mathbf{U}^{H}/F\} &= 0 \\ \mathcal{E}\{\mathbf{X}\mathbf{N}_{\mathbf{Y}}^{H}/F\} &= 0 \end{aligned} \tag{7-232}$$

Hence, $\mathbf{N}\mathbf{N}^T/F$ converges w.p. 1 to

$$C_{\mathbf{Y}}/F = \mathcal{E}\{\mathrm{Re}(\mathbf{N}_{\mathbf{Y}}\mathbf{N}_{\mathbf{Y}}^{H})\}/F = \frac{1}{F}\mathrm{Re}(\sum\nolimits_{k=1}^{F}\sigma_Y^2(k)W_r(k)W_r^H(k)) \tag{7-233}$$

and $\mathbf{X}^{\mathrm{re}}\Pi\mathbf{N}^T/F$ converges w.p. 1 to 0.

We conclude that $\mathrm{range}(C_{\mathbf{Y}}^{1/2}U_{[:,\,1:n_a]})$ converges strongly to $\mathrm{range}(O_r\mathbf{X}^{\mathrm{re}}\Pi)$, which is equal to $\mathrm{range}(O_r)$ (Appendix 7.P). Hence, we have established the strong consistency of $\hat{O}_r$. From versions 2 and 3 of the law of large numbers (see Section 14.9), it follows that $\mathbf{N}\mathbf{N}^T/F$ and $\mathbf{X}^{\mathrm{re}}\Pi\mathbf{N}^T/F$ converge in probability at the rate $O_{\mathrm{p}}(F^{-1/2})$ to their limit value. Hence, this is also valid for $\mathbf{Y}^{\mathrm{re}}\Pi$ and $C_{\mathbf{Y}}^{1/2}U_{[:,\,1:n_a]}$ (7-241) so that $\hat{O}_r$ converges in probability at the rate $O_{\mathrm{p}}(F^{-1/2})$ to O_r. In case of model errors Assumption 7.27 guarantees that the results of the first part of this appendix can be applied to equation (7-241) showing the strong convergence of $\hat{O}_r = \mathrm{range}(C_{\mathbf{Y}}^{1/2}U_{[:,\,1:n_a]})$ to the solution $O_{\mathrm{r}*}$ of the noiseless problem ($\mathbf{N} = 0$ in (7-237)). The convergence rate is also an $O_{\mathrm{p}}(F^{-1/2})$.

Appendix 7.R: Subspace Algorithm for Discrete-Time Systems (Algorithm 7.24)

We first discuss the three basic steps of the subspace algorithm in more details. Next, we present a numerical efficient implementation.

7.R.1 Basic Subspace Algorithm. The three basic steps of the subspace algorithm are (i) estimation of the range space of O_r, (ii) estimation of A and C given $\hat{O}_r$, and (iii) estimation of B and D given $\hat{A}$ and $\hat{C}$.

First Step. As O_r is known only within a right invertible transformation matrix T, see (7-117), it is sufficient to estimate the range space of O_r. If the $F > n_a$ frequencies are distinct, then the matrix $\mathbf{X}$ has rank n_a (see Appendix 7.P), and $\mathrm{range}(O_r\mathbf{X}) = \mathrm{range}(O_r)$. Hence, we can estimate the range of O_r using (7-119) if we can eliminate $S_r\mathbf{U}$ and suppress the influence of the noise $\mathbf{N}_{\mathbf{Y}} - S_r\mathbf{N}_{\mathbf{U}}$.

Because O_r is a real matrix, we are interested in a real range space. Therefore, we convert (7-119) into a set of real equations as

$$\mathbf{Y}^{\mathrm{re}} = O_r\mathbf{X}^{\mathrm{re}} + S_r\mathbf{U}^{\mathrm{re}} + \mathbf{N}_{\mathbf{Y}}^{\mathrm{re}} - S_r\mathbf{N}_{\mathbf{U}}^{\mathrm{re}} \tag{7-234}$$

where $(\ )^{\mathrm{re}}$ locates the real and imaginary parts beside each other, for example,

$$\mathbf{Y}^{\mathrm{re}} = [\mathrm{Re}(\mathbf{Y})\ \mathrm{Im}(\mathbf{Y})] \tag{7-235}$$

The operator $(\ )^{\mathrm{re}}$ should not be confused with $(\ )_{\mathrm{re}}$, which stacks the real and imaginary parts on top of each other. Both operators are related by $X^{\mathrm{re}} = ((X^T)_{\mathrm{re}})^T$.

The term $S_r\mathbf{U}^{\mathrm{re}}$ in (7-234) is eliminated by right multiplication of (7-234) with an orthogonal projection Π

$$\Pi = I_{2F} - \mathbf{U}^{\mathrm{re}T}(\mathbf{U}^{\mathrm{re}}\mathbf{U}^{\mathrm{re}T})^{-1}\mathbf{U}^{\mathrm{re}} \tag{7-236}$$

which has the property $\mathbf{U}^{\mathrm{re}}\Pi = 0$. We get

$$\begin{aligned} \mathbf{Y}^{\mathrm{re}}\Pi &= O_r\mathbf{X}^{\mathrm{re}}\Pi + \mathbf{N} \\ \mathbf{N} &= (\mathbf{N}_{\mathrm{Y}}^{\mathrm{re}} - S_r\mathbf{N}_{\mathrm{U}}^{\mathrm{re}})\Pi \end{aligned} \tag{7-237}$$

If $\mathbf{U}^{\mathrm{re}}\mathbf{U}^{\mathrm{re}T}/F > cI_r$ with $0 < c < \infty$ and where c is independent of F, for any F, ∞ included, the frequencies are distinct, and $F \geq n_a + r$, then $\mathrm{range}(O_r\mathbf{X}^{\mathrm{re}}\Pi) = \mathrm{range}(O_r)$ for any F, ∞ included (see Appendix 7.P).

From (7-237) it follows that the range of O_r can be estimated as $\mathrm{range}(\mathbf{Y}^{\mathrm{re}}\Pi)$. Since the range of a matrix equals the span of the left singular vectors corresponding to the nonzero singular values (see Section 13.4.1), $\mathrm{range}(\mathbf{Y}^{\mathrm{re}}\Pi)$ is calculated via a singular value decomposition (SVD) of $\mathbf{Y}^{\mathrm{re}}\Pi$. The left singular vectors of $\mathbf{Y}^{\mathrm{re}}\Pi$ are consistently estimated if

$$\operatorname*{a.s.lim}_{F\to\infty} \mathbf{X}^{\mathrm{re}}\Pi\mathbf{N}^T/F = 0 \text{ and } \operatorname*{a.s.lim}_{F\to\infty} \mathbf{N}\mathbf{N}^T/F = C_{\mathrm{N}} \text{ with } C_{\mathrm{N}} = \sigma^2 I_r \tag{7-238}$$

(see Appendix 7.Q). The second condition in (7-238) is in general not satisfied and, therefore, the noise $\mathbf{N}$ in (7-237) is whitened by left multiplication of (7-237) with $C_{\mathrm{N}}^{-1/2}$,

$$C_{\mathrm{N}}^{-1/2}\mathbf{Y}^{\mathrm{re}}\Pi = C_{\mathrm{N}}^{-1/2}O_r\mathbf{X}^{\mathrm{re}}\Pi + C_{\mathrm{N}}^{-1/2}\mathbf{N} \tag{7-239}$$

where $C_{\mathrm{N}}^{1/2}$ is a square root of C_{N} (see Section 13.4.4 for the calculation of the square root of a positive (semi-)definite matrix). Because $\mathcal{E}\{C_{\mathrm{N}}^{-1/2}(\mathbf{N}\mathbf{N}^T/F)C_{\mathrm{N}}^{-T/2}\} \to I_r$ for $F \to \infty$, the left singular vectors of $C_{\mathrm{N}}^{-1/2}\mathbf{Y}^{\mathrm{re}}\Pi$ are consistently estimated. From the SVD

$$C_{\mathrm{N}}^{-1/2}\mathbf{Y}^{\mathrm{re}}\Pi = U\Sigma V^T \tag{7-240}$$

we estimate the extended observability matrix O_r as

$$C_{\mathrm{N}}^{-1/2}\hat{O}_r = U_{[:,\,1:n_a]} \text{ or } \hat{O}_r = C_{\mathrm{N}}^{1/2}U_{[:,\,1:n_a]} \tag{7-241}$$

The problem with the proposed algorithm is that $\mathbf{N}$ is a function of the unknown state space parameters, via S_r (see (7-113) and (7-237)), and, hence, C_{N} cannot be calculated. If the input is exactly known, $\mathbf{N}_{\mathrm{U}} = 0$, then $\mathbf{N}$ is independent of S_r and C_{N} can be calculated. If the input observations are disturbed by noise, $\mathbf{N}_{\mathrm{U}} \neq 0$, then we replace $U(k)$, $Y(k)$, $\sigma_U^2(k)$, and $\sigma_Y^2(k)$ everywhere by, respectively, 1, $G(\Omega_k) = Y(k)/U(k)$, 0, and σ_G^2 (7-53). If the worst case input and output signal-to-noise ratio is larger than 10 dB, then the bias on $\mathcal{E}\{Y(k)/U(k)\}$ can be neglected and the variance of the truncated ratio $\underline{G}(\Omega_k)$ is given by

(7-53) (see Section 7.9 for an elaborated discussion). We conclude that from a practical point of view $Y(k)/U(k)$ acts as a zero mean random variable with variance (7-53). From now on, we will assume that $\mathbf{N_U} = 0$. The matrix $C_\mathbf{N}$ is then asymptotically $(F \to \infty)$ given by

$$C_\mathbf{N} = \lim_{F \to \infty} C_\mathbf{Y}/F \text{ with } C_\mathbf{Y} = \mathcal{E}\{\mathbf{N}_\mathbf{Y}^{\text{re}}\mathbf{N}_\mathbf{Y}^{\text{re}T}\} = \text{Re}(\textstyle\sum_{k=1}^{F} \sigma_Y^2(k) W_r(k) W_r^H(k)) \tag{7-242}$$

(see Appendix 7.Q). We conclude that the estimate $\hat{O}_r$ converges w.p. 1 to the true solution O_{r0} under the assumptions of Section 7.6.6 and Assumption 7.14 and that it convergences w.p. 1 to the noiseless solution O_{r*} under the assumptions of Section 7.6.1 and Assumption 7.14. Moreover, the convergence rate is an $O_\text{p}(F^{-1/2})$ (see Appendix 7.Q).

Second Step. Using the estimate $\hat{O}_r$ we calculate from (7-116) and (7-113),

$$\hat{A} = \hat{O}^+_{r[1:r-1,:]}\hat{O}_{r[2:r,:]} \text{ and } \hat{C} = \hat{O}_{r[1,:]} \tag{7-243}$$

Note that $\hat{A}$, $\hat{C}$, and their derivatives w.r.t. $\hat{O}_r$ are continuous functions of $\hat{O}_r$ in a closed and bounded neighborhood of the true value O_{r0} or the noiseless solution O_{r*}. Because $\hat{O}_r$ converges w.p. 1 to the true solution or to the noiseless solution, the estimates $\hat{A}$ and $\hat{C}$ also converge w.p. 1 to the true solution or to the noiseless solution (Lemma 15.31). Since the convergence rate of $\hat{O}_r$ is an $O_\text{p}(F^{-1/2})$, the convergence rate of $\hat{A}$ and $\hat{C}$ is also an $O_\text{p}(F^{-1/2})$ (Lemma 15.34).

Third Step. We choose $W(\xi_k) = \sigma_Y^{-1}(k)$ in the cost function $V_\text{SUB}(C, D, \hat{A}, \hat{C}, Z)$ (7-120). If $N_U(k) \neq 0$ then we replace $Y(k)$, $U(k)$ by $Y(k)/U(k)$, 1 (see First step) and put $W(\xi_k) = \sigma_G^{-1}(k)$ (see (7-53)). This choice would give the smallest uncertainty on the estimates $\hat{C}$ and $\hat{D}$ if $\hat{A}$ and $\hat{C}$ were nonrandom. Under Assumptions 7.14 and 7.16 the linear least squares problem (7-120) is identifiable if and only if the state space realization (7-110) is observable (McKelvey et al., 1996).

Under the assumptions of Sections 7.6.5 and 7.6.6, the estimates $\hat{B}$ and $\hat{D}$ are strongly consistent because $\hat{A}$ and $\hat{C}$ are strongly consistent. To prove this statement, it is sufficient to apply Theorem 16.7 with $w(\theta, \eta(Z), Z) = 0$, $\eta(Z) = [(\text{vec}(\hat{A}))^T\ \hat{C}^T]^T$, and $\eta_* = [(\text{vec}(A_0))^T\ C_0^T]^T$, and to verify that $\mathcal{E}\{V_\text{SUB}(C, D, A_0, C_0, Z)\}$ is minimal in the true parameters C_0, D_0. This last condition is satisfied because for $W(\xi_k) = \sigma_Y^{-1}(k)$,

$$\mathcal{E}\{V_\text{SUB}(C, D, A_0, C_0, Z)\} = V_\text{SUB}(C, D, A_0, C_0, Z_0) + F \tag{7-244}$$

with $V_\text{SUB}(C_0, D_0, A_0, C_0, Z_0) = 0$. Similarly, under the assumptions of Sections 7.6.1 and 7.6.5, the estimates $\hat{B}$ and $\hat{D}$ converge w.p. 1 the noiseless solution (see Theorem 16.5).

Under the assumptions of Section 7.6.2 and 7.6.5 the estimates $\hat{B}$ and $\hat{D}$ converge in probability at the rate $O_\text{p}(F^{-1/2})$ to their limit value. To prove this, it is sufficient to note that the convergence rate of $\hat{A}$ and $\hat{C}$ is an $O_\text{p}(F^{-1/2})$ and to verify that all the conditions of Theorem 16.16 are satisfied.

7.R.2 Numerical Efficient Implementation. The matrix $\mathbf{Y}^\text{re}\Pi$ can be calculated without forming the huge $2F$ by $2F$ matrix Π. This is done as follows. Form the matrix $Z = [\mathbf{U}^{\text{re}T}\ \mathbf{Y}^{\text{re}T}]^T$ and calculate the QR factorization (see Section 13.4.3) $Z^T = QR$ with $Q^TQ = I_{4F}$ and R a $2r$ by $2r$ upper triangular matrix. This factorization can be written as

$$Z = R^T Q^T \text{ or } \begin{bmatrix} \mathbf{U}^{\text{re}} \\ \mathbf{Y}^{\text{re}} \end{bmatrix} = \begin{bmatrix} R_{11}^T & 0 \\ R_{12}^T & R_{22}^T \end{bmatrix} \begin{bmatrix} Q_1^T \\ Q_2^T \end{bmatrix} \tag{7-245}$$

with R_{11} a regular r by r matrix. Using the property $Q_2^T Q_1 = 0$, it is easy to verify that $\mathbf{Y}^{\text{re}}\Pi = R_{22}^T Q_2^T$. The left singular vectors of $\mathbf{Y}^{\text{re}}\Pi$ are the eigenvectors of $\mathbf{Y}^{\text{re}}\Pi(\mathbf{Y}^{\text{re}}\Pi)^T$ (see Exercise 13.16). Using $Q_2^T Q_2 = I_{2F}$ we find that $\mathbf{Y}^{\text{re}}\Pi(\mathbf{Y}^{\text{re}}\Pi)^T = R_{22}^T R_{22}$ is independent of Q_2. It follows that the left singular vectors of $\mathbf{Y}^{\text{re}}\Pi$ and the (asymptotic) covariance matrices $C_{\mathbf{N}}$ and $C_{\mathbf{Y}}$ (7-242) are not influenced by Q_2. Hence, we can calculate (7-240) as

$$C_{\mathbf{Y}}^{-1/2} R_{22}^T = U\Sigma V^T \tag{7-246}$$

Appendix 7.S: Subspace Algorithm for Continuous-Time Systems (Algorithm 7.25)

The algorithm is a three-step procedure. The main differences from Algorithm 7.24 for discrete-time systems are (i) the orthogonalization of the input and output data, (ii) the estimation of a generalized extended observability matrix $O_{r\perp}$ in the first step, and (iii) the estimation of $\hat{A}$ from a generalized shift property of $O_{r\perp}$ in the second step. The third step remains exactly the same. We first explain the orthogonalization, next we discuss the impact of the orthogonalization on the model equation (7-119), and finally we prove the generalized shift property of $O_{r\perp}$. The appendix is concluded with a discussion of the stochastic properties and the numerical implementation.

7.S.1 Orthogonalization Procedure. The data matrices $\mathbf{Y}$ and $\mathbf{U}$ in (7-115) depend on the scalar polynomial basis s^n, $n = 0, 1, \ldots, r-1$. We will construct two scalar orthogonal polynomial bases $p_n(s)$, $n = 0, 1, \ldots, r-1$ and $q_n(s)$, $n = 0, 1, \ldots, r-1$ such that the data matrices $\mathbf{Y}_\perp$, constructed using $p_n(s)$, and $\mathbf{U}_\perp$, constructed using $q_n(s)$, satisfy, respectively, $\text{Re}(\mathbf{Y}_\perp \mathbf{Y}_\perp^H) = I_r$ and $\text{Re}(\mathbf{U}_\perp \mathbf{U}_\perp^H) = I_r$. The matrix $Z_\perp = [\mathbf{U}_\perp^{\text{re}T} \; \mathbf{Y}_\perp^{\text{re}T}]^T$, where $(\;)^{\text{re}}$ locates the real and imaginary parts of the matrix beside each other (see (7-235)), thus has the property

$$Z_\perp Z_\perp^T = \begin{bmatrix} I_r & C_1 \\ C_1^T & I_r \end{bmatrix} \tag{7-247}$$

In Rolain et al. (1995) it has been shown that no other two scalar polynomial bases resulting in a smaller condition number of $Z_\perp Z_\perp^T$ can be found. Hence, $Z_\perp$ is best conditioned for scalar bases.

The matrices $\text{Re}(\mathbf{Y}\mathbf{Y}^H)$ and $\text{Re}(\mathbf{U}\mathbf{U}^H)$ each define an inner product,

$$\begin{aligned} \text{Re}(\mathbf{Y}\mathbf{Y}^H) &\Rightarrow \langle x(s), y(s)\rangle_Y = \text{Re}\left(\textstyle\sum_{k=1}^F x(s_k)\bar{y}(s_k)|Y(k)|^2\right) \\ \text{Re}(\mathbf{U}\mathbf{U}^H) &\Rightarrow \langle x(s), y(s)\rangle_U = \text{Re}\left(\textstyle\sum_{k=1}^F x(s_k)\bar{y}(s_k)|U(k)|^2\right) \end{aligned} \tag{7-248}$$

that is used to calculate, respectively, the bases $p_n(s)$, $n = 0, 1, \ldots, r-1$, and $q_n(s)$, $n = 0, 1, \ldots, r-1$, via a Gram-Schmidt orthogonalization procedure (see Section 13.11). Applying this procedure with inner product $\langle\ ,\ \rangle_Y$ gives

1. Initialization:

$$\begin{aligned} p_0(s) &= 1/\alpha_1 && \text{with } \alpha_1 = \|1\| \\ p_1(s) &= sp_0(s)/\alpha_2 && \text{with } \alpha_2 = \|sp_0(s)\| \end{aligned} \tag{7-249}$$

2. Recursion: for $n = 2$ to $r-1$

$$\begin{aligned} p_n(s) &= (sp_{n-1}(s) + \alpha_n p_{n-2}(s))/\alpha_{n+1} \quad \text{with} \\ \alpha_{n+1} &= \|sp_{n-1}(s) + \alpha_n p_{n-2}(s)\| \end{aligned} \tag{7-250}$$

where $\|t(s)\|^2 = \langle t(s), t(s)\rangle_Y$. The resulting polynomial basis has the property

$$\langle p_n(s), p_m(s)\rangle_Y = \delta_{nm} \tag{7-251}$$

(see Section 13.11). Now define the vector $W_Y(k)$

$$W_Y(k) = \left[p_0(s_k)\ p_1(s_k)\ \ldots\ p_{r-1}(s_k)\right]^T \tag{7-252}$$

and construct $\mathbf{Y}_\perp$ as

$$\mathbf{Y}_\perp = \left[W_Y(1)Y(1)\ W_Y(2)Y(2)\ \ldots\ W_Y(F)Y(F)\right] \tag{7-253}$$

Using (7-251), it is easy to verify that

$$\mathrm{Re}(\mathbf{Y}_\perp \mathbf{Y}_\perp^T) = \mathrm{Re}\left(\sum_{k=1}^{F} |Y(k)|^2 W_Y(k) W_Y^H(k)\right) = I_r \tag{7-254}$$

Making the same calculations with the inner product $\langle\ ,\ \rangle_U$, gives the scalar orthogonal polynomial basis $q_n(s)$ and the numbers β_n, $n = 0, 1, \ldots, r-1$. Similar to (7-252) and (7-253), we define

$$W_U(k) = \left[q_0(s_k)\ q_1(s_k)\ \ldots\ q_{r-1}(s_k)\right]^T \tag{7-255}$$

$$\mathbf{U}_\perp = \left[W_U(1)U(1)\ W_U(2)U(2)\ \ldots\ W_U(F)U(F)\right] \tag{7-256}$$

where $\mathrm{Re}(\mathbf{U}_\perp \mathbf{U}_\perp^T) = I_r$.

7.S.2 Impact of the Orthogonalization on the Model Equation. From the Gram-Schmidt procedure it follows that the orthogonal polynomial bases $p_n(s)$, $n = 0, 1, \ldots, r-1$, and $q_n(s)$, $n = 0, 1, \ldots, r-1$, are related to the basis s^n, $n = 0, 1, \ldots, r-1$, via a lower triangular r by r matrix

$$W_Y(k) = L_Y W_r(k),\ W_U(k) = L_U W_r(k) \text{ with } L_Y, L_U \in \mathbb{R}^{r \times r} \tag{7-257}$$

Applying (7-257) to (7-253) and (7-256) gives, using (7-115),

$$\mathbf{Y}_\perp = L_Y\mathbf{Y} \qquad \mathbf{Y}_\perp^{\mathrm{re}} = L_Y\mathbf{Y}^{\mathrm{re}}$$
$$\mathbf{U}_\perp = L_U\mathbf{U} \qquad \mathbf{U}_\perp^{\mathrm{re}} = L_U\mathbf{U}^{\mathrm{re}} \tag{7-258}$$

Left multiplication of (7-234) with $\mathbf{N}_\mathbf{U} = 0$, by L_Y, gives, using (7-258),

$$\mathbf{Y}_\perp^{\mathrm{re}} = O_{r\perp}\mathbf{X}^{\mathrm{re}} + L_Y S_r L_U^{-1}\mathbf{U}_\perp^{\mathrm{re}} + L_Y\mathbf{N}_\mathbf{Y}^{\mathrm{re}} \tag{7-259}$$

with $O_{r\perp} = L_Y O_r$ the generalized extended observability matrix. Constructing the orthogonal projection $\Pi_\perp$ as in (7-236), where $\mathbf{U}^{\mathrm{re}}$ is replaced by $\mathbf{U}_\perp^{\mathrm{re}}$, makes it possible to eliminate the input term in (7-259)

$$\mathbf{Y}_\perp^{\mathrm{re}}\Pi_\perp = O_{r\perp}\mathbf{X}^{\mathrm{re}}\Pi_\perp + \mathbf{N}_\perp \tag{7-260}$$

with $\mathbf{N}_\perp = L_Y\mathbf{N}_\mathbf{Y}^{\mathrm{re}}\Pi_\perp$. Using the results of Appendix 7.R and (7-253) it follows that $\mathbf{N}_\perp\mathbf{N}_\perp^T/F$ converges w.p. 1 to

$$C_{\mathbf{Y}_\perp}/F = \mathcal{E}\{\mathrm{Re}(\mathbf{N}_{\mathbf{Y}_\perp}\mathbf{N}_{\mathbf{Y}_\perp}^H)\}/F = \frac{1}{F}\mathrm{Re}(\textstyle\sum_{k=1}^{F}\sigma_Y^2(k)W_Y(k)W_Y^H(k)) \tag{7-261}$$

The range space of $O_{r\perp}$ is estimated as in Appendix 7.R: from $C_{\mathbf{Y}_\perp}^{-1/2}\mathbf{Y}_\perp^{\mathrm{re}}\Pi_\perp = U\Sigma V^T$ we get $\hat{O}_{r\perp} = C_{\mathbf{Y}_\perp}^{1/2}U_{[:,1:n_a]}$.

7.S.3 Generalized Shift Property. Because $O_{r\perp} = L_Y O_r$ and $\mathbf{Y}_\perp = L_Y\mathbf{Y}$, it follows that the rows of $O_{r\perp}$ and $\mathbf{Y}_\perp$ can be derived from the respective rows of O_r and $\mathbf{Y}$, using the same linear combinations

$$\mathbf{Y}_{\perp[n,:]} = \textstyle\sum_{m=1}^{n}\gamma_m\mathbf{Y}_{[m,:]}$$
$$O_{r\perp[n,:]} = \textstyle\sum_{m=1}^{n}\gamma_m O_{r[m,:]} \tag{7-262}$$

First, we establish the relationship between the rows of $\mathbf{Y}_\perp$. Next, using (7-262), we show that a similar relationship exists between the rows of $O_{r\perp}$.

Multiplying (7-249) and (7-250), evaluated at s_k, by $Y(k)$ for $k = 1, 2, \ldots, F$, gives the following relationship between the rows of $\mathbf{Y}_\perp$

1. Initialization:

$$\mathbf{Y}_{\perp[1,:]} = \mathbf{Y}_{[1,:]}/(\alpha_1,) \quad \mathbf{Y}_{\perp[2,:]} = \mathbf{Y}_{\perp[1,:]}D_s/\alpha_2 \tag{7-263}$$

with $Y_{[1,:]} = [Y(1) \;\; Y(2) \;\; \ldots \;\; Y(F)]$ and $D_s = \mathrm{diag}(s_1, s_2, \ldots, s_F)$.

2. Recursion: for $n = 3$ to r

$$\mathbf{Y}_{\perp[n,:]} = (\mathbf{Y}_{\perp[n-1,:]}D_s + \alpha_{n-1}\mathbf{Y}_{\perp[n-2,:]})/\alpha_n \tag{7-264}$$

Using the definition (7-248) of $\langle\;,\;\rangle_Y$, it follows that

$$\alpha_1 = \|\mathbf{Y}_{[1,:]}\|_2,\ \alpha_2 = \|\mathbf{Y}_{\perp[1,:]}D_s\|_2,\ \alpha_n = \|\mathbf{Y}_{\perp[n-1,:]}D_s + \alpha_{n-1}\mathbf{Y}_{\perp[n-2,:]}\|_2 \tag{7-265}$$

Using (7-262) we can rewrite (7-264) as

$$\begin{aligned}\mathbf{Y}_{\perp[n,:]} &= \left(\textstyle\sum_{m=1}^{n-1}\gamma_m\mathbf{Y}_{[m,:]}D_s + \alpha_{n-1}\sum_{m=1}^{n-2}\gamma_m\mathbf{Y}_{[m,:]}\right)/\alpha_n \\ &= \left(\textstyle\sum_{m=1}^{n-1}\gamma_m\mathbf{Y}_{[m+1,:]} + \alpha_{n-1}\sum_{m=1}^{n-2}\gamma_m\mathbf{Y}_{[m,:]}\right)/\alpha_n\end{aligned} \tag{7-266}$$

where the last equality is due to the property $\mathbf{Y}_{[m+1,:]} = \mathbf{Y}_{[m,:]}D_s$ (see (7-115)). The second equation of (7-266) is just another way of writing the first equation of (7-262). Hence, the rows of $O_{r\perp}$ should satisfy the same expression

$$\begin{aligned}O_{r\perp[n,:]} &= \left(\textstyle\sum_{m=1}^{n-1}\gamma_m O_{r[m+1,:]} + \alpha_{n-1}\sum_{m=1}^{n-2}\gamma_m O_{r[m,:]}\right)/\alpha_n \\ &= \left(\textstyle\sum_{m=1}^{n-1}\gamma_m O_{r[m,:]}A + \alpha_{n-1}\sum_{m=1}^{n-2}\gamma_m O_{r[m,:]}\right)/\alpha_n\end{aligned} \tag{7-267}$$

where the last equality is due to the shift property $O_{r[m+1,:]} = O_{r[m,:]}A$ (see (7-113)). Using (7-262), the last equation of (7-267) becomes

$$O_{r\perp[n,:]} = (O_{r\perp[n-1,:]}A + \alpha_{n-1}O_{r\perp[n-2]})/\alpha_n \tag{7-268}$$

for $n = 3, 4, \ldots, r$. Following the same lines we get from (7-263)

$$O_{r\perp[1,:]} = O_{r[1,:]}/\alpha_1 \tag{7-269}$$

$$O_{r\perp[2,:]} = O_{r\perp[1,:]}A/\alpha_2 \tag{7-270}$$

From (7-269) it follows that $C = \alpha_1 O_{r\perp[1,:]}$. Writing (7-268) and (7-270) under matrix notation gives the generalized shift property of $O_{r\perp}$

$$[D_1 O_{r\perp[1:r-1,:]}]A = [O_{r\perp[2:r,:]} - b] \tag{7-271}$$

with

$$b = \begin{bmatrix} 0 \\ D_2 O_{r\perp[1:r-2,:]} \end{bmatrix},\quad \begin{aligned} D_1 &= \operatorname{diag}(1/\alpha_2, 1/\alpha_3, \ldots, 1/\alpha_r) \\ D_2 &= \operatorname{diag}(\alpha_2/\alpha_3, \alpha_3/\alpha_4, \ldots, \alpha_{r-1}/\alpha_r) \end{aligned}$$

7.S.4 Discussion. The range space estimation of $O_{r\perp}$ follows exactly the same lines as the range space estimation of O_r. Therefore, $\hat{O}_{r\perp}$ has the same asymptotic ($F \to \infty$) properties as $\hat{O}_r$ in Appendix 7.R. Because $\hat{A}$ and $\hat{C}$ are continuous, differentiable functions of $\hat{O}_{r\perp}$, they have the same asymptotic ($F \to \infty$) properties as $\hat{A}$ and $\hat{C}$ in Appendix 7.R. The third step of both algorithms is identical and, therefore, the properties of $\hat{B}$ and $\hat{D}$ also remain the same.

To orthogonalize the data matrices, we use formulas (7-263) to (7-265) instead of (7-249) and (7-250). Note that the orthogonalization is done without calculating, explicitly, the matrices L_Y and L_U in (7-257). Because $\mathbf{Y}^{\mathrm{re}}\mathbf{Y}^{\mathrm{re}T} = (L_Y^T L_Y)^{-1}$ and $\mathbf{U}^{\mathrm{re}}\mathbf{U}^{\mathrm{re}T} = (L_U^T L_U)^{-1}$, the matrices L_Y and L_U have the same condition numbers as $\mathbf{Y}^{\mathrm{re}}$

and $\mathbf{U}^{\mathrm{re}}$, respectively. Therefore, L_Y and L_U should never be computed. They are used for theoretical derivations only. As $W_Y(k)$ is not explicitly calculated in (7-263) to (7-264), the covariance matrix $C_{\mathbf{Y}_\perp}$ cannot be calculated using (7-261). This is done as follows. Using $\mathbf{N}_{\mathbf{Y}_\perp} = L_Y\mathbf{N}_\mathbf{Y}$ we find

$$C_{\mathbf{Y}_\perp} = L_Y C_\mathbf{Y} L_Y^T = L_Y \mathrm{Re}(\mathbf{C}\mathbf{C}^H) L_Y^T = \mathrm{Re}(\mathbf{C}_\perp \mathbf{C}_\perp^H) \tag{7-272}$$

where $\mathbf{C} = [W_r(1)\sigma_Y(1) W_r(2)\sigma_Y(2) \ldots W_r(F)\sigma_Y(F)]$ and $\mathbf{C}_\perp = L_Y\mathbf{C}$. Because $\mathbf{C}$ has the shift property $\mathbf{C}_{[n+1,:]} = \mathbf{C}_{[n,:]}D_s$ and $\mathbf{C}_\perp = L_Y\mathbf{C}$, the relationship between the rows of $\mathbf{C}_\perp$ is given by (7-263), (7-264).

1. Initialization:

$$\mathbf{C}_{\perp[1,:]} = \mathbf{C}_{[1,:]}/\alpha_1, \ \mathbf{C}_{\perp[2,:]} = \mathbf{C}_{\perp[1,:]}D_s/\alpha_2 \tag{7-273}$$

2. Recursion: for $n = 3$ to r

$$\mathbf{C}_{\perp[n,:]} = (\mathbf{C}_{\perp[n-1,:]}D_s + \alpha_{n-1}\mathbf{C}_{\perp[n-2,:]})/\alpha_n \tag{7-274}$$

The proof follows exactly the same lines as for $O_{r\perp}$ in Section 7.S.3 of this appendix.

The matrix $\mathbf{Y}_\perp^{\mathrm{re}}\Pi_\perp$ is calculated using a QR decomposition of $Z_\perp^T = [\mathbf{U}_\perp^{\mathrm{re}T} \ \ \mathbf{Y}_\perp^{\mathrm{re}T}]$ as explained in Section 7.R.2 of Appendix 7.R.

Appendix 7.T: Sensitivity Estimates to Noise Model Errors

We study the influence of noise model errors on the bias of the GTLS, ML, and BTLS estimates assuming that a true plant model exists and that it belongs to the considered model set (Assumptions 7.16 and 7.17).

7.T.1 Bias of the ML, GTLS, and BTLS Estimators. If the wrong noise (co)variances are used, then Theorem 7.21, except properties 6 to 8 (consistency, bias, and efficiency), is still valid for the GTLS, ML, and BTLS estimates. The estimates are no longer consistent, $\tilde{\theta}(Z_0) \neq \theta_0$ $(\theta_* \neq \theta_0)$, because $\mathcal{E}\{v_F(\theta, N_Z)\}$ is no longer θ independent (use quick analysis tool number 2 of Section 7.5).

An explicit expression of the bias $\tilde{\theta}(Z_0) - \theta_0$ can be found for models (7-7) ($\Omega = z^{-1}$, s, $\sqrt{s}$, and $\tanh(\tau_R s)$) and (7-8) ($\Omega = z^{-1}$) through a Taylor series expansion of the expected value of the cost function at θ_0

$$V_F'^T(\theta) = V_F'^T(\theta_0) + V_F''(\widehat{\theta})(\theta - \theta_0) \tag{7-275}$$

with $\widehat{\theta} = t\theta + (1-t)\theta_0$ and $t \in [0, 1]$. Because $V_F'^T(\tilde{\theta}(Z_0)) = 0$, it follows from (7-275) that

$$\tilde{\theta}(Z_0) - \theta_0 = -V_F''^{-1}(\widehat{\theta})V_F'^T(\theta_0) \tag{7-276}$$

Under Assumptions 7.16 and 7.17 we have $\mathcal{E}\{\,V_F'(\theta_0, Z_0)\,\} = 0$, even if the wrong noise (co)variances are used (see (7-71), (7-79), and (7-97)). Together with (7-19), it facilitates rewriting (7-276) as

$$\tilde{\theta}(Z_0) - \theta_0 = -V_F{}''^{-1}(\widehat{\theta})\,\mathcal{E}\{v_F'(\theta_0, N_Z)\} \tag{7-277}$$

The GTLS, ML, and BTLS estimates are inconsistent because $\mathcal{E}\{\,v_F'(\theta_0, N_Z)\} \neq 0$ if the wrong noise (co)variances are used. The bias term (7-277) will be calculated explicitly for the ML estimator. From the ML cost (7-79) it follows that

$$\mathcal{E}\{\,v_F(\theta, N_Z)\,\} = \frac{1}{F}\sum\nolimits_{k=1}^{F}\frac{\sigma_e^2(\Omega_k, \theta)}{\hat{\sigma}_e^2(\Omega_k, \theta)} = \frac{1}{F}\sum\nolimits_{k=1}^{F}\frac{\sigma_{\bar{Y}}^2(\Omega_k, \theta)}{\hat{\sigma}_{\bar{Y}}^2(\Omega_k, \theta)} \tag{7-278}$$

where $\sigma_e^2(\Omega_k, \theta)$, $\sigma_{\bar{Y}}^2(\Omega_k, \theta)$, and $\hat{\sigma}_e^2(\Omega_k, \theta)$, $\hat{\sigma}_{\bar{Y}}^2(\Omega_k, \theta)$ are calculated as in (7-34), (7-44) using, respectively, the true and wrong noise (co)variances. The second equality in (7-278) is obtained via $\sigma_e^2(\Omega_k, \theta) = |A(\Omega_k, \theta)|^2\sigma_{\bar{Y}}^2(\Omega_k, \theta)$. Using

$$\begin{aligned}
G(\Omega_k, \theta_0) &= G_0(\Omega_k) = Y_0(k)/U_0(k)\\
\frac{\partial|G(\Omega_k, \theta)|^2}{\partial\theta_0} &= 2|G_0(\Omega_k)|^2\mathrm{Re}(\frac{\partial\ln(G(\Omega_k, \theta))}{\partial\theta_0})\\
\mathrm{Re}(\sigma_{\bar{Y}U}^2(k)\frac{\partial\bar{G}(\Omega_k, \theta)}{\partial\theta_0}) &= \mathrm{Re}(\bar{\sigma}_{\bar{Y}U}^2(k)G_0(\Omega_k)\frac{\partial\ln(G(\Omega_k, \theta))}{\partial\theta_0})
\end{aligned}$$

we get

$$\begin{aligned}
\frac{\partial\sigma_{\bar{Y}}^2(\Omega_k, \theta)}{\partial\theta_0} &= 2|Y_0(k)|^2\mathrm{Re}(V_U(k)\frac{\partial\ln(G(\Omega_k, \theta))}{\partial\theta_0})\\
V_U(k) &= \sigma_{\bar{U}}^2(k)/|U_0(k)|^2 - \bar{\sigma}_{\bar{Y}U}^2/(\bar{Y}_0(k)U_0(k))
\end{aligned} \tag{7-279}$$

Using (7-279) and

$$\begin{aligned}
\sigma_{\bar{Y}}^2(\Omega_k, \theta_0) &= |Y_0(k)|^2(V_U(k) + V_Y(k))\\
V_Y(k) &= \sigma_{\bar{Y}}^2(k)/|U_0(k)|^2 - \sigma_{\bar{Y}U}^2/(Y_0(k)\bar{U}_0(k))
\end{aligned} \tag{7-280}$$

the derivative of (7-278) w.r.t. θ at θ_0 equals

$$\mathcal{E}\{\,v_F'(\theta_0, N_Z)\,\} = \frac{2}{F}\sum_{k=1}^{F}\mathrm{Re}\left(\frac{\partial\ln(G(\Omega_k, \theta))}{\partial\theta_0}\frac{V_U(k)\hat{V}_Y(k) - \hat{V}_U(k)V_Y(k)}{(\hat{V}_U(k) + \hat{V}_Y(k))^2}\right) \tag{7-281}$$

where $\hat{V}_U(k)$, $\hat{V}_Y(k)$ equal $V_U(k)$, $V_Y(k)$ evaluated with the wrong noise (co)variances.

From (7-277) and (7-281), it follows that the bias is a function of the difference between the actual and the true noise (co)variances. The same is true for the GTLS and BTLS estimators (see Exercise 7.9). If the noise covariance matrix used $\hat{C}_{N_Z(k)}$ satisfies (7-128), then $\hat{V}_Y(k) = f(k)V_Y(k)$, $\hat{V}_U(k) = f(k)V_U(k)$ and the bias (7-277) is zero.

For model (7-8) with $\Omega = s$, the bias $\theta_* - \theta_0$ is calculated via a Taylor series expansion of $V_*'(\theta)$ at θ_0. Following the same lines as in the previous paragraph, we find

$$\theta_* - \theta_0 = -V_*''^{-1}(\widehat{\theta}) \lim_{F \to \infty} \mathcal{E}\{v_F'(\theta_0, N_Z)\} \tag{7-282}$$

with $\widehat{\theta} = t\theta_* + (1-t)\theta_0$ and $t \in [0, 1]$. Comparing this expression with (7-277), it follows that the conclusions of the previous paragraph also apply to the bias $\theta_* - \theta_0$ of model (7-8) with $\Omega = s$.

7.T.2 Bias of the NLS-FRF and LOG Estimators. The NLS-FRF (7-46) and LOG estimators (7-54) apply to model (7-7). The bias $\tilde{\theta}(Z_0) - \theta_0$ calculation follows the same lines as in the previous section. Therefore, expression (7-276) for the bias is valid with

$$V_F'(\theta_0) = -\frac{2}{F}\sum_{k=1}^{F} \mathrm{Re}\Big(\frac{\partial \ln(G(\Omega_k, \theta))}{\partial \theta_0}\bar{b}(k)g(k)\Big) \tag{7-283}$$

where $g(k) = 1$, $b(k) = \mathcal{E}\{N(k)\}$ and $N(k)$ is defined in (7-55) for the LOG estimator, and $g(k) = |G_0(\Omega_k)|^2$, $b(k) = \mathcal{E}\{N_G(k)\}/G_0(\Omega_k)$ and $N_G(k)$ is defined in (7-47) for the NLS-FRF estimator. For circular complex normally distributed input-output errors, the bias $b(k)$ is given by (7-49) and (7-56). For circular complex noise with even pdf, the bias is a function of the fourth-order moments of the noise (see Appendix 7.G). If the noise is not circular complex then $b(k)$ is a function of the second-order moments of the noise (put $\mathcal{E}\{z^2\} \neq 0$ in (7-177)). □

Appendix 7.U: IWLS Solution in Case of Vector Orthogonal Polynomials

Using the inner product (7-137), the orthogonality condition (7-138), and

$$\begin{bmatrix} A(\Omega, \theta^{(i)}) \\ B(\Omega, \theta^{(i)}) \end{bmatrix} = \sum_{r=0}^{n_a+n_b+1} a_r \begin{bmatrix} p_r(\Omega) \\ q_r(\Omega) \end{bmatrix} \tag{7-284}$$

with $a_{n_a+n_b+1} = 1$, the cost function (7-135) can be written as

$$\left\langle \begin{bmatrix} A(\Omega, \theta^{(i)}) \\ B(\Omega, \theta^{(i)}) \end{bmatrix}, \begin{bmatrix} A(\Omega, \theta^{(i)}) \\ B(\Omega, \theta^{(i)}) \end{bmatrix} \right\rangle = 1 + \sum_{r=0}^{n_a+n_b} a_r^2 \tag{7-285}$$

It follows directly that (7-285) is minimal for $a_0 = a_1 = \ldots = a_{n_a+n_b} = 0$. □

Appendix 7.V: Asymptotic Properties in the Presence of Nonlinear Distortions

Consider the errors-in-variables model (7-145) where $M_U(k)$, $M_Y(k)$, and $N_P(k)$ satisfy the noise assumptions of a time-domain experiment in Section 7.6. Multiplying (7-145) by $e^{-j\angle U_0(k)}$ gives

$$\begin{aligned} Y(k)e^{-j\angle U_0(k)} &= G_R(s_k)|U_0(k)| + N_Y(k)e^{-j\angle U_0(k)} \\ U(k)e^{-j\angle U_0(k)} &= |U_0(k)| + N_U(k)e^{-j\angle U_0(k)} \end{aligned} \tag{7-286}$$

for $k = 1, 2, \ldots, F$. Note that this phase shift, applied to model (7-7) or (7-9), does not change any of the cost functions of Sections 7.8 to 7.14. To prove that Theorem 7.21, with $G_0(s)$ replaced by $G_R(s)$, is valid, it is sufficient to show that the noisy part $N_Z(k)e^{-j\angle U_0(k)}$ of (7-286) satisfies all the assumptions of Section 7.6. First, note that $N_Z(k)e^{-j\angle U_0(k)}$ is independent of the true unknown excitation $|U_0(k)|$. Because $M_U(k)$, $M_Y(k)$ and $N_P(k)$ are independent of $Y_S(k)$ and $U_0(k)$, it follows that $M_U(k)e^{-j\angle U_0(k)}$, $M_Y(k)e^{-j\angle U_0(k)}$, and $N_P(k)e^{-j\angle U_0(k)}$ satisfy the noise assumptions of Section 7.6. $Y_S(k)e^{-j\angle U_0(k)}$ has the same phase as $G_S(k)$ and, therefore, has the same stochastic properties as $G_S(k)$ in Theorems 3.10 (mixing of order infinity) and 3.11 (asymptotic normality). As $Y_S(k)e^{-j\angle U_0(k)}$ is mixing of order four (infinity), all properties of Theorem 7.21 remain valid (proof: see introduction of Appendix 7.E). □

Appendix 7.W: Consistency of the Missing Data Problem

The consistency proof follows the lines of the proof of Theorem 7.21 (Appendix 7.E). We first show the result for discrete-time systems (5-50) and afterward for continuous-time systems (5-49). Because the output error $Y^m(k) - Y^m(\Omega_k, \Theta)$ is linear in the missing samples ψ, we can rewrite the cost function (7-147) as

$$V(\theta, \psi, Z^m) = \frac{1}{2}(\varepsilon_1(\theta, Z^m) + \varepsilon_2(\theta)\psi)^T(\varepsilon_1(\theta, Z^m) + \varepsilon_2(\theta)\psi) \tag{7-287}$$

where $\varepsilon_1(\theta, Z^m) \in \mathbb{R}^N$ is a linear function of the missing data set Z^m, and $\varepsilon_2(\theta) \in \mathbb{R}^{N \times (M_U + M_y)}$ is independent of Z^m. Elimination of ψ in (7-287) gives

$$\psi(\theta, Z^m) = -(\varepsilon_2^T(\theta)\varepsilon_2(\theta))^{-1}\varepsilon_2^T(\theta)\varepsilon_1(\theta, Z^m) \tag{7-288}$$

$$V(\theta, Z^m) = \frac{1}{2}\varepsilon_1^T(\theta, Z^m)P(\theta)\varepsilon_1(\theta, Z^m) \tag{7-289}$$

with $P(\theta) = I_N - \varepsilon_2(\theta)(\varepsilon_2^T(\theta)\varepsilon_2(\theta))^{-1}\varepsilon_2^T(\theta)$ a symmetric idempotent matrix of rank $N - M_u - M_y$. As $\varepsilon_2(\theta)$ lies in the null space of $P(\theta)$, (7-289) can be written as

$$V(\theta, Z^m) = \frac{1}{2}(\varepsilon_1(\theta, Z^m) + \varepsilon_2(\theta)\varphi_Z)^T P(\theta)(\varepsilon_1(\theta, Z^m) + \varepsilon_2(\theta)\varphi_Z) \tag{7-290}$$

with φ_Z the vector ψ (5-48) where the missing input and output samples are replaced by the disturbing noise on these samples. To simplify the notations, we will assume without any loss of generality that the excitation is deterministic. Because the noisy part of $\varepsilon_1(\theta, Z^{\mathrm{m}}) + \varepsilon_2(\theta)\varphi_Z$ contains the DFT spectra of the complete disturbing noise sequences (no missing samples) and the output error $Y^{\mathrm{m}}(k) - Y^{\mathrm{m}}(\Omega_k, \Theta)$ in (7-147) is divided by $\sigma_Y(\Omega_k, \theta)$ (7-44), which contains the (co)variances of the complete noise sequences, we have $\mathrm{Cov}(\varepsilon_1(\theta, Z^{\mathrm{m}}) + \varepsilon_2(\theta)\varphi_Z) = I_N$. Using this result together with $\mathrm{trace}(P(\theta)) = N - M_u - M_y$, the expected value of (7-290) equals (see Exercise 15.2)

$$\mathcal{E}\{\, V(\theta, Z^{\mathrm{m}}) \,\} = \frac{1}{2}\varepsilon_1^T(\theta, Z_0^{\mathrm{m}})P(\theta)\varepsilon_1(\theta, Z_0^{\mathrm{m}}) + \frac{1}{2}(N - M_u - M_y) \tag{7-291}$$

with Z_0^{m} the true (noiseless) missing data set. Under Assumption 7.16 we have

$$P(\theta_0)\varepsilon_1(\theta_0, Z_0^{\mathrm{m}}) = \varepsilon_1(\theta_0, Z_0^{\mathrm{m}}) + \varepsilon_2(\theta_0)\,\psi(\theta_0, Z_0^{\mathrm{m}}) = \varepsilon_1(\theta_0, Z_0^{\mathrm{m}}) + \varepsilon_2(\theta_0)\,\psi_0 = 0 \tag{7-292}$$

so that θ_0 minimizes the expected value of the cost function (7-291). Since $P(\theta) = V\Sigma V^T$, with V an orthogonal matrix and

$$\Sigma = \begin{bmatrix} I_{N - M_u - M_y} & 0 \\ 0 & 0 \end{bmatrix}$$

(see Exercise 13.19), the cost function (7-290) can be written as

$$\begin{aligned} V(\theta, Z^{\mathrm{m}}) &= \frac{1}{2}\varepsilon_3^T(\theta, Z^{\mathrm{m}})\varepsilon_3(\theta, Z^{\mathrm{m}}) \\ \varepsilon_3(\theta, Z^{\mathrm{m}}) &= [I_{N - M_U - M_Y}\ 0]V^T(\varepsilon_1(\theta, Z^{\mathrm{m}}) + \varepsilon_2(\theta)\varphi_Z) \end{aligned} \tag{7-293}$$

Under Assumptions 7.18 and 7.19, the entries of the vector $\varepsilon_3(\theta, Z^{\mathrm{m}})$ are independent Gaussian random variables with variance 1, so that $V(\theta, Z^{\mathrm{m}})/N$ converges ($N \to \infty$) w.p. 1 to its expected value (see Section 14.9, version 2 of the strong law of large numbers). Under Assumption 7.6, applied to $V(\theta, Z^{\mathrm{m}})$ (7-293), this convergence is uniform in a closed and bounded neighborhood of θ_0. Under Assumption 7.7, applied to $V(\theta, Z^{\mathrm{m}})$ (7-293), this implies the strong convergence of the minimizer $\hat{\theta}_{\mathrm{WNLS}}(Z^{\mathrm{m}})$ of (7-293) to the minimizer, θ_0, of (7-291) (see Appendix 15.B). The true coefficients of the $T(\Omega, \theta)$ polynomial in (7-146) are asymptotically zero (Lemma 5.5), thus only the plant model parameters $a_0 a_1 \ldots a_{n_a} b_0 b_1 \ldots b_{n_b}$ in θ are consistently estimated. Note also that the estimate of the missing data $\hat{\psi}_{\mathrm{WNLS}} = \psi(\hat{\theta}_{\mathrm{WNLS}}(Z^{\mathrm{m}}), Z^{\mathrm{m}})$ (see (7-288)) is inconsistent.

The proof for continuous-time systems (5-49) follows exactly the same lines. The only differences are that θ_0 minimizes the limit cost function $\lim_{N \to \infty} \mathcal{E}\{\, V(\theta, Z^{\mathrm{m}}) \,\}/N$ instead of (7-291) and that $V(\theta, Z^{\mathrm{m}})/N$ convergences weakly, instead of strongly, to its expected value. This is due to the presence of the alias term $\delta(s_k)$ in the true output observations (see model (5-49)) which is only in probability ($N \to \infty$) zero (Lemma 5.6). □

Appendix 7.X: Normal Equation for Complex Parameters and Analytic Residuals

Because $\varepsilon(\theta, Z)$ is an analytic function of θ we have

$$\frac{\partial\varepsilon(\theta, Z)}{\partial\mathrm{Re}(\theta)} = \frac{\partial\varepsilon(\theta, Z)}{\partial\theta} \text{ and } \frac{\partial\varepsilon(\theta, Z)}{\partial\mathrm{Im}(\theta)} = j\frac{\partial\varepsilon(\theta, Z)}{\partial\theta} \tag{7-294}$$

so that

$$\begin{aligned}
\frac{\partial\mathrm{Re}(\varepsilon(\theta, Z))}{\partial\mathrm{Re}(\theta)} &= \mathrm{Re}\left(\frac{\partial\varepsilon(\theta, Z)}{\partial\theta}\right) \\
\frac{\partial\mathrm{Im}(\varepsilon(\theta, Z))}{\partial\mathrm{Re}(\theta)} &= \mathrm{Im}\left(\frac{\partial\varepsilon(\theta, Z)}{\partial\theta}\right) \\
\frac{\partial\mathrm{Re}(\varepsilon(\theta, Z))}{\partial\mathrm{Im}(\theta)} &= \mathrm{Re}\left(j\frac{\partial\varepsilon(\theta, Z)}{\partial\theta}\right) = -\mathrm{Im}\left(\frac{\partial\varepsilon(\theta, Z)}{\partial\theta}\right) \\
\frac{\partial\mathrm{Im}(\varepsilon(\theta, Z))}{\partial\mathrm{Im}(\theta)} &= \mathrm{Im}\left(j\frac{\partial\varepsilon(\theta, Z)}{\partial\theta}\right) = \mathrm{Re}\left(\frac{\partial\varepsilon(\theta, Z)}{\partial\theta}\right)
\end{aligned} \tag{7-295}$$

Equations (7-295) are known as the Cauchy-Riemann conditions of an analytic function (Henrici, 1974). Using (7-295) and definition (13-40), we find

$$\frac{\partial\varepsilon_{\mathrm{re}}(\theta, Z)}{\partial\theta_{\mathrm{re}}} = \begin{bmatrix} \dfrac{\partial\mathrm{Re}(\varepsilon(\theta, Z))}{\partial\mathrm{Re}(\theta)} & \dfrac{\partial\mathrm{Re}(\varepsilon(\theta, Z))}{\partial\mathrm{Im}(\theta)} \\ \dfrac{\partial\mathrm{Im}(\varepsilon(\theta, Z))}{\partial\mathrm{Re}(\theta)} & \dfrac{\partial\mathrm{Im}(\varepsilon(\theta, Z))}{\partial\mathrm{Im}(\theta)} \end{bmatrix} = \left(\frac{\partial\varepsilon(\theta, Z)}{\partial\theta}\right)_{\mathrm{Re}} \tag{7-296}$$

Applying (7-296) and Lemma 13.4 to the right-hand side of (7-152) gives

$$J_{\mathrm{re}}(\theta_{\mathrm{re}}^{(i-1)}, Z)\Delta\theta_{\mathrm{re}}^{(i)} = (J(\theta^{(i-1)}, Z))_{\mathrm{Re}}\Delta\theta_{\mathrm{re}}^{(i)} = (J(\theta^{(i-1)}, Z)\Delta\theta^{(i)})_{\mathrm{re}} \tag{7-297}$$

(7-297) together with $\varepsilon_{\mathrm{re}}(\theta_{\mathrm{re}}^{(i-1)}, Z) = (\varepsilon(\theta^{(i-1)}, Z))_{\mathrm{re}}$ shows that (7-152) is equivalent with (7-153). □

Appendix 7.Y: Total Least Squares for Complex Parameters

From Appendix 7.X, it follows that

$$J_{\mathrm{re}}(Z) = \frac{\partial\varepsilon_{\mathrm{re}}(\theta_{\mathrm{re}}, Z)}{\partial\theta_{\mathrm{re}}} = \left(\frac{\partial\varepsilon(\theta, Z)}{\partial\theta}\right)_{\mathrm{Re}} = (J(Z))_{\mathrm{Re}} \tag{7-298}$$

Using (7-298) and Lemma 13.4 we find

$$\begin{aligned} W_{\mathrm{Re}} J_{\mathrm{re}}(Z) &= (WJ(Z))_{\mathrm{Re}} \\ C^T C &= \mathcal{E}\{ W_{\mathrm{Re}}^T j_{\mathrm{re}}^T(N_Z) j_{\mathrm{re}}(N_Z) W_{\mathrm{Re}} \} = (\mathcal{E}\{ W^H j^H(N_Z) j(N_Z) W \})_{\mathrm{Re}} = (C_c^H C_c)_{\mathrm{Re}} \end{aligned} \tag{7-299}$$

with $j(N_z) = J(Z) - J(Z_0)$, and a possible choice for C is $C = (C_c)_{\mathrm{Re}}$. The total least squares solution is calculated via the GSVD of the real matrix pair $(W_{\mathrm{Re}} J_{\mathrm{re}}(Z), C)$ (see Section 7.10). Because $W_{\mathrm{Re}} J_{\mathrm{re}}(Z) = (WJ(Z))_{\mathrm{Re}}$ and $C = (C_c)_{\mathrm{Re}}$, it can also be calculated via the GSVD of the complex matrix pair $(WJ(Z), C_c)$ (see Section 13.8). □

8

Estimation with Unknown Noise Model

Abstract: In the identification schemes that were presented in the previous chapters, it was assumed that the covariance matrix of the noise is known a priori. In practice this information should also be extracted from the experimental data. In this chapter, it is shown that a utilizable nonparametric frequency domain noise model can be obtained from a very small number of repeated experiments. Under these conditions the consistency of the estimates is maintained, while the loss in efficiency is small. Also the classical solution for identifying a parametric noise model together with the plant model is discussed.

8.1 INTRODUCTION

In Chapter 7 a large variety of estimators were discussed, ranging from unweighted linear least squares methods to maximum likelihood estimators. The more advanced estimators such as Markov, GTLS, BTLS, and ML estimators require knowledge of the covariance matrix with the disturbing noise as a function of the frequency. For example, the ML estimator was given as the minimizer of (7-79) which reduces to (8-1) if no transients are added to the model:

$$V_{\text{ML}}(\theta, Z) = \sum_{k=1}^{F} \frac{|e(\Omega_k, \theta, Z(k))|^2}{\sigma_e^2(\Omega_k, \theta)} = \sum_{k=1}^{F} |\varepsilon(\Omega_k, \theta, Z(k))|^2 \tag{8-1}$$

with $e(\Omega_k, \theta, Z(k)) = A(\Omega_k, \theta)Y(k) - B(\Omega_k, \theta)U(k)$, $\sigma_e^2(\Omega_k, \theta) = \text{var}(e(\Omega_k, \theta, N_Z(k)))$,

$$\sigma_e^2(\Omega_k, \theta) = \sigma_Y^2(k)|A(\Omega_k, \theta)|^2 + \sigma_U^2(k)|B(\Omega_k, \theta)|^2 - 2\text{Re}(\sigma_{YU}^2(k)A(\Omega_k, \theta)\bar{B}(\Omega_k, \theta)) \tag{8-2}$$

and $\varepsilon(\Omega_k, \theta, Z(k)) = e(\Omega_k, \theta, Z(k))/\sigma_e(\Omega_k, \theta)$. The noise (co)variances $\sigma_U^2(k)$, $\sigma_Y^2(k)$, and $\sigma_{YU}^2(k)$ were assumed to be known exactly, and under these conditions the properties of the estimators were studied. In practice, this information is not available but should be extracted from the experimental data. In this chapter we will replace the exact noise (co)variances by their

System Identification: A Frequency Domain Approach. Rik Pintelon and Johan Schoukens
Copyright © 2001 Institute of Electrical and Electronics Engineers. ISBN: 0-780-36000-1

sample values. This is possible only if independent, repeated experiments are available. A practical solution consists of applying periodic excitations to the plant and observing M consecutive periods of the steady-state response. Therefore, the simple plant model (7-7) will be used throughout this chapter

$$Y(\Omega_k, \theta) = G(\Omega_k, \theta)U(k) \tag{8-3}$$

with $\Omega = z^{-1}$, s, $\sqrt{s}$, or $\tanh(\tau_R s)$. The M experiments are processed and their DFT spectra

$$U^{[l]}(k), Y^{[l]}(k),\ l = 1, \ldots, M \text{ and } k = 1, \ldots, F \tag{8-4}$$

are calculated as explained in Chapter 2. The sample (co)variances are obtained directly from these measurements, e.g.,

$$\hat{\sigma}_U^2(k) = \frac{1}{M-1}\sum\nolimits_{l=1}^{M}\left|U^{[l]}(k) - \hat{U}(k)\right|^2, \text{ with } \hat{U}(k) = \frac{1}{M}\sum\nolimits_{l=1}^{M} U^{[l]}(k) \tag{8-5}$$

(see (2-31)). These values are used in (8-1) and (8-2) instead of the exact values. To compare this approach with the classical framework that deals with arbitrary excitations (Ljung, 1999), we have to simplify the errors-in-variables framework to a weighted output error problem. This means that only process noise is considered; the measurement noise on the input and the output is assumed to be zero ($\sigma_U^2(k) = 0$ and also $\sigma_{YU}^2(k) = 0$) so that the cost function (8-1) reduces to

$$V_{\text{ML}}(\theta, Z) = \sum_{k=1}^{F} \frac{\left|Y(k) - \dfrac{B(\Omega_k, \theta)}{A(\Omega_k, \theta)}U(k)\right|^2}{\sigma_Y^2(k)} = \sum_{k=1}^{F} \frac{|Y(k) - G(\Omega_k, \theta)U(k)|^2}{\sigma_Y^2(k)} \tag{8-6}$$

Because in this classical framework no repeated measurements are imposed, the sample variance $\hat{\sigma}_Y^2(k)$ cannot be calculated. Instead a parametric noise model $\sigma_Y^2(k) = \sigma^2\left|H(z_k^{-1}, \theta)\right|^2$ is used and the additional noise model parameters are estimated together with the plant model parameters (Ljung, 1999). This poses the question of what approach should be preferred: the parametric or the nonparametric (sample (co)variances) noise modeling approach?

The major advantage of the parametric modeling approach is its applicability to arbitrary excitations. Its major disadvantage is the need for a double model selection problem (plant model and noise model) and a more complex optimization problem. The reader is referred to Ljung (1999) for a comprehensive discussion of these techniques. A brief discussion is also given in this chapter, Section 8.9.

The major disadvantages of the nonparametric approach are the restriction to periodic excitations, and the loss in frequency resolution of a factor M w.r.t. the parametric approach. However, whenever periodic excitations can be applied, significant advantages appear: the nonparametric model is generated automatically, without any user interaction; the errors-in-variables problem can be solved straightforwardly (no equivalent solution is available in the classical approach); the cost function is absolutely interpretable, which simplifies the validation process significantly (see Chapter 9). For these reasons, we prefer to use the nonparametric noise models whenever it is possible to apply periodic excitations, independent of the fact that a time or frequency domain method will be used later on.

8.2 DISCUSSION OF THE DISTURBING NOISE ASSUMPTIONS

8.2.1 Assuming Independent Normally Distributed Noise for Time Domain Experiments

Actually, we will prove the theorems under the frequency domain experiment Assumption 7.18 and Assumption 7.19 (assuming independently normally distributed noise in the frequency domain) while in practice the data are obtained from a time domain experiment (where it is assumed that the disturbing noise is described by a filtered white noise source). Switching to the idealized frequency domain assumptions makes it possible to set up a formal theory to analyze the replacement of the exact variances by their sample values. Moreover, this mixed use of both assumptions is supported by Theorem 14.25 and will be further discussed later.

The following assumptions are necessary to study the asymptotic behavior ($F \to \infty$) of the estimators. First, we require that M independent repeated experiments are available. Next, we make an assumption about the disturbing errors of the lth experiment.

Assumption 8.1 (*M* Independent Repeated Experiments): The measured input-output DFT spectra $U^{[l]}(k)$, $Y^{[l]}(k)$, $k = 1, 2, \ldots, F$ and $l = 1, 2, \ldots, M$, satisfy

$$\begin{aligned} Y^{[l]}(k) &= Y_0(k) + N_Y^{[l]}(k) \\ U^{[l]}(k) &= U_0(k) + N_U^{[l]}(k) \end{aligned} \tag{8-7}$$

where the true unknown deterministic values $U_0(k)$, $Y_0(k)$ are independent of l and where the disturbing input-output errors $N_U^{[l]}(k)$, $N_Y^{[l]}(k)$ are independent over l.

Following the lines of Chapter 7 we introduce the data vector $Z^{[l]}$ (see (7-3))

$$Z^{[l]T} = [Z^{[l]T}(1)Z^{[l]T}(2)\ldots Z^{[l]T}(F)] \text{ with } Z^{[l]T}(k) = [Y^{[l]}(k)\ U^{[l]}(k)] \tag{8-8}$$

and similarly for $N_Z^{[l]}$. It is related to the true values by $Z^{[l]} = Z_0 + N_Z^{[l]}$.

Assumption 8.2 (Zero Mean Normally Distributed Errors): The noise $N_Z^{[l]}(k)$ is independent over the frequency k, has zero mean, and is circular complex normally distributed with covariance matrix

$$C_{N_Z}(k) = \mathcal{E}\{N_Z^{[l]}(k)(N_Z^{[l]}(k))^H\} = \begin{bmatrix} \sigma_Y^2(k) & \sigma_{YU}^2(k) \\ \bar{\sigma}_{YU}^2(k) & \sigma_U^2(k) \end{bmatrix} \tag{8-9}$$

Discussion

In the theory that is set up below, the normal distribution of the noise will be a kernel property. It is asymptotically guaranteed by Theorem 14.25.

It is assumed that the noise is independent from one frequency to the other. Again, this property is only asymptotically met in practice, so that in principle a full covariance matrix should be used, including the covariance over different frequencies. However, this would make the nonparametric approach very intractable because the large full matrix has to be inverted to calculate the cost function. It is shown that under the time domain experiment Assumption 7.3, the nondiagonal terms can be omitted without affecting the asymptotic properties of the estimates, so that the cost function (8-1) still can be used (Schoukens et al., 1999a).

8.2.2 Considering Successive Periods as Independent Realizations

The noise behavior is characterized using the sample mean and sample variance, obtained from a set of repeated measurements. In practice we often obtain these repeated measurements by measuring M successive periods in one record. For each period, we calculate the Fourier coefficients and consider them as independent experiments from one period to the other as formalized in Assumption 8.1. Again, this is only approximately met in practice because some correlation exists between neighboring periods. Because the correlation of filtered white noise (time domain experiment assumption) decays exponentially, the correlation between two neighboring periods disappears in inverse proportion to the length of the period. In practice it can be neglected if the period length is large compared with the correlation length of the noise.

8.3 PROPERTIES OF THE ML ESTIMATOR USING A SAMPLE COVARIANCE MATRIX

8.3.1 The Sample Maximum Likelihood Estimator: Definition of the Cost Function

A new cost function is defined putting $\hat{U}(k)$, $\hat{Y}(k)$ and $\hat{\sigma}_{\hat{U}}^2(k)$, $\hat{\sigma}_{\hat{Y}}^2(k)$, $\hat{\sigma}_{\hat{Y}U}^2(k)$ as the measurements and the variances, respectively, into the cost function (8-1):

$$V_{\mathrm{SML}}(\theta, Z) = \sum_{k=1}^{F} \frac{\left|\hat{e}(\Omega_k, \theta, \hat{Z}(k))\right|^2}{\hat{\sigma}_{\hat{e}}^2(\Omega_k, \theta)} = \sum_{k=1}^{F} \hat{\varepsilon}(\Omega_k, \theta, \hat{Z}(k)) \tag{8-10}$$

with $\hat{\varepsilon}(\Omega_k, \theta, \hat{Z}(k)) = \hat{e}(\Omega_k, \theta, \hat{Z}(k))/\hat{\sigma}_{\hat{e}}(\Omega_k, \theta)$ and

$$\begin{aligned}
\hat{e}(\Omega_k, \theta, \hat{Z}(k)) &= A(\Omega_k, \theta)\hat{Y}(k) - B(\Omega_k, \theta)\hat{U}(k) \\
\hat{\sigma}_{\hat{e}}^2(\Omega_k, \theta) &= \hat{\sigma}_e^2(\Omega_k, \theta)/M \\
\hat{\sigma}_e^2(\Omega_k, \theta) &= \hat{\sigma}_{\hat{Y}}^2(k)|A(\Omega_k, \theta)|^2 + \hat{\sigma}_{\hat{U}}^2(k)|B(\Omega_k, \theta)|^2 - 2\mathrm{Re}(\hat{\sigma}_{\hat{Y}U}^2(k)A(\Omega_k, \theta)\bar{B}(\Omega_k, \theta))
\end{aligned} \tag{8-11}$$

$\hat{\sigma}_e^2(\Omega_k, \theta) = \mathrm{var}(\hat{e}(\Omega_k, \theta, N_Z^{[l]}(k)))$ stands for the variance of the equation error of one experiment, while $\hat{\sigma}_{\hat{e}}^2(\Omega_k, \theta) = \mathrm{var}(\hat{e}(\Omega_k, \theta, \hat{N}_Z(k)))$ is the variance of the sample mean of the equation error.

8.3.2 Properties of the Sample Maximum Likelihood Estimator

The most important concern, when replacing the exact noise (co)variances by their sample values, is the loss in quality of the new estimator $\hat{\theta}_{\text{SML}}(Z)$ with respect to the original estimate $\hat{\theta}_{\text{ML}}(Z)$ due to this change. It turns out that this loss is small, even for a very small number of periods, typically 4 or 7. It will be shown that the sample estimate $\hat{\theta}_{\text{SML}}(Z)$ converges asymptotically $(F \rightarrow \infty)$ to $\hat{\theta}_{\text{ML}}(Z)$ (the estimate obtained with the exact noise (co)-variances). Also the loss in efficiency is small. The covariance matrix of the estimates grows with a factor $(M-2)/(M-3)$. These results are formulated precisely in the next two theorems. The first theorem gives a precise formulation of the properties of the sample estimate. The second describes the relationship between the "sample" estimate and the "exact" estimate.

Theorem 8.3 (Asymptotic Properties $\hat{\theta}_{\text{SML}}(Z)$): Consider model (8-3) with any identifiable parameterization of Section 5.2. Under the assumptions of Section 7.6, Assumptions 8.1 $(M \geq 4)$, and Assumption 8.2 the minimizer $\hat{\theta}_{\text{SML}}(Z)$ of (8-10) has the asymptotic properties of Theorem 7.21 with $V_F(\theta, Z) = V_{\text{SML}}(\theta, Z)/F$. For

1. $M \geq 4$ the stochastic and the deterministic convergence (properties 1, 5, and 6 of Theorem 7.21) are valid.
2. $M \geq 6$ the stochastic convergence rate (properties 2 and 6 of Theorem 7.21) are valid.
3. $M \geq 7$ the systematic and stochastic errors, the asymptotic normality, and the asymptotic bias (properties 3, 4, 6 and 7 of Theorem 7.21) are valid.

Proof. Apply Theorem 7.21 to (8-10), using the results of Appendix 8.D (which guarantees that all moments that appear in the proof exist). □

Theorem 8.4 (Relationship between $\hat{\theta}_{\text{SML}}(Z)$ and $\hat{\theta}_{\text{ML}}(Z)$): Under the conditions of Theorem 8.3, the estimates based on the true (8-1) and the sample (8-10) noise (co)variances are related to each other by:

1. For $M \geq 3$, the expected value of the cost functions,

$$V_{\text{SML}}(\theta) = \frac{M-1}{M-2} V_{\text{ML}}(\theta), \tag{8-12}$$

2. For $M \geq 4$, the asymptotic value of the cost functions,

$$V_{*\text{SML}}(\theta) = \frac{M-1}{M-2} V_{*\text{ML}}(\theta) \tag{8-13}$$

the minimizer of the expected value of the cost functions,

$$\tilde{\theta}_{\text{SML}}(Z_0) = \tilde{\theta}_{\text{ML}}(Z_0) \tag{8-14}$$

and the minimizer of the asymptotic value of the cost functions,

$$\theta_{*\text{SML}} = \theta_{*\text{ML}} \tag{8-15}$$

3. For $M \geq 7$, the parameter uncertainty in the absence on modeling errors, $\tilde{\theta}_{\text{SML}}(Z_0) = \theta_0$,

$$\mathrm{Cov}(\delta_{\theta\mathrm{SML}}(Z)) \approx \frac{M-2}{M-3}\mathrm{Cov}(\delta_{\theta\mathrm{ML}}(Z)), \tag{8-16}$$

where $\hat{\theta}(Z) = \theta_0 + \delta_\theta(Z) + O_\mathrm{p}(F^{-1})$ with $\mathcal{E}\{\delta_\theta(Z)\} = 0$ and $\delta_\theta(Z) = O_\mathrm{p}(F^{-1/2})$, and where $\sqrt{F}\delta_\theta(Z)$ is asymptotically normally distributed.

Proof. See the proof of Theorem 8.3 and Appendices 8.B, 8.F. □

The full proofs of both theorems are in the appendices, but the basic idea is easy to grasp. Consider

$$\begin{aligned} V_{\mathrm{SML}}(\theta, Z) &= \sum_{k=1}^{F} \frac{|\hat{e}(\Omega_k, \theta, \hat{Z}(k))|^2}{\hat{\sigma}_{\hat{e}}^2(\Omega_k, \theta)} = \sum_{k=1}^{F} c_k(\theta) d_k(\theta) \\ c_k(\theta) &= \sigma_{\hat{e}}^2(\Omega_k, \theta)/\hat{\sigma}_{\hat{e}}^2(\Omega_k, \theta) \\ d_k(\theta) &= |\hat{e}(\Omega_k, \theta, \hat{Z}(k))|^2/\sigma_{\hat{e}}^2(\Omega_k, \theta) \end{aligned} \tag{8-17}$$

Observe that $c_k(\theta)$ and $d_k(\theta)$ are independently distributed: the first depends on the sample variance, while the second depends on the sample mean. It is well known that these are independent random variables for normally distributed noise (Stuart and Ord, 1987). In Appendix 8.B it is shown that $\mathcal{E}\{c_k(\theta)\} = (M-1)/(M-2)$, for any $\theta \in \Theta_\mathrm{r}$, so that

$$\mathcal{E}\{V_{\mathrm{SML}}(\theta, Z)\} = \mathcal{E}\{c_k(\theta)\}\mathcal{E}\{V_{\mathrm{ML}}(\theta, Z)\} = \frac{M-1}{M-2}\mathcal{E}\{V_{\mathrm{ML}}(\theta, Z)\} \tag{8-18}$$

This shows that the minimizer of the expected value of the cost is not changed by introducing the sample variance, and this result is the kernel of the classical consistency proof.

8.3.3 Discussion

From Theorem 8.3, it follows that the estimate $\hat{\theta}_{\mathrm{SML}}(Z)$ is consistent and that $\tilde{\theta}_{\mathrm{SML}}(Z_0)$, $\theta_{*\mathrm{SML}}$ are the noiseless solutions when model errors are present (apply quick analysis tools number 2 and 3 of Section 7.5).

Because $V_{\mathrm{SML}}(\lambda\theta, Z) = V_{\mathrm{SML}}(\theta, Z)$, the estimate $\hat{\theta}_{\mathrm{SML}}(Z)$ is independent of the particular constraint $a_i = 1$, $b_i = 1$, or $\|\theta\|_2^2 = 1$ chosen (quick tool number 4).

If the input and output measurements are noisy, the sample covariance $\hat{\sigma}_{\hat{Y}U}^2(k)$ must be calculated, even if it is known that the input and output errors are uncorrelated $\sigma_{\hat{Y}U}^2(k) = 0$. Otherwise the properties of Theorems 8.3 and 8.4 are no longer valid (see the proof of Theorem 8.3). Both theorems show that in practice the unknown exact noise (co)variances can be replaced by the sample noise (co)variances without any problem.

It is not necessary to make a precise measurement, $M = 4$ independent repeated measurements suffice to get consistency, and $M = 7$ independent repeated measurements are enough to guarantee the existence of the covariance matrix of the limiting parameter distribution.

The loss in efficiency is not large (below 12% for $M = 7$) so that this is a very acceptable solution.

In practice it can sometimes be a problem to measure seven periods, especially when the period length becomes very long as is necessary to access very low frequencies. In that case the number of periods can be restricted to $M = 2$, but at that moment the variances should be averaged over seven neighboring frequency lines.

8.3.4 Estimation of Covariance Matrix of the Model Parameters

Equation (8-16) quantifies the loss in efficiency due to the use of the sample variances. However, it does not give an answer about how to calculate $\hat{C}_{\hat{\theta}}$ from the available information. C_{θ} is approximated by

$$\mathrm{Cov}(\underline{\hat{\theta}}_{\mathrm{ML}}(Z)) \approx [2\mathrm{Re}((\varepsilon'(\underline{\hat{\theta}}_{\mathrm{ML}}(Z), Z))^{H}(\varepsilon'(\underline{\hat{\theta}}_{\mathrm{ML}}(Z), Z)))]^{-1}$$

(see Section 7.11.4) and in practice, during the calculations of the covariance matrix, the exact variances in ε' are again replaced by the sample variances, and only $\hat{\varepsilon}'$ is available. Using similar calculations as in (8-18), it turns out that

$$\begin{aligned}&[2\mathrm{Re}((\varepsilon'(\underline{\hat{\theta}}_{\mathrm{ML}}(Z), Z))^{H}(\varepsilon'(\underline{\hat{\theta}}_{\mathrm{ML}}(Z), Z)))]^{-1} \approx \\ &\qquad \frac{M-1}{M-2}[2\mathrm{Re}((\hat{\varepsilon}'(\underline{\hat{\theta}}_{\mathrm{SML}}(Z), \hat{Z}))^{H}(\hat{\varepsilon}'(\underline{\hat{\theta}}_{\mathrm{SML}}(Z), \hat{Z})))]^{-1}\end{aligned} \tag{8-19}$$

so that (8-16) is replaced by

$$\hat{C}_{\hat{\theta}} = \frac{M-1}{M-3}[2\mathrm{Re}((\hat{\varepsilon}'(\underline{\hat{\theta}}_{\mathrm{SML}}(Z), \hat{Z}))^{H}(\hat{\varepsilon}'(\underline{\hat{\theta}}_{\mathrm{SML}}(Z), \hat{Z})))]^{-1} \tag{8-20}$$

8.3.5 Properties of the Cost Function in Its Global Minimum

Due to the availability of the nonparametric noise model it is possible to give a prediction of the value of the cost function that is expected to be observed at the end of the identification process if no model errors are present. To judge the difference between the actual, observed value and the expected cost, the variance of the cost should also be known. Consequently, this information can be used as an undermodeling detection tool during the model selection and validation process by comparing the actual value of the cost function with its expected value (see also Chapter 9). Undermodeling occurs if the orders of the numerator and/or denominator polynomials of the transfer function model are too small (unmodeled dynamics), or if a true linear time-invariant model simply does not exist (for example, nonlinear distortions).

Theorem 8.5 (Mean and Variance of the Global Minimum of the Cost Function): Under the conditions of Theorem 8.3, and for $M \geq 6$, the mean and variance of the global minimum of the cost function $\mathrm{var}(V_{\mathrm{SML}}(\hat{\theta}_{\mathrm{SML}}(Z), Z))$ are given by

$$\begin{aligned}\mathcal{E}\{V_{\mathrm{SML}}(\hat{\theta}_{\mathrm{SML}}(Z), Z)\} &\approx \frac{M-1}{M-2}\mathcal{E}\{V_{\mathrm{ML}}(\hat{\theta}_{\mathrm{ML}}(Z), Z)\} - \frac{M-1}{(M-3)(M-2)}n_{\theta}/2 \\ \mathrm{var}(V_{\mathrm{SML}}(\hat{\theta}_{\mathrm{SML}}(Z), Z)) &\approx \frac{(M-1)^2}{(M-2)(M-3)}\mathrm{var}(V_{\mathrm{ML}}(\tilde{\theta}_{\mathrm{ML}}(Z_0), Z)) \\ &\quad + \frac{(M-1)^2}{(M-2)^2(M-3)}\sum\nolimits_{k=1}^{F}\mathcal{E}\{|\hat{\varepsilon}(\Omega_k, \tilde{\theta}_{\mathrm{ML}}(Z_0), \hat{Z}(k))|^2\}\end{aligned} \tag{8-21}$$

in the presence of model errors $(\tilde{\theta}_{\mathrm{ML}}(Z_0) \neq \theta_0)$, and

$$\begin{aligned}\mathcal{E}\{V_{\text{SML}}(\hat{\theta}_{\text{SML}}(Z), Z)\} &\approx \frac{M-1}{M-2}(F - n_\theta/2)\\ \text{var}(V_{\text{SML}}(\hat{\theta}_{\text{SML}}(Z), Z)) &\approx \frac{(M-1)^3}{(M-2)^2(M-3)}F\end{aligned} \tag{8-22}$$

if no model errors are present ($\tilde{\theta}_{\text{ML}}(Z_0) = \theta_0$).

Proof. See Appendix 8.G. □

8.4 PROPERTIES OF THE GTLS ESTIMATOR USING A SAMPLE COVARIANCE MATRIX

The general form of the cost function of the GTLS estimator is given by (7-71):

$$V_{\text{GTLS}}(\theta, Z) = \frac{\sum_{k=1}^{F}|e(\Omega_k, \theta, Z(k))|^2}{\sum_{k=1}^{F}\sigma_e^2(\Omega_k, \theta)} \tag{8-23}$$

Replacing $Z(k)$ in this expression by the sample mean $\hat{Z}(k)$ and the exact noise (co)variances by the sample noise (co)variances gives the sample GTLS (SGTLS) cost function

$$V_{\text{SGTLS}}(\theta, Z) = \frac{\sum_{k=1}^{F}|\hat{e}(\Omega_k, \theta, \hat{Z}(k))|^2}{\sum_{k=1}^{F}\hat{\sigma}_e^2(\Omega_k, \theta)} \tag{8-24}$$

where $\hat{e}(\Omega_k, \theta, \hat{Z}(k))$, and $\hat{\sigma}_e^2(\Omega_k, \theta)$ are defined in (8-11). The minimizer $\hat{\theta}_{\text{SGTLS}}(Z)$ of (8-24) is not calculated using the iterative Newton-Gauss scheme (7-16) or (7-18) but via the generalized singular value decomposition of the matrix pair $(\hat{J}_{\text{re}}(Z), \hat{C})$ with $\hat{J}(Z) = \partial\hat{e}(\theta, \hat{Z})/\partial\theta$ and $\hat{C}$ a square root of the column covariance matrix of $\hat{j}_{\text{re}}(N_Z)$, with $\hat{j}(N_Z) = \hat{J}(Z) - \hat{J}(Z_0)$, calculated using the sample noise (co)variances (see Section 7.10). Just like the GTLS estimate, $\hat{\theta}_{\text{SGTLS}}(Z)$ suffers from the amplification of the high-frequency errors (see Section 7.10.3). To cope with this problem weighted SGTLS versions can be constructed as in Sections 7.10.3 and 7.12.4.

Theorem 8.6 (Asymptotic Properties $\hat{\theta}_{\text{SGTLS}}(Z)$): Consider model (8-3) with any identifiable parameterization of Section 5.2. Under the assumptions of Section 7.6 and Assumption 8.1 ($M \geq 2$) the minimizer $\hat{\theta}_{\text{SGTLS}}(Z)$ of (8-24) has the asymptotic properties of Theorem 7.21 with $V_F(\theta, Z) = V_{\text{SGTLS}}(\theta, Z)$, where

1. $V_F(\theta)$ and $V_*(\theta)$ are given by

$$V_F(\theta) = \frac{\sum_{k=1}^{F}|e(\Omega_k, \theta, Z_0(k))|^2}{\frac{1}{M}\sum_{k=1}^{F}\sigma_e^2(\Omega_k, \theta)} + 1, \quad V_*(\theta) = \frac{\int_{f_{\min}}^{f_{\max}}|e(\Omega(f), \theta, Z_0(f))|^2 n(f)df}{\frac{1}{M}\int_{f_{\min}}^{f_{\max}}\sigma_e^2(\Omega(f), \theta)n(f)df} + 1 \tag{8-25}$$

2. $\mathcal{E}\{\delta_\theta(Z)\}$ in (7-25) is not necessarily zero or even may not exist.
3. $\delta_\theta(Z)$ in the expression of the covariance matrix (7-26) is replaced by $d_\theta(Z)$ as in Theorem 16.25.

Proof. See Appendix 8.H. □

From Theorem 8.6 it follows that the estimate $\hat{\theta}_{\mathrm{SGTLS}}(Z)$ is consistent and $\tilde{\theta}_{\mathrm{SGTLS}}(Z_0)$, $\theta_{*\mathrm{SGTLS}}$ are the noiseless solutions in case model errors are present (apply quick analysis tools number 2 and 3 of Section 7.5 to $V(\theta)$ and $V_*(\theta)$ in (8-25)). Because $V_{\mathrm{SGTLS}}(\lambda\theta, Z) = V_{\mathrm{SGTLS}}(\theta, Z)$ the estimate $\hat{\theta}_{\mathrm{SGTLS}}(Z)$ is independent of the particular constraint chosen, for example, $a_i = 1$, $b_i = 1$, or $\|\theta\|_2^2 = 1$ (quick tool number 4). The relationship between the asymptotic behavior of the estimates, based on the true (8-23) and the sample (8-24) noise (co)variances, is established in the following theorem.

Theorem 8.7 (Relationship between $\hat{\theta}_{\mathrm{SGTLS}}(Z)$ and $\hat{\theta}_{\mathrm{GTLS}}(Z)$): Under the conditions of Theorem 8.6, the estimates based on the true (8-23) and the sample (8-24) noise (co)variances are related to each other by

$$V_{\mathrm{SGTLS}}(\theta) = V_{\mathrm{GTLS}}(\theta) \text{ and } V_{*\mathrm{SGTLS}}(\theta) = V_{*\mathrm{GTLS}}(\theta) \tag{8-26}$$

$$\tilde{\theta}_{\mathrm{SGTLS}}(Z_0) = \tilde{\theta}_{\mathrm{GTLS}}(Z_0) \text{ and } \theta_{*\mathrm{SGTLS}} = \theta_{*\mathrm{GTLS}} \tag{8-27}$$

In the absence of model errors, $\tilde{\theta}_{\mathrm{SGTLS}}(Z_0) = \theta_0$, we have

$$\mathrm{Cov}(\delta_{\theta\mathrm{SGTLS}}(Z)) = \mathrm{Cov}(\delta_{\theta\mathrm{GTLS}}(Z)) \tag{8-28}$$

where $\hat{\theta}(Z) = \theta_0 + \delta_\theta(Z) + O_p(F^{-1})$ with $\mathcal{E}\{\delta_\theta(Z)\} = 0$ and $\delta_\theta(Z) = O_p(F^{-1/2})$, and where $\sqrt{F}\delta_\theta(Z)$ is asymptotically normally distributed.

Proof. See Appendix 8.I. □

Contrary to the SML solution, it is not necessary to calculate the sample covariance $\hat{\sigma}^2_{YU}(k)$ if it is known that the input and output errors are uncorrelated $\sigma^2_{YU}(k) = 0$. From both theorems it follows that $\hat{\theta}_{\mathrm{SGTLS}}(Z)$ has asymptotic ($F \to \infty$) properties similar to those of $\hat{\theta}_{\mathrm{GTLS}}(Z)$, even if the sample (co)variances are calculated using only $M = 2$ independent repeated experiments. For example, in the absence of model errors, the asymptotic uncertainty of the SGTLS equals that of the GTLS. This is no longer true if model errors are present. The basic reason for the similar asymptotic behavior of $\hat{\theta}_{\mathrm{SGTLS}}(Z)$ and $\hat{\theta}_{\mathrm{GTLS}}(Z)$ is that the "poor quality" sample (co)variances are averaged over the frequency in the cost function (8-24), resulting in a "high quality" estimate of the denominator of the cost function.

8.5 PROPERTIES OF THE BTLS ESTIMATOR USING A SAMPLE COVARIANCE MATRIX

The general form of the cost function of the BTLS estimator is given by (7-97)

$$V_{\text{BTLS}}(\theta^{(i)}, Z) = \frac{\sum_{k=1}^{F} \frac{|e(\Omega_k, \theta^{(i)}, Z(k))|^2}{\sigma_e^{2r}(\Omega_k, \theta^{(i-1)})}}{\sum_{k=1}^{F} \frac{\sigma_e^2(\Omega_k, \theta^{(i)})}{\sigma_e^{2r}(\Omega_k, \theta^{(i-1)})}} \tag{8-29}$$

We recall that $\sigma_e^{2r}(\Omega_k, \theta^{(i-1)})$ and $\sigma_e^2(\Omega_k, \theta^{(i)})$ stem, respectively, from the left W and right C weighting matrices in the total least squares problem (see Sections 7.10 and 7.12.3). Following along the lines of Section 8.4, one could think of replacing the true noise (co)variances everywhere in (8-29) by the sample noise (co)variances

$$\frac{\sum_{k=1}^{F} \frac{|\hat{e}(\Omega_k, \theta^{(i)}, \hat{Z}(k))|^2}{\hat{\sigma}_{\hat{e}}^{2r}(\Omega_k, \theta^{(i-1)})}}{\sum_{k=1}^{F} \frac{\hat{\sigma}_{\hat{e}}^2(\Omega_k, \theta^{(i)})}{\hat{\sigma}_{\hat{e}}^{2r}(\Omega_k, \theta^{(i-1)})}} \tag{8-30}$$

Proceeding in that way, we violate the assumptions of the framework developed in Chapter 16. Indeed, the theorems of Chapter 16 (strong consistency, convergence rate, asymptotic bias, and asymptotic normality) are valid only if the number of stochastic parameters in the weighting remains finite for finite M and $F \to \infty$, and if these parameters converge strongly $(F \to \infty)$ to a nonrandom limit value. This is certainly not the case for the left weighting $\hat{\sigma}_{\hat{e}}^{2r}(\Omega_k, \theta^{(i-1)})$ in (8-30). Therefore, to preserve the strong consistency, the noise (co)variances in the left weighting matrix W $(\hat{\sigma}_{\hat{e}}^{2r}(\Omega_k, \theta^{(i-1)}))$ in (8-30) should be modeled over the frequency using a finite (F-independent) number of parameters α. The estimates $\hat{\alpha}(Z)$ should strongly converge to some nonrandom value α_*. As it is the case for the SGTLS estimator, the right weighting C $(\hat{\sigma}_{\hat{e}}^2(\Omega_k, \theta^{(i)}))$ in (8-30) must still be calculated using the original sample noise (co)variances.

For computational reasons, only noise models that are linear in the parameters are considered. For example,

$$\sigma_n^2(k, \alpha) = \sum_{r=1}^{p_n} \alpha_{nr} h_{nr}(\Omega_k) \qquad n = 1, 2, 3 \tag{8-31}$$

with $\sigma_1^2 = \sigma_U^2$, $\sigma_2^2 = \sigma_Y^2$, and $\sigma_3^2 = \sigma_{YU}^2$ and where $h_{nr}(\Omega)$, $r = 1, 2, \ldots, p_n$, are linear independent basis functions with p_n independent of F. The choice of the parametric noise model is not critical because it does not influence the consistency property. It, however, influences the uncertainty of the estimated plant model parameters. Under Assumption 7.3 or 7.4 and Assumption 8.1 the linear least squares estimate $\hat{\alpha}^T(Z) = [\hat{\alpha}_1^T(Z)\ \hat{\alpha}_2^T(Z)\ \hat{\alpha}_3^T(Z)]$

$$\hat{\alpha}_n(Z) = (H_n^T H_n)^{-1} H_n^T [\hat{\sigma}_n^2(1), \ldots, \hat{\sigma}_n^2(F)]^T \qquad n = 1, 2, 3 \tag{8-32}$$

with $H_{n[k, r]} = h_{nr}(\Omega_k)$ and $\hat{\sigma}_n^2(k)$ the sample noise (co)variances ($\hat{\sigma}_1^2 = \hat{\sigma}_U^2$, $\hat{\sigma}_2^2 = \hat{\sigma}_Y^2$, and $\hat{\sigma}_3^2 = \hat{\sigma}_{YU}^2$) converges strongly $(F \to \infty)$ to a nonrandom value α_*. Note that the estimated parametric noise models $\sigma_n^2(k, \hat{\alpha}(Z))$, $n = 1, 2, 3$, represent a linear projection of an F-dimensional space onto a p_n-dimensional space.

Replacing the true noise (co)variances $\sigma_n^2(k)$ in (8-29) by the estimated parametric noise model $\sigma_n^2(k, \hat{\alpha}(Z))$ in the left weighting W, $\sigma_n^2(k)$ by the sample noise (co)variances $\hat{\sigma}_n^2(k)$ in the right weighting C, and the measurements Z by the sample mean $\hat{Z}$ gives the sample BTLS (SBTLS) cost function

$$V_{\text{SBTLS}}(\theta^{(i)}, Z) = \frac{\sum_{k=1}^{F} \frac{|\hat{e}(\Omega_k, \theta^{(i)}, \hat{Z}(k))|^2}{\sigma_{\hat{e}}^{2r}(\Omega_k, \theta^{(i-1)}, \hat{\alpha}(Z))}}{\sum_{k=1}^{F} \frac{\hat{\sigma}_{\hat{e}}^2(\Omega_k, \theta^{(i)})}{\sigma_{\hat{e}}^{2r}(\Omega_k, \theta^{(i-1)}, \hat{\alpha}(Z))}} \tag{8-33}$$

where $\hat{e}(\Omega_k, \theta, \hat{Z}(k))$ and $\hat{\sigma}_{\hat{e}}^2(\Omega_k, \theta)$ are defined in (8-11) and

$$\begin{aligned} \sigma_{\hat{e}}^2(\Omega_k, \theta, \hat{\alpha}(Z)) &= \sigma_e^2(\Omega_k, \theta, \hat{\alpha}(Z))/M \\ \sigma_e^2(\Omega_k, \theta, \hat{\alpha}(Z)) &= \sigma_Y^2(k, \hat{\alpha}(Z))|A(\Omega_k, \theta)|^2 + \sigma_U^2(k, \hat{\alpha}(Z))|B(\Omega_k, \theta)|^2 \\ &\quad -2\text{Re}(\sigma_{YU}^2(k, \hat{\alpha}(Z))A(\Omega_k, \theta)\bar{B}(\Omega_k, \theta)) \end{aligned} \tag{8-34}$$

Likewise, the SGTLS estimator the minimizer $\hat{\theta}_{\text{SBTLS}}(Z)$ of (8-33) is calculated via the generalized singular value decomposition of the matrix pair $(W_{\text{Re}}\hat{J}_{\text{re}}(Z), \hat{C})$ with $\hat{J}(Z) = \partial\hat{e}(\theta, \hat{Z})/\partial\theta$, W a diagonal matrix with $W_{[k,k]} = \sigma_{\hat{e}}^{-r}(\Omega_k, \theta^{(i-1)}, \hat{\alpha}(Z))$, and $\hat{C}$ a square root of the column covariance matrix of $W_{\text{Re}}\hat{j}_{\text{re}}(N_Z)$, with $\hat{j}(N_Z) = \hat{J}(Z) - \hat{J}(Z_0)$, calculated using the sample noise (co)variances (see Section 7.10). The asymptotic properties of the first step of the iterative procedure (8-33) are analyzed in the following theorem.

Theorem 8.8 (Asymptotic Properties $\hat{\theta}_{\text{SBTLS}}(Z)$): Consider model (8-3) with any identifiable parameterization of Section 5.2. Under the assumptions of Section 7.6 and Assumption 8.1 ($M \geq 2$) the minimizer $\hat{\theta}_{\text{SBTLS}}(Z) = \theta^{(1)}$ of (8-33), with parametric noise model (8-31) and initial guess $\theta^{(0)} = \hat{\theta}(Z)$ satisfying Theorem 7.21, has the asymptotic properties of Theorem 7.21 with $V_F(\theta, Z) = V_{\text{SBTLS}}(\theta, Z)$ and where

1. $V_F(\theta)$ and $V_*(\theta)$ are given by

$$V(\theta) = \frac{\sum_{k=1}^{F} \frac{|e(\Omega_k, \theta, Z_0(k))|^2}{\sigma_e^{2r}(\Omega_k, \theta_*, \alpha_*)}}{\frac{1}{M}\sum_{k=1}^{F} \frac{\sigma_e^2(\Omega_k, \theta)}{\sigma_e^{2r}(\Omega_k, \theta_*, \alpha_*)}} + 1,$$

$$V_*(\theta) = \frac{\int_{f_{\min}}^{f_{\max}} \frac{|e(\Omega(f), \theta, Z_0(f))|^2}{\sigma_e^{2r}(\Omega(f), \theta_*, \alpha_*)} n(f)df}{\frac{1}{M}\int_{f_{\min}}^{f_{\max}} \frac{\sigma_e^2(\Omega(f), \theta)}{\sigma_e^{2r}(\Omega(f), \theta_*, \alpha_*)} n(f)df} + 1 \tag{8-35}$$

2. $\mathcal{E}\{\delta_\theta(Z)\}$ in (7-25) is not necessarily zero or even may not exist,
3. $\delta_\theta(Z)$ in the expression of the covariance matrix (7-26) is replaced by $d_\theta(Z)$ as in Theorem 16.25.

Proof. See Appendix 8.J. □

From Theorem 8.8 it follows that the estimate $\hat{\theta}_{\text{SBTLS}}(Z)$ is consistent and $\tilde{\theta}_{\text{SBTLS}}(Z_0)$, $\theta_{*\text{SBTLS}}$ are the noiseless solutions in case model errors are present (apply quick analysis tools number 2 and 3 of Section 7.5 to $V(\theta)$ and $V_*(\theta)$ in (8-35)). Because $V_{\text{SBTLS}}(\lambda\theta, Z) = V_{\text{SBTLS}}(\theta, Z)$, the estimate $\hat{\theta}_{\text{SBTLS}}(Z)$ is independent of the particular constraint chosen, for example, $a_i = 1$, $b_i = 1$ or $\|\theta\|_2^2 = 1$ (quick tool number 4). The relationship between the asymptotic behavior of $\hat{\theta}_{\text{SBTLS}}(Z)$ and $\hat{\theta}_{\text{BTLS}}(Z)$ is established in the following theorem.

Theorem 8.9 (Relationship between $\hat{\theta}_{\text{SBTLS}}(Z)$ and $\hat{\theta}_{\text{BTLS}}(Z)$): Under the conditions of Theorem 8.8, and assuming that the parametric noise models (8-31), $\sigma_n^2(k, \hat{\alpha}(Z))$ with $n = 1, 2, 3$, are consistent estimates of the true noise (co)variances, $\sigma_n^2(k)$ with $n = 1, 2, 3$, the estimates based on the true (8-29) and the sample (8-33) noise (co)variances are related to each other by

$$V_{\text{SBTLS}}(\theta) = V_{\text{BTLS}}(\theta) \quad \text{and} \quad V_{*\text{SBTLS}}(\theta) = V_{*\text{BTLS}}(\theta) \tag{8-36}$$

$$\tilde{\theta}_{\text{SBTLS}}(Z_0) = \tilde{\theta}_{\text{BTLS}}(Z_0) \quad \text{and} \quad \theta_{*\text{SBTLS}} = \theta_{*\text{BTLS}} \tag{8-37}$$

In the absence of model errors, $\tilde{\theta}_{\text{SBTLS}}(Z_0) = \theta_0$, we have

$$\mathrm{Cov}(\delta_{\theta_{\text{SBTLS}}}(Z)) = \mathrm{Cov}(\delta_{\theta_{\text{BTLS}}}(Z)) \tag{8-38}$$

where $\hat{\theta}(Z) = \theta_0 + \delta_\theta(Z) + O_p(F^{-1})$ with $\mathcal{E}\{\delta_\theta(Z)\} = 0$ and $\delta_\theta(Z) = O_p(F^{-1/2})$, and where $\sqrt{F}\,\delta_\theta(Z)$ is asymptotically normally distributed.

Proof. See Appendix 8.K. □

As with the SGTLS solution, it is not necessary to calculate the sample covariance $\hat{\sigma}_{YU}^2(k)$ if it is known that $\sigma_{YU}^2(k) = 0$. From both theorems it follows that $\hat{\theta}_{\text{SBTLS}}(Z)$ has asymptotic ($F \to \infty$) properties similar to those of $\hat{\theta}_{\text{BTLS}}(Z)$, even for $M = 2$. The price to be paid is the construction of a consistent parametric model for the noise (co)variances in the left weighting. If the parametric noise model is a poor approximation of the true noise model, then the estimated plant model parameters are still strongly consistent (Theorem 8.8 is still valid), but their uncertainty may increase (Theorem 8.9 is no longer valid).

8.6 PROPERTIES OF THE SUB ESTIMATOR USING A SAMPLE COVARIANCE MATRIX

Subspace Algorithms 7.24 and 7.25 identify model (8-3) where $G(\Omega, \theta)$ is parameterized in the state space parameters (A, B, C, D)

$$G(\xi, \theta) = C(\xi I_{n_a} - A)^{-1}B + D \tag{8-39}$$

with $\xi = z$ for discrete-time systems and $\xi = s$ for continuous-time systems. This is done in three steps: (i) estimation of the (generalized) extended observability matrix O_r or $O_{r\perp}$ using the input-output spectra, (ii) estimation of A and C using $\hat{O}_r$ or $\hat{O}_{r\perp}$, and (iii) estimation of B and D using $\hat{A}$, $\hat{C}$, and the input-output spectra. The procedure has been developed for problems where the input is exactly known, $N_U(k) = 0$. If the input observations are noisy, $N_U(k) \neq 0$, then the errors-in-variables problem (8-7) is transformed into an equivalent frequency response function measurement problem

$$Y(k)/U(k) = G_0(\Omega_k) + N_G^{[l]}(k) \text{ with } \mathcal{E}\{N_G^{[l]}(k)\} = 0 \tag{8-40}$$

where $\text{var}(N_G^{[l]}(k)) = \sigma_G^2(k)$ is related to the input-output noise (co)variances by (7-53). For worst case input and output signal-to-noise ratios $|U_0(k)|/\sigma_U(k)$ and $|Y_0(k)|/\sigma_Y(k)$ larger than 10 dB, (8-40) is a very good approximation (see Section 7.9 for an elaborated discussion).

Knowledge of the noise variance $\sigma_Y^2(k)$ or $\sigma_G^2(k)$ is required in the first and third steps of the subspace algorithms. In the first step we need the covariance matrix $C_{\mathbf{Y}}$ or $C_{\mathbf{Y}_\perp}$ for, respectively, discrete-time or continuous-time modeling,

$$\begin{aligned} C_{\mathbf{Y}} &= \text{Re}(\textstyle\sum_{k=1}^F \sigma_Y^2(k) W_r(k) W_r^H(k)) \text{ with } W_r(k) = [1 \ z_k \ \ldots \ z_k^{r-1}]^T \\ C_{\mathbf{Y}_\perp} &= \text{Re}(\textstyle\sum_{k=1}^F \sigma_Y^2(k) W_Y(k) W_Y^H(k)) \text{ with } W_Y(k) = [p_0(s_k) \ p_1(s_k) \ \ldots \ p_{r-1}(s_k)]^T \end{aligned} \tag{8-41}$$

where $p_n(s)$ are scalar orthogonal polynomials of order $n = 0, 1, \ldots, r-1$ (see Appendix 7.S). In the third step the estimates $\hat{B}$ and $\hat{D}$ are obtained by minimizing

$$V_{\text{SUB}}(B, D, \hat{A}, \hat{C}, Z) = \sum_{k=1}^F \frac{|Y(k) - (\hat{C}(\xi I_{n_a} - \hat{A})^{-1}B + D)U(k)|^2}{\sigma_Y^2(k)} \tag{8-42}$$

w.r.t. B and D. Replacing $Z(k)$ by the sample mean $\hat{Z}(k)$, and the exact (co)variances by the sample (co)variances, in the first and the third step of Algorithms 7.24 and 7.25, defines the sample subspace (SSUB) algorithms. Expressions (8-41) and (8-42) then become

$$\begin{aligned} \hat{C}_{\mathbf{Y}} &= \text{Re}(\textstyle\sum_{k=1}^F M^{-1} \hat{\sigma}_Y^2(k) W_r(k) W_r^H(k)) \\ \hat{C}_{\mathbf{Y}_\perp} &= \text{Re}(\textstyle\sum_{k=1}^F M^{-1} \hat{\sigma}_Y^2(k) W_Y(k) W_Y^H(k)) \\ V_{\text{SSUB}}(B, D, \hat{A}, \hat{C}, Z) &= \sum_{k=1}^F \frac{|\hat{Y}(k) - (\hat{C}(\xi I_{n_a} - \hat{A})^{-1}B + D)\hat{U}(k)|^2}{\hat{\sigma}_Y^2(k)/M} \end{aligned} \tag{8-43}$$

For noisy inputs, $\hat{Y}(k)$, $\hat{U}(k)$, and $\hat{\sigma}_Y^2(k)$ are replaced by $\hat{Y}(k)/\hat{U}(k)$, 1, and $\hat{\sigma}_G^2(k)$.

Theorem 8.10 (Asymptotic Properties $\hat{\theta}_{\text{SSUB}}(Z)$): Consider transfer function model (8-39), parameterized in its state space representation, and Algorithms 7.24 and 7.25, where the input-output spectra are replaced by their sample mean, and the exact noise (co)variances by the sample noise (co)variances. The resulting estimate $\hat{\theta}_{\text{SSUB}}(Z)$ has the asymptotic ($F \to \infty$) properties of Theorem 7.28, where

1. $M \geq 2$ for the stochastic convergence and stochastic convergence rate (properties 1, 2, and 3 of Theorem 7.28) of $\hat{A}$ and $\hat{C}$ (assumptions of Sections 7.6.1 and 7.6.5 and Assumptions 7.26, 7.27 and 8.1).
2. $M \geq 4$ for the stochastic convergence and the stochastic convergence rate (properties 1, 2, and 3 of Theorem 7.28) of $\hat{B}$ and $\hat{D}$ (assumptions of Sections 7.6.1 and 7.6.5, and Assumptions 7.26, 7.27, 8.1, and 8.2).

Proof. See Appendix 8.L. □

From Theorem 8.10, it follows that $\hat{\theta}_{\text{SSUB}}(Z)$ is consistent and that $\theta_{*\text{SSUB}}$ is the noiseless solution in case model errors are present. Note that estimation of the poles, which depend on $\hat{A}$ only, has a fundamentally different stochastic behavior than the estimation of the zeros, which depend on $\hat{A}$, $\hat{B}$, $\hat{C}$, and $\hat{D}$. Indeed, the poles can be estimated consistently under the same noise assumptions as those when the noise (co)variances are known, while consistent estimation of the zeros requires that the disturbing noise is independent (over the frequency) and normally distributed (Assumption 8.2). The relationship between the asymptotic behavior of $\hat{\theta}_{\text{SSUB}}(Z)$ and $\hat{\theta}_{\text{SUB}}(Z)$ is established in the following theorem.

Theorem 8.11 (Relationship between $\hat{\theta}_{\text{SSUB}}(Z)$ and $\hat{\theta}_{\text{SUB}}(Z)$): Under the conditions of Theorem 8.10, the estimates based on the true (8-41), (8-42), and the sample (8-43) noise (co)variances are related to each other by

1. For $M \geq 2$, $A_{*\text{SSUB}} = A_{*\text{SUB}}$ and $C_{*\text{SSUB}} = C_{*\text{SUB}}$
2. For $M \geq 4$, $B_{*\text{SSUB}} = B_{*\text{SUB}}$ and $D_{*\text{SSUB}} = D_{*\text{SUB}}$

Proof. See Appendix 8.L. □

8.7 IDENTIFICATION IN THE PRESENCE OF NONLINEAR DISTORTIONS

We recall that the steady-state response of the nonlinear system $y(t) = G[u_0(t)]$ to a random phase multisine (see Definition 3.2) $u_0(t)$ is given by

$$Y(k) = G_R(s_k)U_0(k) + Y_S(k) \tag{8-44}$$

with $Y_S(k)$ the stochastic nonlinear contributions (see Section 5.8) and where $y_s(t) = \text{IDFT}(Y_S(k))$ has the same periodicity as the input signal $u_0(t)$. Calculating the sample (co)variances over several periods of the output signal $y(t)$ will, hence, give no information about the stochastic nonlinear contributions $Y_S(k)$. Experiments with different realizations of the random phases $\varphi_k = \angle U_0(k)$ of the input signal are necessary to get the contribution of $Y_S(k)$ to the sample (co)variances. To distinguish between the signal part $G_R(s_k)U_0(k)$ and the noise part $Y_S(k)$ in (8-44), the input and output DFT spectra of each experiment must be turned back with the corresponding phases $\varphi_k = \angle U_0(k)$ of the input. Because the observations of the input are in general corrupted by measurement noise, a reference signal $r(t)$ is required to perform this operation (see Figure 8-1). This leads to the following measurement strategy.

1. Choose the amplitude spectrum of the random phase multisine (see Definition 3.2).
2. Make a random choice of the phases φ_k of the random phase multisine (see Definition 3.2) and calculate the corresponding time signal $r(t)$.
3. Apply the excitation to the plant and measure $P \geq 1$ periods of the steady-state response $u(t)$, $y(t)$.
4. Repeat steps 2 and 3 $M \geq 4$ times.
5. Calculate the DFT spectra of the input $u(t)$, output $y(t)$, and reference $r(t)$ signals for each experiment at the excited DFT frequencies. This gives M sets of the reference $R^{[m]}(k)$, the noisy input $U^{[m]}(k)$, and the noisy output $Y^{[m]}(k)$ spectra, $k = 1, 2, \ldots, F$ and $m = 1, 2, \ldots, M$.

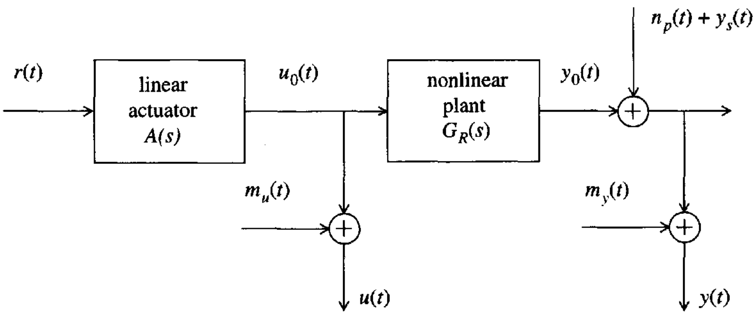


Figure 8-1. Measurement of the best linear approximation $G_R(s)$ of a nonlinear device: $u_0(t)$, $y_0(t)$ are the true input/output signals, $m_u(t)$, $m_y(t)$ are the input-output measurement errors, $y_s(t)$ is the zero mean stochastic nonlinear contribution, and $r(t)$ is the reference signal (typically the waveform stored in the arbitrary waveform generator).

6. Project the input and output spectra on the reference spectrum

$$Y_R^{[m]}(k) = Y^{[m]}(k)/R^{[m]}(k),\ U_R^{[m]}(k) = U^{[m]}(k)/R^{[m]}(k) \tag{8-45}$$

and finally calculate the sample mean and sample (co)variances

$$\hat{Y}(k) = \frac{1}{M}\sum_{m=1}^{M} Y_R^{[m]}(k),\ \hat{U}(k) = \frac{1}{M}\sum_{m=1}^{M} U_R^{[m]}(k) \tag{8-46}$$

$$\begin{aligned}\hat{\sigma}_Y^2(k) &= \frac{1}{M-1}\sum_{m=1}^{M}\left|\hat{Y}(k) - Y_R^{[m]}(k)\right|^2\\ \hat{\sigma}_U^2(k) &= \frac{1}{M-1}\sum_{m=1}^{M}\left|\hat{U}(k) - U_R^{[m]}(k)\right|^2\\ \hat{\sigma}_{YU}^2(k) &= \frac{1}{M-1}\sum_{m=1}^{M}(\hat{Y}(k) - Y_R^{[m]}(k))\overline{(\hat{U}(k) - U_R^{[m]}(k))}\end{aligned} \tag{8-47}$$

Using the sample means (8-46) and sample (co)variances (8-47), we can calculate the SML (8-11), SGTLS (8-24), and SBTLS (8-33) estimates of the related linear dynamic system $G_R(s)$. Because Assumption 8.2 is asymptotically valid ($F \to \infty$) for $Y_S(k)$ (see Section 5.8 and Theorems 3.10 and 3.11) it still makes sense to study the properties of $\hat{\theta}_{SML}(Z)$ under the idealized assumptions of Theorems 8.3 and 8.4, where $G_0(s)$ is replaced by $G_R(s)$. Because $Y_S(k)$ has similar properties as the measurement and process noise in a time domain experiment (see Sections 5.8 and 7.6), Theorems 8.6 and 8.7 and Theorems 8.8 and 8.9, where $G_0(s)$ is replaced by $G_R(s)$, remain valid for, respectively, $\hat{\theta}_{SGTLS}(Z)$ and $\hat{\theta}_{SBTLS}(Z)$ (proof: similar to Appendix 7.V). Similar conclusions hold for the SSUB algorithms of Section 8.6.

It may happen that the nonlinear plant loads the generator nonlinearly or that the actuator itself is nonlinear, creating nonlinear distortions at the input of the plant. Figure 8-2 shows the corresponding block diagram. Applying the measurement strategy to this situation gives an estimate $\hat{G}_R(s_k) = \hat{Y}(k)/\hat{U}(k)$ that converges strongly ($M \to \infty$) to $\mathcal{E}\{Y_R(k)\}/\mathcal{E}\{U_R(k)\} = T_R(s_k)/A_R(s_k)$ where $T_R(s_k)$ and $A_R(s_k)$ are the related linear dynamic systems of the nonlinear systems $T[.]$ and $A[.]$ respectively (see Appendix 8.M). Note that in general $T_R(s_k)/A_R(s_k) \neq G_R(s_k)$; however, if the nonlinear distortions at the input are small, $\text{var}(U_S(k)) \ll |A_R(s_k)R(k)|^2$, or if the related linear dynamic system $G_R(s_k)$ is not very sensitive to (small) variations of the input power spectrum, then

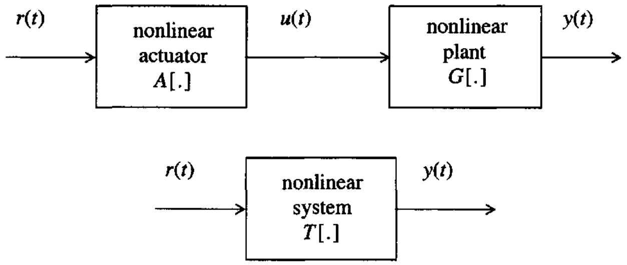


Figure 8-2. Schematic representation of a nonlinear device loading nonlinearly the generator: the input $u(t)$ of the device is nonlinearly related to the reference signal $r(t)$. The overall system from the reference $r(t)$ to the output of the plant is denoted by $T[.]$.

$T_R(s_k)/A_R(s_k) \approx G_R(s_k)$. The same is true for the parametric estimate $G(s, \hat{\theta})$ because it converges strongly $(F \to \infty)$ to the related linear dynamic system.

8.8 ILLUSTRATION AND OVERVIEW OF THE PROPERTIES

8.8.1 Real Measurement Example

We illustrate the SML (8-10), SGTLS (8-24), "full" $(r = 1)$ SBTLS (8-33), and SSUB (Algorithm 7.25 with $\hat{\sigma}_Y^2(k)$ and $r = 70$) estimators on the second measurement example (flight flutter data) of Section 7.15.3. The sample noise (co)variances are calculated using three independent burst swept-sine experiments (see Figure 8-3). Because the three experiments were not synchronized, a postsynchronization was first executed before calculating the sample noise (co)variances. The postsynchronization consists of estimating the delay of the second and third experiments with respect to the first experiment and adding a corresponding phase shift $e^{j\omega\tau^{[l]}}$, $l = 1, 2$, to DFT spectra of the second and the third experiment. Although $M = 3$ independent repeated experiments are not sufficient to use the sample noise (co)variances within the SML and SSUB framework (see Theorem 8.3 and Theorem 8.10), we still calculated the SML and SSUB estimates to show that they are rather robust w.r.t. to this condition. For SBTLS the noise (co)variances in the left weighting are modeled by a constant: put $h_{nr}(\Omega) = 1$ and $p_n = 1$, $n = 1, 2, 3$, in (8-31). Figure 8-4 shows the estimation results for a rational form in s of order $n_b = 11$ over $n_a = 10$. It can be seen that the SBTLS and SSUB estimates have SML quality. The SGTLS estimate misses the second reso-

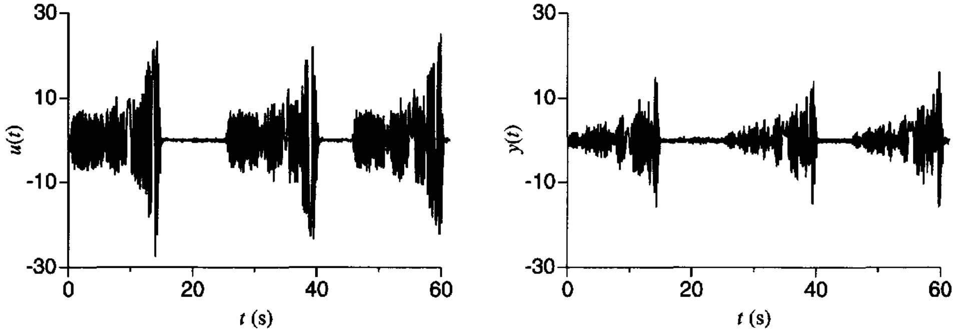


Figure 8-3. Input and output signals of the flight flutter data showing the three independent burst swept-sine experiments.

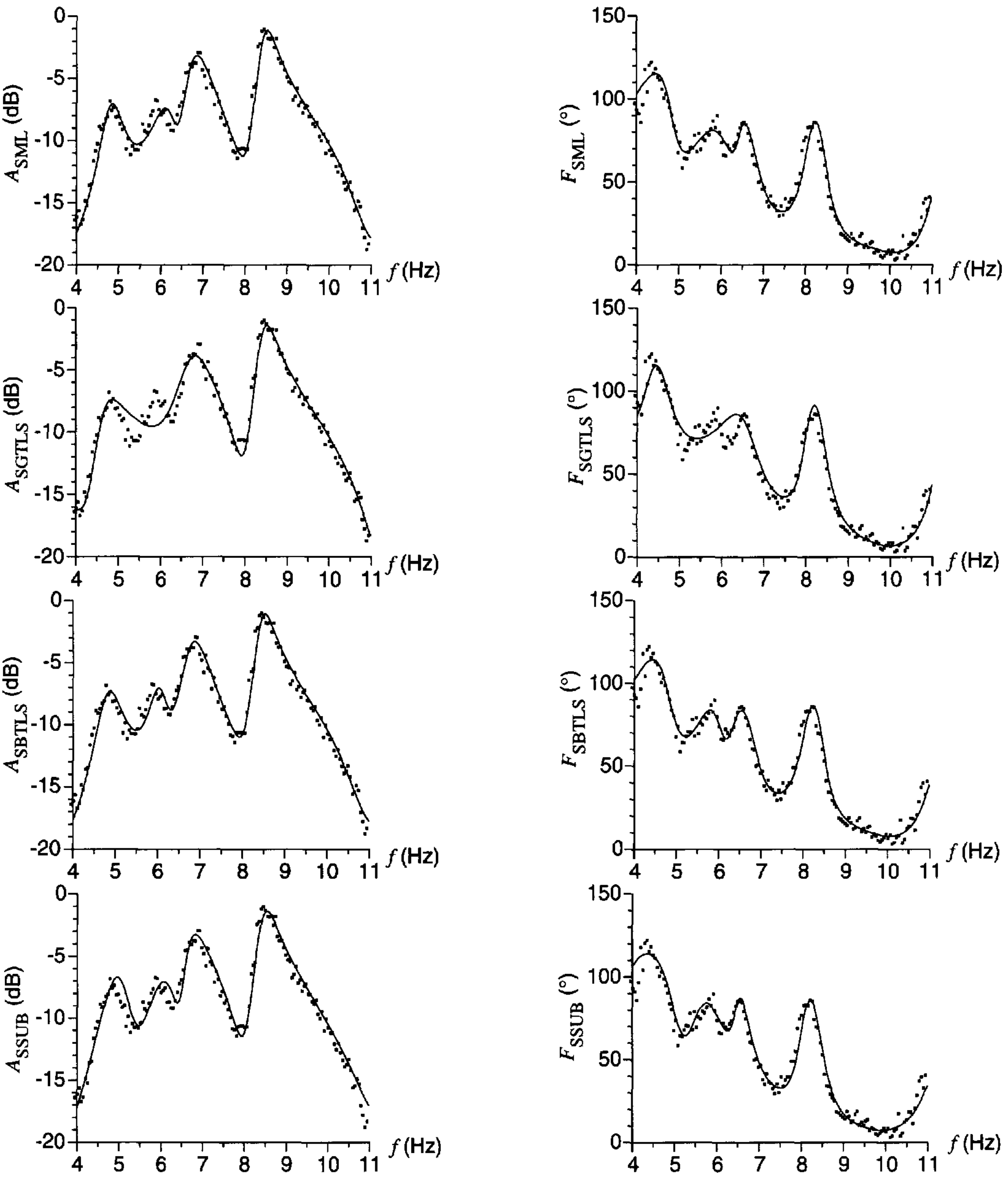


Figure 8-4. Comparison between the measurements (dots) and the estimates using the sample noise (co)variances (solid line) of the flight flutter data (model $n_a = 10$, $n_b = 11$). From left to right amplitude and phase.

nance peak, which can be explained by its inappropriate frequency weighting (see also Section 7.10.3). Comparing Figure 7-13 on page 236 and Figure 8-4, it follows that SGTLS, SBTLS, and SSUB perform (much) better than GTLS, BTLS, and SUB. The basic reason for this is that in Figure 7-13 the three burst swept-sine excitations were treated as independent nonsynchronized experiments, whereas in Figure 8-4 the set of three postsynchronized input-output DFT spectra have been averaged (see also Section 11.3.4.5). The averaging improves the signal-to-noise ratio (compare the measured frequency response functions in both figures), which explains the better behavior of SGTLS, SBTLS, and SSUB.

8.8.2 Overview of the Properties

The SML, SGTLS, SBTLS, and SSUB estimators have the same basic properties as the ML, GTLS, BTLS, and SUB estimators (see Table 7-5 on page 238), except that no prior noise information is required. Table 8-1 gives an overview of the differences and the similarities between the estimates using the sample and the true noise (co)variances

TABLE 8-1 Comparison of the Estimates Using the Sample (Subscript S) and the True Noise (Co)Variances in the General Case of Input-Output Errors

Estimator	$\theta_{*S} = \theta_*$?	$R = \frac{\text{var}(\underline{f}(\hat{\theta}_S))}{\text{var}(\underline{f}(\hat{\theta}))}$ (1)	Same Noise Assumption	Estim. $\hat{\sigma}^2_{YU}$ if $\sigma^2_{YU} = 0$	Sens. to Noise Model Errors
SML	Yes(2)	$\frac{M-2}{M-3}$ (3)	No(4)	Yes(5)	Excellent(6)
SGTLS	Yes(2)	1(3)	Yes(4)	No	Excellent(6)
SBTLS	Yes(2)	1(3)	Yes(4)	No	Very good(6)
SSUB	Yes(2)	—	Yes/no(4)	No/yes(5)	Excellent(6)

See Table 7-5 on page 238 for an overview of the basic properties of ML, GTLS, BTLS, and SUB.

1. $f(.)$ represents any invariant of the model $G(\Omega, \theta)$, for example, the frequency response function, the poles, or the zeros. In order to ensure the existence of the variance, the function $f(.)$ is truncated as $\hat{\theta}(Z)$ in (7-24).
2. The required number of independent repeated experiments is $M \geq 4$ for SML and $M \geq 2$ for SGTLS and SBTLS. The SSUB estimator requires $M \geq 2$ for the poles and $M \geq 4$ for the zeros. If the input is observed with errors, then SSUB uses $G(\Omega_k) = Y(k)/U(k)$ as primary data, and $\theta_{*S} = \theta_*$ is "practically valid" if the worst case input-output signal-to-noise ratio is larger than 10 dB (see Section 7.9).
3. If no model errors are present, otherwise the uncertainty is larger. For SBTLS the parametric noise model in the left weighting should be a consistent estimate of the true noise model. The required number of independent repeated experiments is $M \geq 6$ for SML and $M \geq 2$ for SGTLS and SBTLS.
4. SML requires $M \geq 6$ and independent (over the frequency k), normally distributed errors $N_Z(k)$, whereas SGTLS and SBTLS require $M \geq 2$ and the noise assumptions are exactly the same as when the noise model is known. The picture is somewhat more complicated for SSUB: the estimation of the poles requires $M \geq 2$ and the noise assumptions are the same as when the noise model is known, while the estimation of the zeros requires $M \geq 4$ and independent (over the frequency k), normally distributed errors $N_Z(k)$.
5. Even if it is known that the input-output errors are uncorrelated, the sample noise (co)variance must be estimated, otherwise SML is no longer consistent. This is also the case for the SSUB estimates of the zeros, but not of the poles.
6. All the estimates are sensitive to the assumption that the repeated experiments are independent. On top of that, the uncertainty of the SBTLS estimate is sensitive to the quality of the parametric noise model in the left weighting.

If the sample noise (co)variances are regularized for $M = 2, 3$, then the estimate $\hat{\theta}_{\text{SML}}(Z)$ satisfies Theorem 7.21 and is "practically consistent." The same can be done for the SSUB estimates of B and D.

8.9 IDENTIFICATION OF PARAMETRIC NOISE MODELS

Without an additional assumption, it is impossible to identify a (non)parametric noise model within an errors-in-variables framework. A first possibility consists of using periodic excitation signals as discussed in Sections 8.1 to 8.8. A second possibility, which is often applicable in control applications, consists of assuming that the arbitrary input is exactly known $u(t) = u_0(t)$ (see Figure 8-5). To avoid the mathematically difficult problem of treating continuous-time random processes, $n_y(t) = n_p(t) + m_y(t)$ is modeled at the sampling instances as filtered white discrete-time noise (see Section 5.7.3). Solving the resulting combined plant-noise transfer function models (5-64) and (5-65) for $E(k)$ defines the prediction error $\varepsilon(\Omega_k, \theta)$ ($\Omega = s$ or z^{-1})

$$\varepsilon(\Omega_k, \theta) = H^{-1}(z_k^{-1}, \theta)(Y(k) - G(\Omega_k, \theta)U(k) - T(\Omega_k, \theta) - T_H(z_k^{-1}, \theta)) \tag{8-48}$$

which is the difference between the noisy output $Y(k)$ and the one-step-ahead prediction $\hat{Y}(\Omega_k, \theta)$ of the noisy output

$$\begin{aligned}\hat{Y}(\Omega_k, \theta) = {} & H^{-1}(z_k^{-1}, \theta)(G(\Omega_k, \theta)U(k) + T(\Omega_k, \theta) + T_H(z_k^{-1}, \theta)) \\ & + (1 - H^{-1}(z_k^{-1}, \theta))Y(k)\end{aligned} \tag{8-49}$$

(Ljung, 1999). In (8-48) and (8-49) the input $U(k)$ and output $Y(k)$ DFT spectra stem from a time domain experiment (see Section 7.2); the plant transfer function $G(\Omega, \theta)$ and the plant transient term $T(\Omega, \theta)$ can take any parameterization of Sections 5.2 and 5.3; and the noise transfer function $H(z^{-1}, \theta)$ and the noise transient term $T_H(z^{-1}, \theta)$ are parameterized as in Section 5.7.3.

The prediction error (PE) method minimizes

$$V_{\text{PE}}(\theta, Z) = \sum_{k \in \mathbb{F}} |\varepsilon(\Omega_k, \theta)|^2 \tag{8-50}$$

w.r.t. θ, with $\mathbb{F}$ a subset of the DFT frequencies $k = 0, 1, \ldots, N-1$, F the number of frequencies in the set $\mathbb{F}$, and $\varepsilon(\Omega_k, \theta)$ as in (8-48). The following cases can be distinguished.

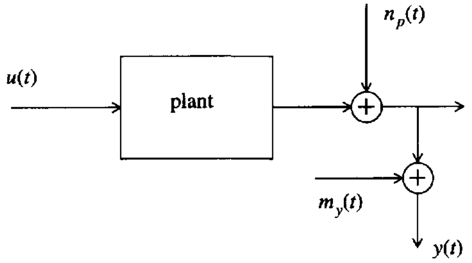


Figure 8-5. General output error framework: the arbitrary input is exactly known, $u(t) = u_0(t)$, and the output is disturbed by process noise $n_p(t)$ and measurement noise $m_y(t)$.

1. For $\Omega = z^{-1}$ and $\mathbb{F} = \{0, 1, ..., N-1\}$ (8-50) is the frequency domain equivalent of the prediction error method used in classical time domain identification. We refer the reader to Ljung (1999) for an excellent account of these techniques. According to the parameterization of the plant and the noise model in (8-48), it boils down to the identification of an ARX, ARMAX, ARMA, OE, or BJ model structure (see Section 5.7.3).
2. For $\Omega = z^{-1}$ and $\mathbb{F} \neq \{0, 1, ..., N-1\}$ (8-50) is the prediction error method with noncausal filtering (removal of DFT lines).
3. For $\Omega = s$, (8-50) reduces to the identification of a hybrid BJ (Box-Jenkins) model. With some additional effort (8-50) also allows hybrid ARMAX modeling. The denominators of the plant and noise model are then parameterized in their poles, and the poles of $H(z^{-1}, \theta)$ are related to those of $G(s, \theta)$ by the impulse invariant transformation $z = \exp(sT_s)$. Starting values are obtained via the hybrid BJ model.

A comprehensive study of the properties of the minimizer $\hat{\theta}_{\text{PE}}(Z)$ of (8-50) with $\Omega = z^{-1}$ and $\mathbb{F} = \{0, 1, ..., N-1\}$ (case 1) can be found in Ljung (1999), and under exactly the same assumptions as in Ljung (1999), the minimizer $\hat{\theta}_{\text{PE}}(Z)$ of (8-50) has exactly the same properties as the prediction error method. To study the general case with the aid of Theorem 7.21, we must add the following assumptions to Sections 7.6.6, 7.6.7 and 7.6.8 (assumptions for consistency, bias, and efficiency).

Assumption 8.12 (Generalized Output Error Framework): The input $u(t)$ is observed without measurement errors $m_u(t) = 0$.

Assumption 8.13 (Existence of a True Linear Discrete-Time Noise Model): There is an identifiable parameterization $\theta_0 \in \Theta_r$ such that $H(z_k^{-1}, \theta_0)E(k) + T_H(z_k^{-1}, \theta_0)$, with $H(z^{-1}, \theta_0)$ a stable and inversely stable monic ($c_0 = d_0 = 1$) filter, represents the true output error $N_Y(k)$.

Assumption 8.14 (Selected DFT Frequencies): The set $\mathbb{F}$ is the union of disjoint sets of indices k such that the corresponding z_k values of each set cover uniformly the unit circle:

$$\mathbb{F} = \bigcup_{r=1}^{R} \mathbb{F}_r \text{ with } \mathbb{F}_i \cap \mathbb{F}_j = \varnothing \text{ for } i \neq j$$
$$\mathbb{F}_0 = \{0, 1, ..., N-1\}$$
$$\mathbb{F}_r = \{k_m = o_r + \Delta_r m, m = 0, 1, ..., N_r - 1 \,|\, o_r, \Delta_r, N_r \in \mathbb{F}_0 \text{ and } \Delta_r N_r = N\}$$

Theorem 8.15 (Asymptotic Properties $\hat{\theta}_{\text{PE}}(Z)$ in Open Loop Identification): Consider the one-step-ahead predictor (8-49) with any identifiable parameterization of Sections 5.2 and 5.7.3. Let the assumptions of Section 7.6 be valid, where Assumptions 8.12, 8.13, and 8.14 are added to Sections 7.6.6, 7.6.7 and 7.6.8. The minimizer $\hat{\theta}_{\text{PE}}(Z)$ of (8-50) then has the asymptotic properties of Theorem 7.21 with $V_F(\theta, Z) = V_{\text{PE}}(\theta, Z)/F$ and with the following extensions:

1. If the plant and noise models are independently parameterized, then the estimated plant model parameters $\hat{a}_{\text{PE}}(Z)$, $\hat{b}_{\text{PE}}(Z)$ are consistent, even if Assumptions 8.13 and 8.14 are not satisfied: a true noise model either does not exist or does not belong to the model set (Assumption 8.13 is not fulfilled) or is inconsistently estimated (Assumption 8.14 is not fulfilled).

2. The estimates are inconsistent if the input is observed with measurement errors (Assumption 8.12 is not fulfilled).
3. In case of plant model errors, the limit values $a_{*\mathrm{PE}}$ and $b_{*\mathrm{PE}}$ depend, in general, on the noise level.

Proof. See Appendix 8.N. □

Note that the limit values $a_{*\mathrm{PE}}$ and $b_{*\mathrm{PE}}$ are, in general, not the noiseless solution, whereas the limit value θ_* of the SML, SGTLS, SBTLS, and SSUB estimates is the noiseless solution. Only for OE model structures, $H(z^{-1}, \theta) = 1$, $a_{*\mathrm{PE}}$ and $b_{*\mathrm{PE}}$ are independent of the noise level (Assumption 8.12 should be satisfied, see Appendix 8.N). The study of the constraint (in)dependence is somewhat more subtle than for the other estimators. Indeed, the estimation of the noise model is consistent only if $H(z^{-1}, \theta)$ is monic $c_0 = d_0 = 1$ (see the proof of Theorem 8.15). If the plant and the noise model are independently parameterized (BJ or OE model structure), then the prediction error cost function (8-50) satisfies $V_{\mathrm{PE}}(\lambda a, \lambda b, \lambda i, Z) = V_{\mathrm{PE}}(a, b, i, Z)$ so that the estimate $\hat{\theta}_{\mathrm{PE}}(Z)$ is independent of the particular constraint chosen, for example, $a_i = 1$, or $b_i = 1$, or $\|a\|_2^2 + \|b\|_2^2 = 1$ (quick tool number 4). For the other model structures (ARX, ARMA, and ARMAX) the consistency constraints on the noise model parameters $c_0 = d_0 = 1$ make the problem identifiable.

Although the input $u(t)$ is band-limited in the hybrid BJ model, the summation in the PE cost function (8-50) is taken over all the DFT frequencies, unless Assumption 8.14 is satisfied. Otherwise, the estimate of the noise model parameters would not be consistent. The estimate of the plant model parameters is consistent in open loop (Assumption 7.3 is satisfied) even if the summation is taken only over a subset of the DFT frequencies, for example, the bandwidth of the excitation. Assumption 8.14 allows filtering of the input and output DFT spectra (for example, removal of every second DFT line) without affecting the consistency of the noise model parameters.

Although the equivalent initial conditions of the plant and noise model are not consistently estimated, they are added in the cost function (8-50) to improve the finite sample behavior of the estimated plant and noise model parameters.

A disadvantage of the parametric noise model is that its quality (strongly) depends on quality of the estimated plant model (the noise model tries to follow every systematic deviation from the true plant model). This is not the case for the sample noise (co)variances because they are independent of the estimated plant model.

8.10 IDENTIFICATION IN FEEDBACK

Consider the linear feedback experiment of Figure 7-14 on page 243. For periodic reference signals $r(t)$, the SML (8-10), SGTLS (8-23), SBTLS (8-33), and SSUB (Section 8.6) estimates remain strongly consistent (see Theorems 8.3, 8.6, 8.8, and 8.10). The basic reason for this is that the sample mean and sample (co)variance calculation makes a natural separation between the random and the periodic parts of the input $u(t)$ and output $y(t)$ signals without requiring knowledge of $r(t)$.

For arbitrary reference signals $r(t)$, the technical difficulty is that the input is correlated with the process noise (Ljung, 1999). The prediction error estimate (8-50) of the plant model parameters is consistent if the true noise model belongs to the considered model set (see Theorem 8.19). Other solutions have been developed that do not require the construction of a consistent parametric noise model: see Forssell and Ljung (1999) and Van den Hof and Schrama (1995) for an overview of the classical time domain methods, which all require that the input and output signals are observed without errors ($m_u(t) = 0$, $m_y(t) = 0$), and see

Schoukens et al. (1999b) for the full errors-in-variables problem. Several of the solutions assume that the reference signal is exactly known and consider the controller and process noise as disturbing errors. The influence of these errors is eliminated by projection of the input and output signals on the reference signal. Note that in the periodic case the distinction between the signal originating from $r(t)$ and the disturbing noise can be made without explicit knowledge of the reference signal.

To study the properties of the prediction error estimate (8-50) in feedback with the aid of Theorem 7.21, we must add the following assumptions to Sections 7.6.6, 7.6.7 and 7.6.8 (assumptions for consistency, bias, and efficiency).

Assumption 8.16 (No Measurement Errors): The input and output signals are observed without measurement errors: $m_u(t) = 0$ and $m_y(t) = 0$ (see Figure 7-14 on page 243).

Assumption 8.17 (Delay Assumption): Either the plant or the controller contains a delay: $\lim_{z \to \infty} G(z^{-1}, \theta) = 0$ or $\lim_{z \to \infty} C_0(z^{-1}) = 0$.

Assumption 8.18 (Existence of a True Noise Model): There is an identifiable parameterization $\theta_0 \in \Theta_r$ such that $H(z_k^{-1}, \theta_0)E(k) + T_H(z_k^{-1}, \theta_0)$, with $H(z^{-1}, \theta_0)$ an inversely stable monic ($c_0 = d_0 = 1$) filter, represents the true output error $N_Y(k)$.

Note that these assumptions are slightly different from the open loop case; for example, the output must be observed without errors and the noise model may be unstable. Note also that the condition $\lim_{z \to \infty} G(z^{-1}, \theta) = 0$ in Assumption 8.17 is satisfied for step-invariant transfer function models $G(z^{-1}, \theta) = (1 - z^{-1})Z\{G(s)/s\}$ of most physical continuous-time plants $G(s)$ (see also Example 5.3).

Theorem 8.19 (Asymptotic Properties $\hat{\theta}_{\text{PE}}(Z)$ in Closed Loop Identification): Consider the identification of a plant within a linear stabilizing feedback loop (see Figure 7-14 on page 243). Consider, furthermore, the one-step-ahead predictor (8-49) with any identifiable parameterization of Sections 5.2 and 5.7.3. Let the assumptions of Section 7.6 be valid, where Assumptions 8.14 and 8.16 to 8.18 are added to Sections 7.6.6, 7.6.7 and 7.6.8 and where the independence of the true input and the disturbing noise has been removed in Assumption 7.3. The minimizer $\hat{\theta}_{\text{PE}}(Z)$ of (8-50) then has the asymptotic properties of Theorem 7.21 with $V_F(\theta, Z) = V_{\text{PE}}(\theta, Z)/F$, and with the following extensions

1. Continuous-time plant models $\Omega = s$: the estimate $\hat{\theta}_{\text{PE}}(Z)$ is inconsistent.
2. Discrete-time plant models $\Omega = z^{-1}$:
   - 2a. Stable plants: the estimate $\hat{\theta}_{\text{PE}}(Z)$ is strongly consistent for any model structure of Section 5.7.3, for example, ARX, ARMAX, OE, and BJ.
   - 2b. Unstable plants: the estimate $\hat{\theta}_{\text{PE}}(Z)$ is strongly consistent for ARX and ARMAX models and is inconsistent for OE and BJ models.
3. If only the process noise $n_p(t)$ excites the plant ($r(t) = 0$ and $n_c(t) = 0$) then the estimated plant model converges to $-1/C_0(\Omega)$, with $C_0(\Omega)$ the controller transfer function.
4. In case of plant model errors, the limit values $a_{*\text{PE}}$ and $b_{*\text{PE}}$ depend, in general, on the noise level.

Proof. See Appendix 8.N. □

In contrast to the open loop case, the consistency of the plant model parameters requires the consistency of the noise model (compare extension 1 of Theorem 8.15 with extension 2 of Theorem 8.19), no output measurement errors are allowed (compare Assumption 8.16 with Assumption 8.12), and either the plant or the controller should contain a delay (Assumption 8.17). The limit values $a_{*\text{PE}}$ and $b_{*\text{PE}}$ are independent of the noise level for OE model structures, $H(z^{-1}, \theta) = 1$, if the true noise model is white, $H(z^{-1}, \theta_0) = 1$, and if Assumptions 8.14 and 8.16 to 8.18 are satisfied (see Appendix 8.N).

If the plant is unstable, then the prediction error estimate $\hat{\theta}_{\text{PE}}(Z)$ (8-50) of OE and BJ models is inconsistent (extension 2b of Theorem 8.19). The estimate can be made consistent by multiplying the noise models in the OE and BJ model structures with the following monic all-pass filter:

$$H_a(z^{-1}, \theta) = \frac{F_a^*(z^{-1}, \theta)}{F_a(z^{-1}, \theta)} = \frac{\sum_{r=0}^{n_f} f_{n_f - r} f_{n_f}^{-1} z^{-r}}{\sum_{r=0}^{n_f} f_r z^{-r}} \text{ with } f_0 = 1 \tag{8-51}$$

where $F_a(z^{-1}, \theta)$ is the monic unstable part of $A(z^{-1}, \theta)$, and $F_a^*(z^{-1}, \theta)$ is the monic stabilized $F_a(z^{-1}, \theta)$ polynomial (see Forssell and Ljung, 2000a and Appendix 8.N).

8.11 APPENDIXES

Appendix 8.A: Expected Value and Variance of the Inverse of Chi-Square Random Variable

Let x be a $\chi^2(n)$ distributed random variable. Starting from the results of an F-distribution, it is found that

$$\mathscr{E}\{x^{-1}\} = \frac{1}{n-2} \quad \text{and} \quad \text{var}(x^{-1}) = \frac{2}{(n-2)^2(n-4)} \tag{8-52}$$

Appendix 8.B: First and Second Moments of the Ratio of the True and the Sample Variance of the Equation Error

Consider $c_k(\theta)$ defined in (8-17). Under Assumptions 8.1 and 8.2, the expected value $\mathscr{E}\{c_k(\theta)\}$ is independent of θ and equals:

$$\mathscr{E}\{c_k(\theta)\} = \frac{M-1}{M-2}, \text{ for } M > 2 \quad \text{and} \quad \text{var}(c_k(\theta)) = \frac{(M-1)^2}{(M-2)^2(M-3)}, \text{ for } M > 3 \tag{8-53}$$

Proof. Consider

$$\begin{aligned} 2(M-1)\frac{1}{c_k(\theta)} &= 2(M-1)\frac{\hat{\sigma}_{\hat{e}}^2(\theta, \Omega_k)}{\sigma_{\hat{e}}^2(\theta, \Omega_k)} \\ &= 2(M-1)\frac{\frac{1}{M}\frac{1}{M-1}\sum_{l=1}^{M}\left|e(\Omega_k, \theta, Z^{[l]}(k)) - \hat{e}(\Omega_k, \theta, \hat{Z}(k))\right|^2}{\sigma_e^2(\theta, \Omega_k)/M} \\ &= \sum_{l=1}^{M}(\text{Re}(z_k^{[l]}(\theta)))^2 + (\text{Im}(z_k^{[l]}(\theta)))^2 \end{aligned} \tag{8-54}$$

with $z_k^{[l]}(\theta) = (e(\Omega_k, \theta, Z^{[l]}(k)) - \hat{e}(\Omega_k, \theta, \hat{Z}(k)))/(\sqrt{2}\sigma_e(\theta, \Omega_k))$. It follows that (8-54) consist of the sum of two independent central χ^2-distributed variables with $M-1$ degrees of freedom resulting in a χ^2-distribution with $2M-2$ degrees of freedom. Using the results of Appendix 8.A, we find

$$\mathcal{E}\{c_k(\theta)\} = \frac{M-1}{M-2} \quad \text{and} \quad \operatorname{var}(c_k(\theta)) = \frac{(M-1)^2}{(M-2)^2(M-3)} \tag{8-55}$$

Note that the moments of $c_k(\theta)$ are θ independent due to the fact that $\operatorname{Re}(z_k^{[l]}(\theta))$ and $\operatorname{Im}(z_k^{[l]}(\theta))$ have a θ-independent distribution.

Corollary 8.20: $\hat{\sigma}_{\hat{e}}^2/\sigma_{\hat{e}}^2$ has a $\chi^2(2M-2)$ distribution. The moments $\mathcal{E}\{1/\hat{\sigma}_{\hat{e}}^k\}$ are finite if $M \geq k/2 + 2$.

Proof. The $\chi^2(2M-2)$ distributions follows directly from (8-54). The second claim is a direct result of the $\chi^2(\nu)$ distribution:

$$\chi^2(\nu) \sim \left(\frac{1}{2}x^2\right)^{\nu/2-1} e^{-x^2/2} \tag{8-56}$$

Appendix 8.C: Calculation of Some First- and Second-Order Moments

The results in this appendix depend strongly on the number of degrees of freedom that appear in the χ^2 of the sample variance. M measured periods result in M observations of the Fourier coefficient at a given frequency. After subtracting the mean value, this results in $M-1$ degrees of freedom for the sample variance on real or imaginary parts, and those are quadratically combined so that finally the sample variance has $2(M-1)$ degrees of freedom.

For notational simplicity, the dependence of $c_k(\theta)$ and $\hat{\sigma}_{\hat{e}}(\Omega_k, \theta)$ on θ, k, and Ω_k is omitted.

(i) c has finite first- and second-order moments for $M > 3$. From (8-52) and (8-53):

$$\mathcal{E}\{c\} = \frac{M-1}{M-2}$$

$$\mathcal{E}\{c^2\} = (\mathcal{E}\{c\})^2 + \sigma_c^2 = \left(\frac{M-1}{M-2}\right)^2 + 2\frac{4(M-1)^2}{(2M-4)^2(2M-6)} = \frac{(M-1)^2}{(M-2)(M-3)}$$

(ii) c' has finite first- and second-order moments for $M \geq 5$.

Because $\mathcal{E}\{c\}$ is independent of θ, it follows directly that $\mathcal{E}\{c'\} = 0$.

$\mathcal{E}\{(\partial c/\partial\theta_{[l]})^2\} = \mathcal{E}\{[\hat{\sigma}_{\hat{e}}^{-2}\partial\sigma_{\hat{e}}^2/\partial\theta_{[l]} - (\sigma_{\hat{e}}^2/\hat{\sigma}_{\hat{e}}^4)\partial\hat{\sigma}_{\hat{e}}^2/\partial\theta_{[l]}]^2\}$ with $\theta_{[l]}$ the lth element of θ. This is an expression of the form $\mathcal{E}\{(a-b)^2\}$. Applying the inequality $-\mathcal{E}\{a^2\} - \mathcal{E}\{b^2\} \leq 2\,\mathcal{E}\{ab\} \leq \mathcal{E}\{a^2\} + \mathcal{E}\{b^2\}$ shows that it is sufficient to prove that the second-order moments of a and b are finite. This is obvious for a because c_k has finite moments. Bounding $\mathcal{E}\{b^2\}$ is more involved. Calculating the derivative of $\hat{\sigma}_{\hat{e}}^2$ w.r.t. $\theta_{[l]}$, a parameter of the numerator $B(\Omega, \theta)$, gives

$$\partial\hat{\sigma}^2_{\hat{e}}/\partial\theta_{[l]} = \frac{2}{M}\mathrm{Re}((B\hat{\sigma}^2_{\hat{U}} - A\hat{\sigma}^2_{\hat{Y}U})\overline{\Omega}^l)$$

and

$$\begin{aligned}(\partial\hat{\sigma}^2_{\hat{e}}/\partial\theta_{[l]})^2 &= 4[\mathrm{Re}(B\hat{\sigma}^2_{\hat{U}} - A\hat{\sigma}^2_{\hat{Y}U})\overline{\Omega}^l]^2/M^2 \\ &\le 4|\Omega|^{2l}\left|B\hat{\sigma}^2_{\hat{U}} - A\hat{\sigma}^2_{\hat{Y}U}\right|^2/M^2 \\ &\le 4|\Omega|^{2l}\hat{\sigma}^2_{\hat{U}}(|B|^2\hat{\sigma}^2_{\hat{U}} + |A|^2\left|\hat{\sigma}^2_{\hat{Y}U}\right|^2/\hat{\sigma}^2_{\hat{U}} - 2\mathrm{Re}(A\overline{B}\hat{\sigma}^2_{\hat{Y}U}))/M^2 \\ &\le 4|\Omega|^{2l}\hat{\sigma}^2_{\hat{U}}\hat{\sigma}^2_{\hat{e}}/M\end{aligned}$$

where the last inequality is due to $\left|\hat{\sigma}^2_{\hat{Y}U}\right|^2 \le \hat{\sigma}^2_{\hat{U}}\hat{\sigma}^2_{\hat{Y}}$ and (8-11). Following the same lines, we find $(\partial\hat{\sigma}^2_{\hat{e}}/\partial\theta_{[l]})^2 \le 4|\Omega|^{2l}\hat{\sigma}^2_{\hat{Y}}\hat{\sigma}^2_{\hat{e}}/M$ if θ_l is a parameter of the denominator $A(\Omega, \theta)$. Without loss of generality, the first situation will be considered here and it will be shown that the second moment of

$$b^2 = \sigma^4_{\hat{e}}(1/\hat{\sigma}^2_{\hat{e}})^4(\partial\hat{\sigma}^2_{\hat{e}}/\partial\theta_{[l]})^2 \le \frac{4}{M}\sigma^4_{\hat{e}}|\Omega|^{2l}\hat{\sigma}^2_{\hat{U}}/(\hat{\sigma}^2_{\hat{e}})^3$$

is bounded. For notational simplicity, the dependence on k will be omitted.

The sample variances $(\hat{\sigma}^2_{\hat{U}}, \hat{\sigma}^2_{\hat{e}})$ have a Wishart distribution that also depends on the cross-correlation $\hat{\sigma}^2_{\hat{U}\hat{e}}$ and is noticed as $dF(\hat{\sigma}^2_{\hat{U}}, \hat{\sigma}^2_{\hat{e}}, \rho)$ with $\rho = \sigma^2_{\hat{U}\hat{e}}/\sqrt{\sigma^2_{\hat{U}}\sigma^2_{\hat{e}}}$ (Kendall and Stuart, 1979). The domain of the integral to calculate the expected value of b^2 can be split into two parts: with respect to the variable $\hat{\sigma}^2_{\hat{e}}$: $D_\varepsilon = \{\hat{\sigma}^2_{\hat{e}} | \hat{\sigma}^2_{\hat{e}} < \varepsilon\}$ and $D_r = \{\hat{\sigma}^2_{\hat{e}} | \hat{\sigma}^2_{\hat{e}} \ge \varepsilon\}$. The integral over D_r will be finite because $\sigma^2_{\hat{U}}$ has finite moments and the denominator is bounded. It can be shown after some calculations that on D_ε the marginal density function $dH(\hat{\sigma}^2_{\hat{U}}, \hat{\sigma}^2_{\hat{e}}) = \int_{D_\rho} dF(\hat{\sigma}^2_{\hat{U}}, \hat{\sigma}^2_{\hat{e}}, \rho)$ is bounded by

$$dH(\hat{\sigma}^2_{\hat{U}}, \hat{\sigma}^2_{\hat{e}}) \le C(\nu)e^{(|\rho|\sigma_U\sqrt{\varepsilon})(\nu+1)/2}dG(\nu, \hat{\sigma}^2_{\hat{U}})dG(\nu, \hat{\sigma}^2_{\hat{e}}) \tag{8-57}$$

Here $dG(\nu, \hat{\sigma}^2_{\hat{U}})$ is a χ^2 distribution with respect to $\hat{\sigma}^2_{\hat{U}}$ with ν degrees of freedom and similarly for $G(\nu, \hat{\sigma}^2_{\hat{e}})$, and $C(\nu)$ a constant with respect to $(\hat{\sigma}^2_{\hat{U}}, \hat{\sigma}^2_{\hat{e}})$, depending on ν. Because the density function $dG(\nu, \hat{\sigma}_{\hat{e}})$ of a χ^2 distribution is proportional to $(\hat{\sigma}^2_{\hat{e}}/2)^{\nu/2-1}e^{-\hat{\sigma}^2_{\hat{e}}/2}$ it is clear that the expected value of b^2 over D_ε will be bounded if $M \ge 5$ (Corollary 8.20).

Remark. To obtain (8-57), it had to be assumed that $|\rho| < 1 - \delta$, with $\delta > 0$. For the singular case that $|\rho| = 1$, it is straightforwardly shown that c_k becomes θ independent and, hence, the derivatives are zero.

(iii) $c'' = \sigma^{2''}_{\hat{e}}/\hat{\sigma}^2_{\hat{e}} - (\sigma^{2'T}_{\hat{e}}\hat{\sigma}^{2'}_{\hat{e}} + \hat{\sigma}^{2'T}_{\hat{e}}\sigma^{2'}_{\hat{e}})/\hat{\sigma}^4_{\hat{e}} - \sigma^2_{\hat{e}}\hat{\sigma}^{2''}_{\hat{e}}/\hat{\sigma}^4_{\hat{e}} + 2(\hat{\sigma}^{2'T}_{\hat{e}}\hat{\sigma}^{2'}_{\hat{e}})\sigma^2_{\hat{e}}/\hat{\sigma}^6_{\hat{e}}$ has finite first- and second-order moments for $M \ge 6$.

The proof is completely similar to the previous one, noticing that $(\partial\hat{\sigma}^2_{\hat{e}}/(\partial\theta_{[r]}\partial\theta_{[s]})) = 2\hat{\sigma}^2_{\hat{U}}\mathrm{Re}(\Omega^r\overline{\Omega}^s)$, $2\hat{\sigma}^2_{\hat{Y}U}\mathrm{Re}(\Omega^r\overline{\Omega}^s)$ or $2\hat{\sigma}^2_{\hat{Y}}\mathrm{Re}(\Omega^r\overline{\Omega}^s)$, and $\left|\hat{\sigma}^2_{\hat{Y}U}\right|^2 \le (\hat{\sigma}^4_{\hat{U}} + \hat{\sigma}^4_{\hat{Y}})/2$. This time contributions of $(1/\hat{\sigma}^2_{\hat{e}})^4$ should be bounded, requiring that $M \ge 6$ (Corollary 8.20).

Appendix 8.D: Proof of Theorem 8.3

Because the noise is, by assumption, independent over the frequency (Assumption 8.2), the proof of Theorem 7.21 for a frequency domain experiment can be applied to the cost function (8-10) provided that the necessary moments of the cost function and its derivatives w.r.t. θ are finite in Θ_r. We need the first- and second-order moments of $V_{\text{SML}}(\theta, Z)$ for properties 1, 5, and 6 of Theorem 7.21; in addition, the first- and second-order moments of $V_{\text{SML}}'(\theta, Z)$ and $V_{\text{SML}}''(\theta, Z)$ for properties 2 and 6 of Theorem 7.21; in addition, the first- and second-order moments of $V_{\text{SML}}'''(\theta, Z)$ for properties 3, 6, and 7 of Theorem 7.21; and in addition, the third-order moments of $V_{\text{SML}}(\theta, Z)$ and $V_{\text{SML}}'(\theta, Z)$ for property 4 of Theorem 7.21 (the moment $2 + \delta$ in Assumption 7.13 is bounded above by a third-order moment). The moments of the cost function $V_{\text{SML}}(\theta, Z)$ and its derivatives w.r.t. θ exist, if the moments of $|\hat{\varepsilon}(\Omega_k, \theta, \hat{Z}(k))|^2$ and its derivatives w.r.t. θ exist. These moments are calculated in the following. For notational simplicity, we dropped the arguments of $|\hat{\varepsilon}|^2$, $|\hat{e}|^2$ and $\hat{\sigma}^2$.

1. The moments $\mathcal{E}\{|\hat{\varepsilon}|^{2l}\}$ are finite with $l = 1, M \geq 3$; $l = 2, M \geq 4$; $l = 3, M \geq 5$.

   Proof: It follows directly from Corollary 8.20 in Appendix 8.B.

2. The moments $\mathcal{E}\{(\partial|\hat{\varepsilon}^2|/\partial\theta_{[r]})^l\}$ are finite with $l = 1, M \geq 4$; $l = 2, M \geq 5$; $l = 3, M \geq 6$.

   Proof:

$$\frac{\partial|\hat{\varepsilon}^2|}{\partial\theta_{[r]}} = \frac{1}{\hat{\sigma}_{\hat{e}}^2}\frac{\partial|\hat{e}^2|}{\partial\theta_{[r]}} - |\hat{e}^2|\frac{1}{\hat{\sigma}_{\hat{e}}^4}\frac{\partial\hat{\sigma}_{\hat{e}}^2}{\partial\theta_{[r]}} \tag{8-58}$$

   In this expression, it is the last term that is critical in bounding its expected value. Using the same technique as in Appendix 8.C, and noticing that $\hat{e}$ is independent of $\hat{\sigma}_{\hat{e}}$, it turns out that it is enough to bound $\mathcal{E}\{1/\hat{\sigma}_{\hat{e}}^{3l}\}$. The claim then follows directly from Corollary 8.20.

3. The moments $\mathcal{E}\{[\partial^2|\hat{\varepsilon}^2|/(\partial\theta_{[r]}\partial\theta_{[s]})]^l\}$ are finite with $l = 1, M \geq 4$; $l = 2, M \geq 6$.

   Proof: This time the term that is critical in bounding the expected value becomes:

$$|\hat{e}^2|\frac{1}{\hat{\sigma}_{\hat{e}}^6}\frac{\partial\hat{\sigma}_{\hat{e}}^2}{\partial\theta_{[r]}}\frac{\partial\hat{\sigma}_{\hat{e}}^2}{\partial\theta_{[s]}} \tag{8-59}$$

   Its lth order moment is again bounded if $\mathcal{E}\{1/\hat{\sigma}_{\hat{e}}^{4l}\}$ is finite. The claim then follows again from Corollary 8.20.

4. The moments $\mathcal{E}\{(\partial^3|\hat{\varepsilon}^2|/(\partial\theta_{[r]}\partial\theta_{[s]}\partial\theta_{[t]}))^l\}$ are finite with $l = 1, M \geq 5$; $l = 2, M \geq 7$.

Proof. This time the critical term in bounding the expected value becomes:

$$|\hat{e}^2|\frac{1}{\hat{\sigma}_{\hat{e}}^8}\frac{\partial\hat{\sigma}_{\hat{e}}^2}{\partial\theta_{[r]}}\frac{\partial\hat{\sigma}_{\hat{e}}^2}{\partial\theta_{[s]}}\frac{\partial\hat{\sigma}_{\hat{e}}^2}{\partial\theta_{[t]}} \tag{8-60}$$

Its lth order moment is again bounded if $\mathcal{E}\{1/\hat{\sigma}_{\hat{e}}^{5l}\}$ is finite. The claim then follows again from Corollary 8.20.

Appendix 8.E: Approximation of the Derivative of the Cost Function

Replacing N_Z by υN_Z (C_{N_Z} by $\upsilon^2 C_{N_Z}$) and $\hat{e}(\Omega_k, \tilde{\theta}(Z_0), \hat{Z}_0(k))$ by $\mu\hat{e}(\Omega_k, \tilde{\theta}(Z_0), \hat{Z}_0(k))$ makes it possible to analyze $V_{\text{SML}}'(\theta, Z)$ for small noise levels $\upsilon \to 0$ (large signal-to-noise ratios) and small model errors $\mu \to 0$ (see Section 7.5, quick analysis tool 6)

$$V_{\text{SML}}'(\tilde{\theta}(Z_0), Z) \approx \sum_{k=1}^{F} c_k(\tilde{\theta}(Z_0)) d_k'(\tilde{\theta}(Z_0)) \tag{8-61}$$

Proof. Using (8-17), $V_{\text{SML}}'(\theta, Z)$ can be written as

$$\frac{1}{F} V_{\text{SML}}'(\tilde{\theta}(Z_0), Z) = \frac{1}{F}\sum_{k=1}^{F} c_k'(\tilde{\theta}(Z_0)) d_k(\tilde{\theta}(Z_0)) + \frac{1}{F}\sum_{k=1}^{F} c_k(\tilde{\theta}(Z_0)) d_k'(\tilde{\theta}(Z_0)) \tag{8-62}$$

Because $c_k(\tilde{\theta}(Z_0)) = O(\upsilon^0\mu^0)$, $c_k'(\tilde{\theta}(Z_0)) = O(\upsilon^0\mu^0)$, $d_k(\tilde{\theta}(Z_0)) = \upsilon^{-2}O(\mu^2 + \mu\upsilon + \upsilon^2)$, and $d_k'(\tilde{\theta}(Z_0)) = \upsilon^{-2}O(\mu + \upsilon)$, it follows that for small model errors $\mu \to 0$ and large signal-to-noise ratios $\upsilon \to 0$, the term $c_k(\tilde{\theta}(Z_0)) d_k'(\tilde{\theta}(Z_0))$ in (8-62) dominates over $c_k'(\tilde{\theta}(Z_0)) d_k(\tilde{\theta}(Z_0))$. Summing over the frequencies will not change this behavior because the two sums in the right-hand side of (8-62) are both an $O_{\text{p}}(F^{-1/2})$. The last statement follows from the fact that the terms in both sums are independent zero mean random variables with bounded first- and second-order moments (see Section 14.9, version 2 of the law of large numbers).

Appendix 8.F: Loss in Efficiency of the Sample Estimator

In both sections we use the following properties: (i) $V_{\text{ML}}(\theta, Z) = \sum_{k=1}^{F} d_k(\theta)$ and (ii) $c_k(\theta)$ and $d_k(\theta)$ are mutually independent random variables (see (8-17)). Property (i) follows directly from the fact that $d_k(\theta)$ contains the true noise (co)variances (see (8-17)). Property (ii) is shown as follows: $c_k(\theta)$ depends on the sample variance, while $d_k(\theta)$ depends on the sample mean. It is well known that the sample variance and the sample mean are independent random variables for normally distributed noise (Stuart and Ord, 1987). Hence, this is also the case for $c_k(\theta)$ and $d_k(\theta)$.

8.F.1 Approximate Expression for the Parameter Deviations. Under the assumptions of Theorem 8.3, formula (7-25) of Theorem 7.21 is valid for $M \geq 7$ (Theorem 8.3)

$$\begin{aligned} \hat{\theta}_{\text{SML}}(Z) &= \tilde{\theta}_{\text{SML}}(Z_0) + \delta_{\theta\text{SML}}(Z) + O_{\text{p}}(F^{-1}) \\ \delta_{\theta\text{SML}}(Z) &= -V_{\text{SML}}''^{-1}(\tilde{\theta}_{\text{SML}}(Z_0)) V_{\text{SML}}'^{T}(\tilde{\theta}_{\text{SML}}(Z_0), Z) \end{aligned} \tag{8-63}$$

Using notation (8-17), the Hessian $V_{\text{SML}}''(\theta)$ can be written as

$$V_{\text{SML}}''(\theta) = \sum_{k=1}^{F} \mathcal{E}\{c_k''(\theta) d_k(\theta) + 2c_k'^{T}(\theta) d_k'(\theta) + c_k(\theta) d_k''(\theta)\} \tag{8-64}$$

Because $d_k(\theta)$ is independent of $c_k(\theta)$, $\mathcal{E}\{c_k(\theta)\} = (M-1)/(M-2)$, $\mathcal{E}\{c_k'(\theta)\} = 0$ and $\mathcal{E}\{c_k''(\theta)\} = 0$ (see Appendix 8.C), and $V_{\text{ML}}(\theta, Z) = \sum_{k=1}^{F} d_k(\theta)$, (8-64) becomes

$$V_{\mathrm{SML}}''(\theta) = \frac{M-1}{M-2}\sum_{k=1}^{F}\mathcal{E}\{d_k''(\theta)\} = \frac{M-1}{M-2}V_{\mathrm{ML}}''(\theta) \tag{8-65}$$

Using $\tilde{\theta}_{\mathrm{SML}}(Z_0) = \tilde{\theta}_{\mathrm{ML}}(Z_0)$ (Theorem 8.4), approximation (8-61) (see Appendix 8.E), we finally get

$$\begin{aligned}\delta_{\theta\mathrm{SML}}(Z) &\approx -\frac{M-2}{M-1}V_{\mathrm{ML}}''^{-1}(\tilde{\theta}_{\mathrm{ML}}(Z_0))\sum_{k=1}^{F}c_k(\tilde{\theta}_{\mathrm{ML}}(Z_0))d_k'^{T}(\tilde{\theta}_{\mathrm{ML}}(Z_0))\\ \mathcal{E}\{\delta_{\theta\mathrm{SML}}(Z)\} &= -\frac{M-2}{M-1}V_{\mathrm{ML}}''^{-1}(\tilde{\theta}_{\mathrm{ML}}(Z_0))V_{\mathrm{ML}}'^{T}(\tilde{\theta}_{\mathrm{ML}}(Z_0)) = 0\end{aligned} \tag{8-66}$$

Note that (8-66) is also valid in the presence of model errors.

8.F.2 An Approximate Expression for the Covariance Matrix. Using (8-66), $\mathcal{E}\{c_k^2(\theta)\} = (M-1)^2/((M-2)(M-3))$ for $M \geq 6$ (see Appendix 8.C), and the fact that $c_k(\theta)$ and $d_k(\theta)$ are mutually independent and independent over the frequency k, the following asymptotic ($F \to \infty$) expression is found:

$$\begin{aligned}\hat{C}_{\theta} &= \mathcal{E}\{\delta_{\theta\mathrm{SML}}(Z)\delta_{\theta\mathrm{SML}}^{T}(Z)\}\\ &= \frac{M-2}{M-3}V_{\mathrm{ML}}''^{-1}(\tilde{\theta}_{\mathrm{ML}}(Z_0))\sum_{k=1}^{F}\mathcal{E}\{d_k'^{T}(\tilde{\theta}_{\mathrm{ML}}(Z_0))d_k'(\tilde{\theta}_{\mathrm{ML}}(Z_0))\}V_{\mathrm{ML}}''^{-1}(\tilde{\theta}_{\mathrm{ML}}(Z_0))\end{aligned} \tag{8-67}$$

For independent (over the frequency) distributed errors, (7-26) reduces to

$$\begin{aligned}C_{\theta} &= \mathcal{E}\{\delta_{\theta\mathrm{ML}}(Z)\delta_{\theta\mathrm{ML}}^{T}(Z)\}\\ &= V_{\mathrm{ML}}''^{-1}(\tilde{\theta}_{\mathrm{ML}}(Z_0))\sum_{k=1}^{F}\mathcal{E}\{d_k'^{T}(\tilde{\theta}_{\mathrm{ML}}(Z_0))d_k'(\tilde{\theta}_{\mathrm{ML}}(Z_0))\}V_{\mathrm{ML}}''^{-1}(\tilde{\theta}_{\mathrm{ML}}(Z_0))\end{aligned} \tag{8-68}$$

where $V_{\mathrm{ML}}(\theta, Z) = \sum_{k=1}^{F}d_k(\theta)$. Hence, from (8-67) and (8-68) it follows that

$$\hat{C}_{\theta} \approx \frac{M-2}{M-3}C_{\theta} \tag{8-69}$$

for $M \geq 7$ ((8-66) is valid for $M \geq 7$), with C_{θ} the covariance matrix of the parameters when the noise (co)variances are exactly known.

Appendix 8.G: Mean and Variance of the Sample Cost in Its Global Minimum

Some precautions should be taken when calculating the expected value and the variance of $V_{\mathrm{SML}}(\hat{\theta}_{\mathrm{SML}}(Z), Z)$ since $c_k(\hat{\theta}_{\mathrm{SML}}(Z))$ and $d_k(\hat{\theta}_{\mathrm{SML}}(Z))$ (see (8-17)) are no longer independent because they both depend on $\hat{\theta}_{\mathrm{SML}}(Z)$. To get around this problem, a Taylor series expansion is made around the asymptotic value $\tilde{\theta}_{\mathrm{SML}}(Z_0) = \tilde{\theta}_{\mathrm{ML}}(Z_0)$ (Theorem 8.4), which is denoted as $\tilde{\theta}$ for simplicity of notation

$$V_{\mathrm{SML}}(\hat{\theta}_{\mathrm{SML}}(Z), Z) \approx V_{\mathrm{SML}}(\tilde{\theta}, Z) + V_{\mathrm{SML}}'(\tilde{\theta}, Z)\delta + \frac{1}{2}\delta^T V_{\mathrm{SML}}''(\tilde{\theta}, Z)\delta \tag{8-70}$$

with $\delta = \hat{\theta}_{\mathrm{SML}}(Z) - \tilde{\theta}_{\mathrm{ML}}(Z_0)$. Under the assumptions of Theorem 8.3, the Hessian $V_{\mathrm{SML}}''(\tilde{\theta}, Z)/F$ converges w.p. 1 at the rate $O_p(F^{-1/2})$ to its expected value $V_{\mathrm{SML}}''(\tilde{\theta})/F$ (proof: apply version 2 of the strong law of large numbers (14-69)). Hence, using (8-65), (8-70) can be written as

$$V_{\mathrm{SML}}(\hat{\theta}_{\mathrm{SML}}(Z), Z) \approx V_{\mathrm{SML}}(\tilde{\theta}, Z) + V_{\mathrm{SML}}'(\tilde{\theta}, Z)\delta + \frac{1}{2}\frac{M-1}{M-2}\delta^T V_{\mathrm{ML}}''(\tilde{\theta})\delta \tag{8-71}$$

Substituting approximation (8-66) for δ in (8-71) gives, using (8-65),

$$\begin{aligned} V_{\mathrm{SML}}(\hat{\theta}_{\mathrm{SML}}(Z), Z) &\approx V_{\mathrm{SML}}(\tilde{\theta}, Z) - \Delta V(\tilde{\theta}, Z) \\ \Delta V(\tilde{\theta}, Z) &= \frac{1}{2}\sum_{k=1}^{F} c_k(\tilde{\theta})d_k'(\tilde{\theta})\frac{M-2}{M-1}V_{\mathrm{ML}}''^{-1}(\tilde{\theta})\sum_{k=1}^{F} c_k(\tilde{\theta})d_k'^T(\tilde{\theta}) \end{aligned} \tag{8-72}$$

8.G.1 Calculation of the Mean Value. Using $\mathcal{E}\{c_k^2(\theta)\} = (M-1)^2/((M-2)(M-3))$ for $M \geq 6$ (see Appendix 8.C), and the fact that $c_k(\tilde{\theta})$ and $d_k(\tilde{\theta})$ are mutually independent over the frequency k, the expected value of the second term in the right-hand side of (8-72) equals

$$\begin{aligned} \Delta V(\tilde{\theta}, Z) &= \frac{1}{2}\frac{M-1}{M-3}\sum\nolimits_{k=1}^{F}\mathcal{E}\{d_k'(\tilde{\theta})V_{\mathrm{ML}}''^{-1}(\tilde{\theta})d_k'^T(\tilde{\theta})\} \\ &= \frac{1}{2}\frac{M-1}{M-3}\mathrm{trace}(V_{\mathrm{ML}}''^{-1}(\tilde{\theta})\sum\nolimits_{k=1}^{F}\mathcal{E}\{d_k'^T(\tilde{\theta})d_k'(\tilde{\theta})\}) \end{aligned} \tag{8-73}$$

For small model errors ($\mu \to 0$) and large signal-to-noise ratios ($\upsilon \to 0$) we have

$$V_{\mathrm{ML}}''(\tilde{\theta}) \approx \mathcal{E}\{\sum\nolimits_{k=1}^{F} d_k'^T(\tilde{\theta})d_k'(\tilde{\theta})\} \Rightarrow \Delta V(\tilde{\theta}, Z) \approx \frac{1}{2}\frac{M-1}{M-3}n_\theta \tag{8-74}$$

with n_θ the number of free model parameters (apply quick analysis tool 6 of Section 7.5). Collecting (8-72) and (8-74), using $\mathcal{E}\{c_k(\tilde{\theta})\} = (M-1)/(M-2)$ (see Appendix 8.C), gives

$$\mathcal{E}\{V_{\mathrm{SML}}(\hat{\theta}_{\mathrm{SML}}(Z), Z)\} \approx \frac{M-1}{M-2}\mathcal{E}\{V_{\mathrm{ML}}(\tilde{\theta}, Z)\} - \frac{M-1}{M-3}n_\theta/2 \tag{8-75}$$

Because $\mathcal{E}\{V_{\mathrm{ML}}(\tilde{\theta}, Z)\} = \mathcal{E}\{V_{\mathrm{ML}}(\tilde{\theta}, Z_0)\} + F$, $\tilde{\theta} = \tilde{\theta}_{\mathrm{ML}}(Z_0)$ and

$$\mathcal{E}\{V_{\mathrm{ML}}(\hat{\theta}_{\mathrm{ML}}(Z), Z)\} \approx \mathcal{E}\{V_{\mathrm{ML}}(\hat{\theta}_{\mathrm{ML}}(Z_0), Z_0)\} + F - n_\theta/2 \tag{8-76}$$

(see Theorem 17.12), (8-75) can be written as

$$\mathcal{E}\{V_{\mathrm{SML}}(\hat{\theta}_{\mathrm{SML}}(Z), Z)\} \approx \frac{M-1}{M-2}\mathcal{E}\{V_{\mathrm{ML}}(\hat{\theta}_{\mathrm{ML}}(Z), Z)\} - \frac{M-1}{(M-3)(M-2)}n_\theta/2 \tag{8-77}$$

If no model errors are present ($\tilde{\theta}_{\mathrm{ML}}(Z_0) = \theta_0$), then $\mathcal{E}\{V_{\mathrm{ML}}(\hat{\theta}_{\mathrm{ML}}(Z), Z)\} = F - n_\theta/2$ (see Theorem 17.12).

8.G.2 Calculation of the Variance. For the variance, it is mostly sufficient to have a rough estimate, which can be obtained by neglecting the influence of δ and calculating

$$\mathrm{var}(V_{\mathrm{SML}}(\hat{\theta}_{\mathrm{SML}}(Z), Z)) \approx \mathrm{var}(V_{\mathrm{SML}}(\tilde{\theta}_{\mathrm{SML}}(Z_0), Z)) \tag{8-78}$$

Using $\mathrm{var}(xy) = \sigma_x^2\sigma_y^2 + \sigma_x^2(\mathcal{E}\{y\})^2 + (\mathcal{E}\{x\})^2\sigma_y^2$, we find after some calculations

$$\begin{aligned}\mathrm{var}(V_{\mathrm{SML}}(\hat{\theta}_{\mathrm{SML}}(Z), Z)) \approx{} & \frac{(M-1)^2}{(M-2)(M-3)}\mathrm{var}(V_{\mathrm{ML}}(\tilde{\theta}_{\mathrm{ML}}(Z_0), Z)) \\ & + \frac{(M-1)^2}{(M-2)^2(M-3)}\sum_{k=1}^{F}\mathcal{E}\{|\hat{\varepsilon}(\Omega_k, \tilde{\theta}_{\mathrm{ML}}(Z_0), \hat{Z}(k))|^2\}\end{aligned} \tag{8-79}$$

In the absence of model errors, we have $\tilde{\theta}_{\mathrm{ML}}(Z_0) = \theta_0$, $\mathcal{E}\{|\hat{\varepsilon}(\Omega_k, \theta_0, \hat{Z}(k))|^2\} = 1$ and $\mathrm{var}(V_{\mathrm{ML}}(\theta_0, Z)) = F$, so that (8-79) reduces to (8-22).

Appendix 8.H: Asymptotic Properties of the SGTLS Estimator (Theorem 8.6)

Following the same lines as in the proof of Theorem 7.21 (see Appendix 7.E), it is sufficient to verify that all the conditions of the theorems in Chapter 16 are fulfilled. The cost function (8-24) is of the form

$$V_{\mathrm{SGTLS}}(\theta, Z) = f_F(\theta, w(\theta, Z), Z) \tag{8-80}$$

where $w(\theta, Z) = \frac{1}{F}\sum_{k=1}^{F}\hat{\sigma}_{\hat{e}}^2(\Omega_k, \theta)$. Because $w(\theta, Z)$ is quadratic in the measurements Z and quadratic in the model parameters θ, the theorems of Chapter 15 are valid under the assumptions of Section 7.6: $w^{(k)}(\theta, Z)$, $k = 0, 1, 2$, converge uniformly w.p. 1 (in prob.) to their expected value at the rate $O_p(F^{-1/2})$ in Θ_r; and $w^{(k)}(\theta, Z)$, $k = 0, 1$, converge in law to a Gaussian random variable at the rate $F^{-1/2}$. Hence, $w(\theta, Z)$ satisfies all the assumptions of Chapter 16. The cost function $f_F(\theta, w(\theta), Z)$, with $w(\theta) = \mathcal{E}\{w(\theta, Z)\}$, is quadratic in Z and also satisfies all the assumptions of Chapter 16. We conclude that all the theorems of Chapter 16 apply to the SGTLS cost function (8-24). From Theorems 16.5 and 16.6 follows that $\tilde{\theta}_{\mathrm{SGTLS}}(Z_0)$ and $\theta_{*\mathrm{SGTLS}}$ are the minimizers of $V_F(\theta) = \mathcal{E}\{f_F(\theta, w(\theta), Z)\}$ and $V_*(\theta) = \lim_{F \to \infty}\mathcal{E}\{f_F(\theta, w(\theta), Z)\}$ respectively. Therefore, Theorem 7.21 is valid with three modifications: (i) to calculate $V_F(\theta)$ and $V_*(\theta)$ we first replace $w(\theta, Z)$ by its expected value $w(\theta) = \frac{1}{F}\sum_{k=1}^{F}\sigma_e^2(\Omega_k, \theta)/M$ before taking the expected value of the cost function; (ii) the expected value $\mathcal{E}\{\delta_\theta(Z)\}$ is, in general, not zero as $\tilde{\theta}_{\mathrm{SGTLS}}(Z_0)$ is not the minimizer of $\mathcal{E}\{f_F(\theta, w(\theta, Z), Z)\}$ (see Theorem 16.18); and (iii) $\delta_\theta(Z)$ is replaced by $d_\theta(Z)$ in the expression (7-26) of the covariance matrix

$$\begin{aligned} d_\theta(Z) &= -V_F''^{-1}(\tilde{\theta})d_F(Z) \\ d_F(Z) &= g_F(\tilde{\theta}, w(\tilde{\theta}), w'(\tilde{\theta}), Z) + \left.\frac{\partial g_F(\tilde{\theta}, w(\tilde{\theta}), w'(\tilde{\theta}))}{\partial x}\right|_{x = x_*}(\tilde{x}(Z) - x_*) \end{aligned} \tag{8-81}$$

with $x^T = [w(\tilde{\theta})\ w'(\tilde{\theta})]$, $x_*^T = [w(\tilde{\theta})\ w'(\tilde{\theta})]$, $\tilde{x}^T(Z) = [\tilde{w}(\tilde{\theta}, Z)\ \tilde{w}'(\tilde{\theta}, Z)]$,

$$\begin{aligned} g_F(\theta, w(\theta, Z), w'(\theta, Z), Z) &= f_F'(\theta, w(\theta, Z), Z) \\ g_F(\theta, w, w_1) &= \mathcal{E}\{g_F(\theta, w, w_1, Z)\} \end{aligned} \tag{8-82}$$

and $\tilde{\theta} = \tilde{\theta}_{\mathrm{SGTLS}}(Z_0)$, and where w and w_1 are deterministic variables that replace the random variables $w(\theta, Z)$ and $w'(\theta, Z)$ respectively (see Theorem 16.25).

Appendix 8.I: Relationship between the GTLS and the SGTLS Estimates (Theorem 8.7)

Formulas (8-26) and (8-27) follow directly from (8-25), (7-71), Exercise 7.6 and the fact that the (co)variances of the mean value equal the (co)variances of one realization divided by M.

In the absence of model errors, $\tilde{\theta}(Z_0) = \theta_0$, we have $e(\Omega_k, \theta_0, Z_0(k)) = 0$. Calculating (8-82), where $f_F(\ )$ and $w(\theta, Z)$ are defined in (8-80), gives

$$\begin{aligned} g_F(\theta, w, w_1) = {} & \frac{2M}{Fw}\sum_{k=1}^{F}\mathrm{Re}(e'(\Omega_k, \theta, Z_0(k))\bar{e}(\Omega_k, \theta, Z_0(k))) \\ & -\frac{Mw_1}{Fw^2}\sum_{k=1}^{F}|e(\Omega_k, \theta, Z_0(k))|^2 \end{aligned} \tag{8-83}$$

It follows that $g_F(\theta_0, w, w_1) = 0$ for any w and w_1. Hence, the derivatives of $g_F(\theta_0, w, w_1)$ w.r.t. w and w_1 are zero, so that the second term in the right-hand side of (8-81) is zero. The remaining term in $d_F(Z)$ is the term one would have if the true noise (co)variances were used instead of the sample noise (co)variances. This concludes the proof of (8-28). □

Appendix 8.J: Asymptotic Properties of SBTLS Estimator (Theorem 8.8)

The proof is similar to that of Theorem 8.6 (see Appendix 8.H). The cost function (8-33), with $\theta^{(0)} = \hat{\theta}(Z)$ and $\theta = \theta^{(1)}$, is of the form

$$V_{\mathrm{SBTLS}}(\theta, Z) = f_F(\theta, \eta(Z), w(\theta, \eta(Z), Z), Z) \tag{8-84}$$

where

$$\begin{aligned} \eta^T(Z) &= [\hat{\theta}^T(Z)\ \hat{\alpha}^T(Z)] \\ w(\theta, \eta(Z), Z) &= \frac{1}{F}\sum_{k=1}^{F} \frac{\hat{\sigma}_{\hat{e}}^2(\Omega_k, \theta)}{\sigma_{\hat{e}}^{2r}(\Omega_k, \hat{\theta}(Z), \hat{\alpha}(Z))} \end{aligned} \tag{8-85}$$

satisfy all the conditions of the theorems in Chapter 16. □

Appendix 8.K: Relationship between the BTLS and the SBTLS Estimates (Theorem 8.9)

The proof is similar to that of Theorem 8.7 (see Appendix 8.I). Because $\sigma_n^2(k, \hat{\alpha}(Z))$, $n = 1, 2, 3$, is a consistent estimate of the true noise model $\sigma_n^2(k)$, $n = 1, 2, 3$, we have that $\sigma_e^2(\Omega_k, \theta_*, \alpha_*) = \sigma_e^2(\Omega_k, \theta_*)$, which proves (8-36) and (8-37).

In the absence of model errors, $\tilde{\theta}(Z_0) = \theta_0$, we have $e(\Omega_k, \theta_0, Z_0(k)) = 0$. Calculating

$$\begin{aligned} g_F(\theta, \eta(Z), w(\theta, \eta(Z), Z), w'(\theta, \eta(Z), Z), Z) &= f_F'(\theta, \eta(Z), w(\theta, \eta(Z), Z), Z) \\ g_F(\theta, \eta, w, w_1) &= \mathcal{E}\{g_F(\theta, \eta, w, w_1, Z)\} \end{aligned} \tag{8-86}$$

where $f_F(\ )$, $\eta(Z)$, and $w(\theta, \eta(Z), Z)$ are defined in (8-84) and where η, w, and w_1 are deterministic variables that replace the random variables $\eta(Z)$, $w(\theta, \eta(Z), Z)$, and $w'(\theta, \eta(Z), Z)$, respectively (see Lemma 16.14), gives

$$\begin{aligned} g_F(\theta, \eta, w, w_1) = {} & \frac{2M}{Fw}\sum_{k=1}^{F} \frac{\mathrm{Re}(e'(\Omega_k, \theta, Z_0(k))\bar{e}(\Omega_k, \theta, Z_0(k)))}{\sigma_e^{2r}(\Omega_k, \eta)} \\ & - \frac{Mw_1}{Fw^2}\sum_{k=1}^{F} \frac{|e(\Omega_k, \theta, Z_0(k))|^2}{\sigma_e^{2r}(\Omega_k, \eta)} \end{aligned} \tag{8-87}$$

It follows that $g_F(\theta_0, \eta, w, w_1) = 0$ for any η, w, and w_1. Hence, the derivatives of $g_F(\theta_0, \eta, w, w_1)$ w.r.t. η, w, and w_1 are zero, so that the second term in $d_F(Z)$ (16-17) is zero. Because $\sigma_e^2(\Omega_k, \theta_*, \alpha_*) = \sigma_e^2(\Omega_k, \theta_*)$, the remaining term in $d_F(Z)$ is the term one would have if the true noise (co)variances were used instead of the sample noise (co)variances. This concludes the proof of (8-38).

Appendix 8.L: Asymptotic Properties of SSUB Algorithms (Theorem 8.10)

Because the second step of the subspace Algorithms 7.24 and 7.25 does not require any noise information, it is sufficient to analyze the first and the third steps only.

8.L.1 Step 1: Estimation of the Extended Observability Matrix. The results of Appendix 7.Q, estimation of the extended observability matrix O_r, remain valid if we can show that $\hat{C}_Y/F$ converges w.p. 1 to its expected value C_Y/F and if the convergence rate is an $O_p(F^{-1/2})$. From this result it follows directly that the limit value O_{r*}, calculated with the sample variance $\hat{\sigma}_Y^2(k)$, equals that calculated with the true variance $\sigma_Y^2(k)$.

Under the assumptions of Section 7.6.1, $|N_Y(k)|^2$ is independently distributed for a frequency domain experiment and converges for $F \to \infty$ to a random variable that is mixing of order 2 for time domain experiment. This is also valid for

$$\hat{\sigma}_Y^2(k) = \frac{1}{M-1}\sum_{l=1}^{M}|Y^{[l]}(k) - \hat{Y}(k)|^2 = \frac{1}{M-1}\sum_{l=1}^{M}\left|N_Y^{[l]}(k) - \frac{1}{M}\sum_{m=1}^{M} N_Y^{[m]}(k)\right|^2 \quad (8\text{-}88)$$

and, therefore, $\hat{C}_{\mathrm{Y}}/F$

$$\hat{C}_{\mathrm{Y}}/F = \mathrm{Re}\left(\frac{1}{F}\sum_{k=1}^{F}(\hat{\sigma}_Y^2(k)/M)W_r(k)W_r^H(k)\right) \quad (8\text{-}89)$$

converges w.p. 1 to its expected value at the rate $O_{\mathrm{p}}(F^{-1/2})$ (see Section 14.9, versions 2 and 3 of the law of large numbers). Because $\mathcal{E}\{\hat{\sigma}_Y^2(k)\} = \sigma_Y^2(k)$, the expected value of $\hat{C}_{\mathrm{Y}}$ equals C_{Y}.

The proof for $\hat{C}_{\mathrm{Y}_\perp}$ is somewhat more complicated because the scalar orthogonal basis $p_n(s)$, $n = 0, 1, \ldots, r-1$ depends on the measurements $\hat{Y}(k)$, $k = 1, 2, \ldots, F$. Indeed, the orthogonal basis is calculated using the inner product (see Appendix 7.S)

$$\langle x(s), y(s)\rangle_Y = \mathrm{Re}\left(\sum_{k=1}^{F}|\hat{Y}(k)|^2 x(s_k)\bar{y}(s_k)\right) \quad (8\text{-}90)$$

which clearly depends on $\hat{Y}(k)$. Under the assumptions of Section 7.6.1, the inner product $\langle x(s), y(s)\rangle_Y/F$ converges w.p. 1 at the rate $O_{\mathrm{p}}(F^{-1/2})$ to its expected value

$$\mathcal{E}\{\langle x(s), y(s)\rangle_Y/F\} = \mathrm{Re}\left(\frac{1}{F}\sum_{k=1}^{F}(|Y_0(k)|^2 + \sigma_Y^2(k)/M)x(s_k)\bar{y}(s_k)\right) \quad (8\text{-}91)$$

Hence, the polynomials $p_n(s)$ also converge w.p. 1 at the rate $O_{\mathrm{p}}(F^{-1/2})$ to its limit value $\tilde{p}_n(s)$ (see recursion formulas (7-249) and (7-250)). We conclude that $W_Y(k)$ converges w.p. 1 at the rate $O_{\mathrm{p}}(F^{-1/2})$ to $\tilde{W}_Y(k)$. Applying Corollary 15.35 to

$$\hat{C}_{\mathrm{Y}_\perp}/F = \mathrm{Re}\left(\frac{1}{F}\sum_{k=1}^{F}(\hat{\sigma}_Y^2(k)/M)W_Y(k)W_Y^H(k)\right) \quad (8\text{-}92)$$

shows that $\hat{C}_{\mathrm{Y}_\perp}/F$ converges w.p. 1 at the rate $O_{\mathrm{p}}(F^{-1/2})$ to

$$\mathrm{Re}\left(\frac{1}{F}\sum_{k=1}^{F}(\sigma_Y^2(k)/M)\tilde{W}_Y(k)\tilde{W}_Y^H(k)\right) \quad (8\text{-}93)$$

This concludes the proof because (8-93) is the asymptotic expression of the covariance matrix when the noise variance is known.

8.L.2 Step 3: Estimation of B and D. From step 1 it follows that $\hat{A}$ and $\hat{C}$ converge w.p. 1 at the rate $O_{\mathrm{p}}(F^{-1/2})$ to their limit values A_0, C_0 or, in case of model errors, to A_*, C_* (see Appendix 7.R). The cost function

$$V_{\text{SSUB}}(B, D, A_1, C_1, Z) = \sum_{k=1}^{F} \frac{\left|\hat{Y}(k) - (C_1(\xi_k I_{n_a} - A_1)^{-1}B + D)\hat{U}(k)\right|^2}{\hat{\sigma}_Y^2(k)/M} \tag{8-94}$$

where A_1 and C_1 are deterministic and where, by assumption, the input is known ($\hat{U}(k) = U_0(k)$), has exactly the same stochastic structure as the SML cost function (8-10). Therefore, under the same assumptions $V_{\text{SSUB}}(B, D, A_1, C_1, Z)$ has the same asymptotic properties as $V_{\text{SML}}(\theta, Z)$. The only difference is that the denominator in (8-94) is independent of θ, which decreases the value of M from 6 to 4 for the convergence rate (see Appendix 8.D). Hence, for $M \geq 4$, (8-94) converges w.p. 1 at the rate $O_p(F^{-1/2})$ to its expected value (see proof of Theorem 8.3). We conclude that $V_{\text{SSUB}}(B, D, \hat{A}, \hat{C}, Z)$ satisfies all the assumptions of Theorems 16.5, 16.7 and 16.16. Hence, for $M \geq 4$, $\hat{B}$ and $\hat{D}$ converge w.p. 1 to their limit values B_0, D_0 (Theorem 16.7) or, in case of model errors, to B_*, D_* (Theorem 16.5), and the convergence rate is an $O_p(F^{-1/2})$ (Theorem 16.16). Note that the limit values A_*, C_*, B_*, and D_*, calculated with the sample variance $\hat{\sigma}_Y^2(k)$, equal those calculated with the true variance $\sigma_Y^2(k)$ since this is also the case for O_{r*}.

Appendix 8.M: Related Linear Dynamic System of a Cascade of Nonlinear Systems

Applying (8-44) to the nonlinear operators $T[.]$ and $A[.]$ gives, taking into account the measurement errors $M_U(k)$, $M_Y(k)$ and the process noise $N_P(k)$,

$$\begin{aligned} Y(k) &= T_R(s_k)R(k) + N_Y(k) \\ U(k) &= A_R(s_k)R(k) + N_U(k) \end{aligned} \tag{8-95}$$

with $N_Y(k) = N_P(k) + Y_S(k) + M_Y(k)$, $N_U(k) = U_S(k) + M_U(k)$, and $Y_S(k)$, $U_S(k)$ the zero mean nonlinear distortions, which are independent of $R(k)$. Dividing both sides of (8-95) by $R(k)$ and taking the expected value w.r.t. to the measurement noise and the random phase φ_k of $R(k)$ shows that $E\{Y_R(k)\}/E\{U_R(k)\} = T_R(s_k)/A_R(s_k)$, where $Y_R(k)$, $U_R(k)$ are defined as in (8-45).

Appendix 8.N: Asymptotic Properties of the Prediction Error Estimate (Theorem 8.15)

The prediction error cost function is quadratic-in-the-measurements Z so that properties 1 to 5 of Theorem 7.21 are valid, even if contrary to Assumption 7.3 the excitation $u(t)$ and the output error $n_y(t)$ are not independent (the proof follows the lines of the time domain experiment in Theorem 7.21). To prove properties 6 to 8 of Theorem 7.21 (consistency, bias, and efficiency) we need only verify that the estimates are consistent under the assumptions of Section 7.6.6 and Assumptions 8.12 to 8.14 (open loop identification) or Assumptions 8.14 and 8.16 to 8.18 (closed loop identification). To prove the consistency, it is sufficient to show that the true parameters θ_0 minimize the expected value of the cost function $V_F(\theta)$, $\tilde{\theta}(Z_0) = \theta_0$, or its limit value $V_*(\theta)$, $\theta_* = \theta_0$.

Under Assumption 7.16, and Assumptions 8.12 and 8.13 (open loop) or Assumptions 8.16 and 8.18 (closed loop), the observed output $Y(k)$ can be written as

$$Y(k) = G(s_k, \theta_0)U(k) + T(s_k, \theta_0) + H(z_k^{-1}, \theta_0)E(k) + T_H(z_k^{-1}, \theta_0) + \delta(s_k) \tag{8-96}$$

$$Y(k) = G(z_k^{-1}, \theta_0)U(k) + T(z_k^{-1}, \theta_0) + H(z_k^{-1}, \theta_0)E(k) + T_H(z_k^{-1}, \theta_0) \tag{8-97}$$

where $H(z_k^{-1}, \theta_0)E(k)$ models the error $N_Y(k) = N_P(k) + M_Y(k)$ (open loop) or $N_Y(k) = N_P(k)$ (closed loop). These expressions converge $(F \to \infty)$ uniformly in prob. for $\Omega = s$ and w.p. 1 for $\Omega = z^{-1}$ to

$$Y(k) = G(\Omega_k, \theta_0)U(k) + H(z_k^{-1}, \theta_0)E(k) \tag{8-98}$$

at the rate $O_p(F^{-1/2})$ (see Lemmas 5.7 and 14.23). It shows that $V_{\text{PE}}(\theta, Z)/F$ converges uniformly in prob. for $\Omega = s$ and w.p. 1 for $\Omega = z^{-1}$ to

$$V_{1F}(\theta, Z) = \frac{1}{F}\sum_{k \in \mathbb{F}} \left| \frac{\Delta G(\Omega_k, \theta)U(k) + H(z_k^{-1}, \theta_0)E(k)}{H(z_k^{-1}, \theta)} \right|^2 \tag{8-99}$$

with $\Delta G(\Omega_k, \theta) = G(\Omega_k, \theta_0) - G(\Omega_k, \theta)$. Adding and subtracting $H(z_k^{-1}, \theta)E(k)$ in the numerator of (8-99) gives

$$\begin{aligned} V_{1F}(\theta, Z) = {} & \frac{1}{F}\sum_{k \in \mathbb{F}} \left| \frac{\Delta G(\Omega_k, \theta)U(k) + \Delta H(z_k^{-1}, \theta)E(k)}{H(z_k^{-1}, \theta)} \right|^2 + \frac{1}{F}\sum_{k \in \mathbb{F}} |E(k)|^2 \\ & + \frac{2}{F}\operatorname{Re}\left(\sum_{k \in \mathbb{F}} \frac{\Delta G(\Omega_k, \theta)}{H(z_k^{-1}, \theta)} U(k)\bar{E}(k) \right) + \frac{2}{F}\operatorname{Re}\left(\sum_{k \in \mathbb{F}} \frac{\Delta H(z_k^{-1}, \theta)}{H(z_k^{-1}, \theta)} |E(k)|^2 \right) \end{aligned} \tag{8-100}$$

with $\Delta H(z_k^{-1}, \theta) = H(z_k^{-1}, \theta_0) - H(z_k^{-1}, \theta)$. Under Assumption 7.3 and Assumption 8.12 (open loop) or Assumption 8.16 (closed loop), $E(k)$ has zero mean and variance σ^2. Hence, the expected value $V_{1F}(\theta) = \mathcal{E}\{V_{1F}(\theta, Z)\}$ of (8-100) becomes

$$\begin{aligned} V_{1F}(\theta) = {} & \frac{1}{F}\sum_{k \in \mathbb{F}} \begin{bmatrix} \Delta G(\Omega_k, \theta) \\ \Delta H(z_k^{-1}, \theta) \end{bmatrix}^H \Phi(k) \begin{bmatrix} \Delta G(\Omega_k, \theta) \\ \Delta H(z_k^{-1}, \theta) \end{bmatrix} + \sigma^2 \\ & + \frac{2}{F}\operatorname{Re}\left(\sum_{k \in \mathbb{F}} \frac{\Delta G(\Omega_k, \theta)}{H(z_k^{-1}, \theta)} \mathcal{E}\{U(k)\bar{E}(k)\} \right) + \frac{2\sigma^2}{F}\operatorname{Re}\left(\sum_{k \in \mathbb{F}} \frac{\Delta H(z_k^{-1}, \theta)}{H(z_k^{-1}, \theta)} \right) \end{aligned} \tag{8-101}$$

with

$$\Phi(k) = \frac{1}{|H(z_k^{-1}, \theta)|^2} \begin{bmatrix} \mathcal{E}\{|U(k)|^2\} & \mathcal{E}\{\bar{U}(k)E(k)\} \\ \mathcal{E}\{U(k)\bar{E}(k)\} & \sigma^2 \end{bmatrix} \tag{8-102}$$

a positive (semi-)definite matrix. We first analyze (8-101) for the two following cases: $u(t)$, $n_y(t)$ are independent (open loop identification), and $u(t)$, $n_y(t)$ are linearly dependent (linear closed loop identification). Next, the influence of input measurement errors on the estimates is studied. Finally, we study (8-101) in case of plant model errors.

8.N.1 Open Loop Identification. In this section we assume that $u(t)$, $n_y(t)$ are independent (Assumption 7.3 is valid), so that $\mathcal{E}\{U(k)\bar{E}(k)\} = 0$. Hence, (8-101) becomes

$$\begin{aligned} V_{1F}(\theta) &= \frac{1}{F}\sum_{k \in \mathbb{F}}\left|\frac{\Delta G(\Omega_k, \theta)}{H(z_k^{-1}, \theta)}\right|^2 \mathcal{E}\{|U(k)|^2\} + \frac{\sigma^2}{F}\sum_{k \in \mathbb{F}}\left|\frac{\Delta H(z_k^{-1}, \theta)}{H(z_k^{-1}, \theta)}\right|^2 \\ &+ \sigma^2 + \frac{2\sigma^2}{F}\mathrm{Re}\left(\sum_{k \in F}\frac{\Delta H(z_k^{-1}, \theta)}{H(z_k^{-1}, \theta)}\right) \end{aligned} \tag{8-103}$$

From (8-103) it follows directly that $V_{1F}(\theta)$ is minimal in the true plant model parameters $a_{0,}$ b_0 if the plant and noise models are independently parameterized. Note that this is true for an arbitrary subset $\mathbb{F}$ of the DFT frequencies, even if a true noise model does not exist. It proves the weak $(\Omega = s)$ and strong $(\Omega = z^{-1})$ consistency of the estimated plant model parameters (extension 1 of Theorem 8.15).

$H(z^{-1}, \theta_0)$ and $H(z^{-1}, \theta)$ are stable and inversely stable monic filters so that

$$\frac{\Delta H(z^{-1}, \theta)}{H(z^{-1}, \theta)} = F(z^{-1}) = \sum\nolimits_{r=0}^{\infty} f_r z^{-r} \text{ for any } |z| \geq 1 \text{ with } f_0 = 0 \tag{8-104}$$

Note that $f_0 = 0$ is a direct consequence of the constraint $c_0 = d_0 = 1$, and is the key property to show the consistency. It is easy to verify that under Assumption 8.14

$$\sum\nolimits_{k \in \mathbb{F}} z_k^{-r} = 0 \qquad r = 1, 2, \ldots \tag{8-105}$$

Take, for example, $\mathbb{F} = \{0, 1, \ldots, N-1\}$ or $\mathbb{F} = \{2k+1 | k = 0, 1, \ldots, N/2-1\}$. Using (8-105) we get

$$\sum_{k \in \mathbb{F}}\frac{\Delta H(z_k^{-1}, \theta)}{H(z_k^{-1}, \theta)} = \sum\nolimits_{r=0}^{\infty} f_r \sum_{k \in \mathbb{F}} z_k^{-r} = F f_0 = 0 \tag{8-106}$$

because $f_0 = 0$. Collecting (8-103) and (8-106) gives

$$V_{1F}(\theta) = \frac{1}{F}\sum_{k \in \mathbb{F}}\left|\frac{\Delta G(\Omega_k, \theta)}{H(z_k^{-1}, \theta)}\right|^2 \mathcal{E}\{|U(k)|^2\} + \frac{\sigma^2}{F}\sum_{k \in \mathbb{F}}\left|\frac{\Delta H(z_k^{-1}, \theta)}{H(z_k^{-1}, \theta)}\right|^2 + \sigma^2 \tag{8-107}$$

It follows that (8-107) is minimal in the true plant and noise model parameters θ_0, which shows the weak $(\Omega = s)$ and strong $(\Omega = z^{-1})$ consistency of the estimate $\hat{\theta}_{\mathrm{PE}}(Z)$ in the general case that the plant and noise model parameters may have common parameters.

8.N.2 Closed Loop Identification. Consider the linear feedback experiment of Figure 7-14 on page 243 for arbitrary reference signals $r(t)$, where $m_u(t) = 0$ and $m_y(t) = 0$ so that $n_u(t) = 0$ and $n_y(t) = n_p(t)$. Contrary to Assumption 7.3, the true excitation $u(t) = u_1(t)$ and the process noise $n_p(t)$ are linearly dependent. Expressions (8-99) to (8-102) remain valid with $N_p(k) = H(z_k^{-1}, \theta_0)E(k)$ the true noise model. $U(k)$ and $N_P(k)$ are related by

$$U(k) = \frac{R(k) - (N_C(k) + C_0(\Omega_k)N_P(k))}{1 + G(\Omega_k, \theta_0)C_0(\Omega_k)} \tag{8-108}$$

see (7-143) and (7-144) with $M_U(k) = 0$ and $M_Y(k) = 0$, so that

$$\mathcal{E}\{U(k)\bar{E}(k)\} = -\frac{C_0(\Omega_k)H(z_k^{-1}, \theta_0)}{1 + G(\Omega_k, \theta_0)C_0(\Omega_k)}\sigma^2 \tag{8-109}$$

Using (8-109), the expected value (8-101) can be written as

$$\begin{aligned} V_{1F}(\theta) = {} & \frac{1}{F}\sum_{k\in\mathbb{F}}\begin{bmatrix}\Delta G(\Omega_k, \theta) \\ \Delta H(z_k^{-1}, \theta)\end{bmatrix}^H \Phi(k)\begin{bmatrix}\Delta G(\Omega_k, \theta) \\ \Delta H(z_k^{-1}, \theta)\end{bmatrix} + \sigma^2 \\ & + \frac{2\sigma^2}{F}\mathrm{Re}\left(\sum_{k\in\mathbb{F}} F(\Omega_k, z_k^{-1}, \theta)\right) \end{aligned} \tag{8-110}$$

with

$$F(\Omega, z^{-1}, \theta) = \frac{H(z^{-1}, \theta_0)}{1 + G(\Omega, \theta_0)C_0(\Omega)}\frac{1 + G(\Omega, \theta)C_0(\Omega)}{H(z^{-1}, \theta)} - 1 \tag{8-111}$$

For continuous-time models $\Omega = s$ we have in general $\sum_{k\in\mathbb{F}} F(\Omega_k, z_k^{-1}, \theta) \neq 0$, so that the true model parameters θ_0 do not minimize (8-110). For discrete-time models $\Omega = z^{-1}$ we have $\lim_{z\to\infty} F(z^{-1}, z^{-1}, \theta) = 0$ because $H(z^{-1}, \theta_0)$ and $H(z^{-1}, \theta)$ are monic and either $\lim_{z\to\infty} G(z^{-1}, \theta) = 0$ or $\lim_{z\to\infty} C_0(z^{-1}) = 0$. If $F(z^{-1}, z^{-1}, \theta)$ is stable then

$$F(z^{-1}, z^{-1}, \theta) = \sum_{r=0}^{\infty} f_r z^{-r} \text{ for any } |z| \geq 1 \text{ with } f_0 = 0 \tag{8-112}$$

and similarly to (8-106),

$$\sum_{k\in\mathbb{F}} F(z_k^{-1}, z_k^{-1}, \theta) = \sum_{r=0}^{\infty} f_r \sum_{k\in\mathbb{F}} z_k^{-r} = F f_0 = 0 \tag{8-113}$$

Collecting (8-110) and (8-113) gives

$$V_{1F}(\theta) = \frac{1}{F}\sum_{k\in\mathbb{F}}\begin{bmatrix}\Delta G(\Omega_k, \theta) \\ \Delta H(z_k^{-1}, \theta)\end{bmatrix}^H \Phi(k)\begin{bmatrix}\Delta G(\Omega_k, \theta) \\ \Delta H(z_k^{-1}, \theta)\end{bmatrix} + \sigma^2 \tag{8-114}$$

The quadratic form in (8-114) is minimal in the true plant and noise model parameters θ_0 if the matrix $\Phi(k)$ (8-102) is positive definite for almost every k. Using (8-109), it is easy to verify that the condition $\det(\Phi(k)) > 0$ is satisfied for almost every k if and only if $R(k) \neq N_C(k)$ for almost every k. We conclude that $\hat{\theta}_{\mathrm{PE}}(Z)$ is strongly consistent provided that $F(z^{-1}, z^{-1}, \theta)$ is stable. We will now verify under which conditions $F(z^{-1}, z^{-1}, \theta)$ is stable. We distinguish between two cases: the plant is stable or the plant is unstable.

If the plant $G(z^{-1}, \theta_0)$ is stable, then the true noise model $H(z^{-1}, \theta_0)$ should be stable, and there exists a closed and bounded neighborhood Θ_r of θ_0 such that $G(z^{-1}, \theta)$ and $H(z^{-1}, \theta)$ are stable for any $\theta \in \Theta_r$. By assumption, the controller $C_0(z^{-1})$ stabilizes the closed loop system so that $1/(1 + G_0C_0)$ is stable. Because the noise filters $H(z^{-1}, \theta_0)$ and

$H(z^{-1}, \theta)$ are stable and inversely stable, it follows that $F(z^{-1}, z^{-1}, \theta)$ (8-111) is stable. We conclude that the prediction error estimate $\hat{\theta}_{\text{PE}}(Z)$ of any model structure of Section 5.7.3 is consistent for stable plants.

If the plant $G(z^{-1}, \theta_0)$ is unstable, then $G(z^{-1}, \theta)$ is also unstable in a closed and bounded neighborhood Θ_r of θ_0. The controller $C_0(z^{-1})$ stabilizes the closed loop system so that $1/(1 + G_0C_0)$ is still stable. There are two possibilities now: the true noise model H_0 is unstable or stable.

1. If H_0 is unstable, then it should have the same unstable poles as the plant G_0, otherwise $H_0/(1 + G_0C_0)$ in (8-111) would be unstable. The factor $(1 + GC_0)/H$ in (8-111) is stable in Θ_r only if H and G also have exactly the same unstable poles. This condition can never be met for a Box-Jenkins (BJ) model structure because the plant $G(z^{-1}, \theta)$ and noise $H(z^{-1}, \theta)$ models are independently parameterized (see Section 5.7.3). For ARX and ARMAX model structures, the plant $G(z^{-1}, \theta)$ and the noise $H(z^{-1}, \theta)$ model have the same denominator (see Section 5.7.3), so that $(1 + GC_0)/H$ and $F(z^{-1}, z^{-1}, \theta)$ (8-111) are stable. We conclude that (8-112) is valid for unstable plants and ARX and ARMAX models.
2. If H_0 is stable, then $H(z^{-1}, \theta)$ is stable in a closed and bounded neighborhood Θ_r of θ_0. Hence, the factor $H_0/(1 + G_0C_0)$ in (8-111) is stable, while $(1 + GC_0)/H$ is unstable in Θ_r. We conclude that (8-112) is no longer valid for stable noise models and unstable plants.

We conclude that the prediction error estimate $\hat{\theta}_{\text{PE}}(Z)$ of ARX and ARMAX models is consistent for unstable plants, but that of BJ and OE models is inconsistent.

If Assumption 8.18 is not satisfied, then (8-114) is minimal for $\Delta H \neq 0$. Because the matrix $\Phi(k)$ is not diagonal, it also follows directly that $\Delta G \neq 0$. If Assumption 8.14 is not satisfied, then, (8-113) is no longer valid, and (8-114) is not minimal in θ_0.

If $R(k) = N_C(k)$ for every k, then, $\Phi(k)$ is singular (positive semidefinite) for very k, and we may no longer conclude from (8-114) that θ_0 minimizes the expected value of the cost function $V_{1F}(\theta)$. Indeed, putting the condition $R(k) = N_C(k)$ for any k in (8-99) gives, using (8-108),

$$V_{1F}(\theta, Z) = \frac{1}{F}\sum_{k \in \mathbb{F}} \left|\frac{1 + G(z_k^{-1}, \theta)C_0(z_k^{-1})}{1 + G(z_k^{-1}, \theta_0)C_0(z_k^{-1})}\right|^2 \left|\frac{H(z_k^{-1}, \theta_0)}{H(z_k^{-1}, \theta)}\right|^2 |E(k)|^2 \tag{8-115}$$

Clearly, the expected value of (8-115) is minimal for $G(z^{-1}, \theta) = -1/C_0(z^{-1})$. We conclude that $\hat{\theta}_{\text{PE}}(Z)$ is no longer consistent if $R(k) = N_C(k)$ for every k.

If the output measurement errors are not zero $m_y(t) \neq 0$, then, $H(z_k^{-1}, \theta_0)E(k)$ models $N_Y(k) = N_P(k) + M_Y(k)$ where $N_P(k)$ and $M_Y(k)$ are, by assumption, independent random variables. Using $N_P(k) = H_1(z_k^{-1})E_1(k)$, the expected value $\mathcal{E}\{U(k)\bar{E}(k)\}$ can be written as

$$\mathcal{E}\{U(k)\bar{E}(k)\} = -\frac{C_0(z_k^{-1})|H_1(z_k^{-1})|^2 \quad \sigma_1^2}{1 + G(z_k^{-1}, \theta_0)C_0(z_k^{-1})\bar{H}(z_k^{-1}, \theta_0)} \tag{8-116}$$

Due to the factor $|H_1(z_k^{-1})|^2/\bar{H}(z_k^{-1}, \theta_0)$, the sum of the third and the fourth term in (8-101) cannot be zero. We conclude that $\hat{\theta}_{\text{PE}}(Z)$ is no longer consistent if $m_y(t) \neq 0$.

8.N.3 Influence of Input Measurement Errors. If the input is observed with measurement errors $m_u(t)$, then (8-99) becomes

$$V_{1F}(\theta, Z) = \frac{1}{F}\sum_{k \in \mathbb{F}} \left| \frac{\Delta G(\Omega_k, \theta) U_1(k) + H(z_k^{-1}, \theta_0) E(k) + G(\Omega_k, \theta) N_U(k)}{H(z_k^{-1}, \theta)} \right|^2 \tag{8-117}$$

where $U_1(k)$ is the true input of the plant (see Figure 7-14 on page 243). Assume now for simplicity of notation and without loss of generality that the measurement errors $m_u(t)$ are independent of the process noise $n_p(t)$ and the output measurement error $m_y(t)$. The expected value of (8-117) then equals (8-101) plus

$$\frac{1}{F}\sum_{k \in \mathbb{F}} \left| \frac{G(\Omega_k, \theta)}{H(z_k^{-1}, \theta)} \right|^2 \sigma_U^2(k) \tag{8-118}$$

Because (8-118) is, in general, not minimal in θ_0, it is also valid for the expected value of (8-117), and, hence, the estimates are inconsistent.

8.N.4 Plant Model Errors. Using the notations of Assumption 7.14, the expected value of (8-101) can be written as

$$\begin{aligned} V_*(\theta) = {} & \int_{f_{\min}}^{f_{\max}} \begin{bmatrix} \Delta G(\Omega(f), \theta) \\ \Delta H(e^{-j2\pi f T_s}, \theta) \end{bmatrix}^H \Phi(f) \begin{bmatrix} \Delta G(\Omega(f), \theta) \\ \Delta H(e^{-j2\pi f T_s}, \theta) \end{bmatrix} n(f) df + \sigma^2 \\ & + 2\mathrm{Re}\left(\int_{f_{\min}}^{f_{\max}} \frac{\Delta G(\Omega(f), \theta)}{H(e^{-j2\pi f T_s}, \theta)} \Phi_{UE}(f) n(f) df \right) \\ & + 2\sigma^2 \mathrm{Re}\left(\int_{f_{\min}}^{f_{\max}} \frac{\Delta H(e^{-j2\pi f T_s}, \theta)}{H(e^{-j2\pi f T_s}, \theta)} n(f) df \right) \end{aligned} \tag{8-119}$$

where

$$\Phi(f) = \frac{1}{|H(e^{-j2\pi f T_s}, \theta)|^2} \begin{bmatrix} \Phi_{UU}(f) & \bar{\Phi}_{UE}(f) \\ \Phi_{UE}(f) & \sigma^2 \end{bmatrix} \tag{8-120}$$

with $\Phi_{UU}(f) = \mathcal{E}\{|U(f)|^2\}$ and $\Phi_{UE}(f) = \mathcal{E}\{U(f)\bar{E}(f)\}$. Note that the first term in (8-119) depends on the noise level σ through $\Phi(f)$ (8-120). Therefore, in case of plant model errors and if a parametric noise model is identified $(H(z^{-1}, \theta) \neq 1)$, the minimizing arguments $a_{*\mathrm{PE}}$ and $b_{*\mathrm{PE}}$ of $V_*(\theta)$ (8-119) depend on σ, even in open loop $\Phi_{UE}(f) = 0$, and even if the last two terms in (8-119) are zero. It is easy to verify that $a_{*\mathrm{PE}}$ and $b_{*\mathrm{PE}}$ are independent of the noise level for OE model structures, $H(z^{-1}, \theta) = 1$, identified in

1. Open loop, $\Phi_{UE}(f) = 0$, even if Assumptions 8.13 and 8.14 are not satisfied.
2. Closed loop, $\Phi_{UE}(f) \neq 0$, if Assumption 8.14 is satisfied (the third term in (8-119) is then zero) and if the true noise model is white, $H(z^{-1}, \theta_0) = 1$ (ΔH in (8-119) is then zero).

8.N.5 Closed Loop Identification with an Enhanced Noise Model. Unstable plants identified using OE and BJ models lead to inconsistent estimates in the frequency domain (extension 2b of Theorem 8.19) and not computable estimates in the time domain (Forssell and Ljung, 2000a). To cope with this problem (guarantee the stability of the prediction error $\varepsilon(t, \theta)$), Forssell and Ljung (2000a) propose to multiply the noise model $H(z^{-1}, \theta)$ in the OE and BJ model structures by the all-pass filter $H_a(z^{-1}, \theta)$ (8-51), giving the enhanced noise model $H_1(z^{-1}, \theta)$

$$H_1(z^{-1}, \theta) = H(z^{-1}, \theta)H_a(z^{-1}, \theta) \tag{8-121}$$

Replacing $H(z^{-1}, \theta)$ by $H_1(z^{-1}, \theta)$ in (8-48) guarantees the stability of the prediction error $\varepsilon(z_k^{-1}, \theta)$: the unstable poles of $G(z^{-1}, \theta)$ are canceled by the nonminimum phase zeros of $F_a(z^{-1}, \theta)$, and $H_1^{-1}(z^{-1}, \theta)$ is stable (the zeros of $F_a^*(z^{-1}, \theta)$ are the zeros of $F_a(z^{-1}, \theta)$ reflected into the unit circle). In the sequel we show that the prediction error estimate with the enhanced OE or BJ noise model (8-121) is strongly consistent.

The term $H(z_k^{-1}, \theta_0)E(k)$ in (8-97) can be written as $H_1(z_k^{-1}, \theta_0)E_1(k)$, with $E_1(k) = H_a^{-1}(z_k^{-1}, \theta_0)E(k)$ the DFT spectrum of a stable white noise sequence, and where $H_1(z_k^{-1}, \theta_0)$ is defined in (8-121). Because $|H_a(z_k^{-1}, \theta)| = 1$, $E_1(k)$ has the following properties:

$$|E_1(k)| = |E(k)| \quad \text{and} \quad \mathcal{E}\{E_1(k)\bar{E}_1(l)\} = \sigma^2\delta(k-l) \tag{8-122}$$

Using the enhanced noise models $H_1(z^{-1}, \theta_0)$ and $H_1(z^{-1}, \theta)$, and properties (8-122), it follows immediately that the expression (8-110), where H is replaced by H_1, is still valid. Using the definitions of H_1 (8-121) and H_a (8-51), it is easy to see that $F(z^{-1}, z^{-1}, \theta)$ is stable for the enhanced OE and BJ model structures. Hence, the third term in (8-110) is zero (proof: follow the lines of Section 8.N.2 of this appendix), and $V_{1F}(\theta)$ is minimal in the true model parameters: $\Delta G(z_k^{-1}, \theta_0) = 0$ and $\Delta H_1(z_k^{-1}, \theta_0) = 0$. □

9

Model Selection and Validation

Abstract: A critical step in the identification process is the quality assessment of the identified model. A model without error bounds has no value. For this reason, we need tools to check whether all linear dynamics in the raw data are captured and tools to quantify the remaining model errors. Also, the presence of nonlinear distortions should be detected, qualified, and quantified. Finally, the validity of the disturbing noise models should be tested.

This chapter provides dedicated tools to test for over- and undermodeling. This information is used not only to validate the final model but also to guide the model selection process during the identification. The methods vary from a simple visual inspection (does the transfer function fit the FRF measurements well enough for the intended application?) to an advanced statistical analysis of the residuals. In the case of undermodeling, the remaining model error is quantified so that the user can decide whether the final model is acceptable for his or her application.

9.1 INTRODUCTION

At the end of an identification run, two important questions remain to be answered. What is the quality of the model? Can this model be used to solve my problem? Whereas the first question is an absolute one, the second question shows that in practice the applicability of an identified model strongly depends on the intended application. Each model is only an approximation of reality, and often the existence of a "true" model is only a fiction, in the mind of the experimenter. The deviations between the model and the system that generated the measurements are partitioned in two parts following their nature: systematic errors and stochastic errors. If the experiment is repeated under the same conditions, the systematic errors will be the same, but the stochastic errors vary from one realization to the other. Model validation is directed toward the quantification of the remaining model errors. Once the level of the systematic errors is known, the user should decide whether or not they are acceptable. It is not evident at all that one is looking for the lowest error level; often it is sufficient to push them below a given upper bound. In order to decide whether the errors are systematic, it is necessary to know the uncertainty on the estimated model. In this book we use probabilistic uncertainty bounds (e.g., 95% bounds) that describe how the individual realizations are scattered

System Identification: A Frequency Domain Approach. Rik Pintelon and Johan Schoukens
Copyright © 2001 Institute of Electrical and Electronics Engineers. ISBN: 0-780-36000-1

around their mean values. Errors that are outside this bound are considered to be unlikely, so that they are most probably due to systematic deviations.

This short discussion shows, clearly, that model validation starts with the generation of good uncertainty bounds. These bounds can be used in a second step to check for the presence of significant (from a statistical point of view) systematic errors. This two-step approach is developed in the course of this chapter. First, it is shown how error bounds on the transfer function and the poles can be generated starting from the covariance matrix of the parameters C_θ (generated by the estimation algorithm). Next, it will be explained how the presence of systematic errors can be detected. This information will be used to develop an automatized model selection procedure.

Note that using the measured frequency response function (FRF) in the validation step is theoretically not the best choice in an errors-in-variables concept (noise on the input and output measurements). However, as shown in Chapter 2, the FRF measurements are very usable for sufficient high signal-to-noise ratios (SNRs). Although we still prefer to do the identification from the measured input-output data (no user decision needed to check whether the measurements are good enough to jump to the FRF; no increase in complexity for the user; nonincreased computation time), using the FRF simplifies the validation process significantly because it is much easier to interpret, allows a simple visual inspection, and is often one of the major results that is needed.

9.2 ASSESSING THE MODEL QUALITY: QUANTIFYING THE STOCHASTIC ERRORS

As mentioned in the introduction, the first step in the validation process is the partitioning into stochastic and systematic errors. The stochastic error bounds are not only a tool to detect systematic errors, they are also intensively used to describe the overall quality of the model once it is known that systematic errors are no longer dominating. The basic "uncertainty" information is delivered under the form of the covariance matrix on the estimated parameters. The actual covariance matrix is mostly too difficult to calculate. But in most cases the Cramér-Rao lower bound (see Sections 1.3.2 and 14.12) can be used for asymptotically efficient estimators. Also, for weighted least squares estimators, approximative expressions to calculate the covariance matrix are available (see, for example, Theorem 7.21). An approximation of both expressions can be calculated easily at the end of the identification process. However, in many applications the user is not interested in the estimated parameters and their uncertainty but wants to calculate from these parameters other system characteristics such as the transfer function or the pole positions. In Section 14.12 it is shown that the Cramér-Rao lower bound of these derived quantities is generated by simple transformation laws, obtained from the first-order derivatives of the actual transformation. Remark that the same laws also apply to the approximated covariance matrices such as those obtained in Theorem 7.21:

$$\operatorname{Cov}(f(x)) \approx \left.\frac{\partial f(x)}{\partial x}\right|_{x=\mu_x} \operatorname{Cov}(x) \left(\left.\frac{\partial f(x)}{\partial x}\right|_{x=\mu_x}\right)^H \tag{9-1}$$

In practice, this works very well as long as the transformations are not heavily nonlinear (e.g., transfer function calculation), but sometimes it fails. A typical example of such a failure is the generation of the uncertainty regions on the estimated poles/zeros. Although the Cramér-Rao bounds (or the approximate covariance matrix) are correct, the actual uncertainties can significantly differ due to the fact that the asymptotic properties on these estimates are not yet reached for practical signal-to-noise ratios. For this case, we present a more precise numerical method to generate these bounds. In the following, the uncertainty on the transfer function, the transfer function residuals (difference between the measured and the estimated transfer function), and the poles are studied in detail. It is shown how to calculate these quantities, starting from the covariance matrix on the parameters C_θ. To evaluate the expression in practice, C_θ is replaced by the covariance matrix estimate, delivered by the identification algorithm, and the estimated parameter values are used instead of the limiting values parameters.

9.2.1 Uncertainty Bounds on the Calculated Transfer Functions

The transfer function interpretation of an identified model $G(\Omega, \hat{\theta})$ is used very intensively. It is also important to know the reliability of the estimated transfer function as a function of the frequency. Applying Eq. (9-1) gives the variance of the transfer function due to the noise sensitivity of the parameter estimates:

$$\operatorname{var}(G(\Omega, \hat{\theta})) \approx \left.\frac{\partial G(\Omega, \theta)}{\partial \theta}\right|_{\theta = \hat{\theta}} C_\theta \left(\left.\frac{\partial G(\Omega, \theta)}{\partial \theta}\right|_{\theta = \hat{\theta}}\right)^H \qquad (9\text{-}2)$$

9.2.2 Uncertainty Bounds on the Residuals

A very simple, but popular, validation test is to compare the differences between the measured FRF, $G(\Omega_k)$, and the modeled transfer function, $G(\Omega_k, \hat{\theta})$. In order to decide whether these residuals $G(\Omega_k) - G(\Omega_k, \hat{\theta})$ are significantly different from zero, their variance should be calculated. Equation (9-2) of the previous section cannot be applied here directly because $G(\Omega_k) - G(\Omega_k, \hat{\theta})$ depends now not only on $\hat{\theta}(Z)$ but also on the raw data $G(\Omega_k)$. Note that $\hat{\theta}(Z)$ and $G(\Omega_k)$ are correlated stochastic variables because they both depend on the same noise distortions N_Z. The extended expression (Eq. 17-37) is repeated here for the readers' convenience for complex-valued $f(Z, \theta)$ and Z:

$$\begin{aligned}\operatorname{Cov}(f(Z, \hat{\theta}(Z))) \approx {} & \left(\frac{\partial f(Z, \hat{\theta}(Z))}{\partial Z}\right) C_{N_Z} \left(\frac{\partial f(Z, \hat{\theta}(Z))}{\partial Z}\right)^H + \left(\frac{\partial f(Z, \theta)}{\partial \hat{\theta}(Z)}\right) \operatorname{Cov}(\underline{\hat{\theta}}(Z)) \left(\frac{\partial f(Z, \theta)}{\partial \hat{\theta}(Z)}\right)^H \\ & + 2\operatorname{herm}\left(\left(\frac{\partial f(Z, \hat{\theta}(Z))}{\partial Z}\right) \operatorname{Cov}(N_Z, \hat{\theta}(Z) - \tilde{\theta}(Z_0)) \left(\frac{\partial f(Z, \theta)^H}{\partial \hat{\theta}(Z)}\right)\right) \qquad (9\text{-}3)\end{aligned}$$

$$\operatorname{Cov}(N_Z, \hat{\theta}(Z) - \tilde{\theta}(Z_0)) \approx -C_{N_Z} \left(\frac{\partial \varepsilon(\hat{\theta}(Z), Z)}{\partial Z}\right)^H \left(\frac{\partial \varepsilon(\theta, Z)}{\partial \hat{\theta}(Z)}\right) \operatorname{Cov}(\underline{\hat{\theta}}(Z))$$

We apply this to the residual $G(\Omega_k) - G(\Omega_k, \hat{\theta})$, for deterministic excitations (the expected value with respect to a random input disappears), assuming that there is no input noise, and $\hat{\theta} = \hat{\theta}_{\mathrm{ML}}$ considering $G(\Omega_k)$ as the raw data (see 6-1). The following expression is obtained:

$$\begin{aligned}\operatorname{var}(G(\Omega_k) - G(\Omega_k, \hat{\theta})) = {} & \sigma_G^2(k) - 2\left(\frac{\partial G(\Omega_k, \theta)}{\partial \hat{\theta}}\right) V_{\mathrm{ML}}{}''^{-1}(\hat{\theta}) \left(\frac{\partial G(\Omega_k, \theta)}{\partial \hat{\theta}}\right)^H \\ & + x^H(\Omega_k, \hat{\theta}) \left(\sum_{k=1}^{F} \frac{1}{\sigma_G^2(k)} \left(\frac{\partial G(\Omega_k, \theta)}{\partial \hat{\theta}}\right)^H \left(\frac{\partial G(\Omega_k, \theta)}{\partial \hat{\theta}}\right)\right) x(\Omega_k, \hat{\theta})\end{aligned} \qquad (9\text{-}4)$$

with

$$x(\Omega_k, \hat{\theta}) = V_{\mathrm{ML}}{}''^{-1}(\hat{\theta})\left(\frac{\partial G(\Omega_k, \theta)}{\partial \hat{\theta}}\right)^H$$

$$V_{\mathrm{ML}}{}''^{-1}(\hat{\theta}) = \sum_{k=1}^{F} \frac{1}{\sigma_G^2(k)}\left(\frac{\partial G(\Omega_k, \theta)}{\partial \hat{\theta}}\right)^H \left(\frac{\partial G(\Omega_k, \theta)}{\partial \hat{\theta}}\right) + \sum_{k=1}^{F} \frac{G(\Omega_k) - G(\Omega_k, \hat{\theta})}{\sigma_G^2(k)} \frac{\partial^2 G(\Omega_k, \theta)}{\partial \hat{\theta}^2}$$

and where the second term in $V_{\mathrm{ML}}{}''^{-1}(\hat{\theta})$ disappears if there are no modeling errors. Noticing that

$$\sigma_G^2(\Omega_k, \hat{\theta}) = x^H(\Omega_k, \hat{\theta})\left(\sum_{k=1}^{F} \frac{1}{\sigma_G^2(k)}\left(\frac{\partial G(\Omega_k, \theta)}{\partial \hat{\theta}}\right)\left(\frac{\partial G(\Omega_k, \theta)}{\partial \hat{\theta}}\right)^H\right) x(\Omega_k, \hat{\theta}) \tag{9-5}$$

Eq. (9-4) finally becomes

$$\begin{aligned} \operatorname{var}(G(\Omega_k) - G(\Omega_k, \hat{\theta})) &= \sigma_G^2(k) - \sigma_G^2(\Omega_k, \hat{\theta}) - \Delta_G(k) \\ \Delta_G(k) &= x^H(\Omega_k, \hat{\theta})\left(\sum_{k=1}^{F} (G(\Omega_k) - G(\Omega_k, \hat{\theta})) \frac{\partial^2 G(\Omega_k, \theta)}{\partial \hat{\theta}^2}\right) x(\Omega_k, \hat{\theta}) \end{aligned} \tag{9-6}$$

If there are no model errors, this expression can be further reduced to

$$\operatorname{var}(G(\Omega_k) - G(\Omega_k, \hat{\theta})) = \sigma_G^2(k) - \sigma_G^2(\Omega_k, \hat{\theta}) \tag{9-7}$$

Practical Application. In general, $\sigma_G^2(\Omega_k, \hat{\theta}) \ll \sigma_G^2(k)$ so that the compensation in (9-7) can be neglected. $\sigma_G^2(\Omega_k, \hat{\theta})$ can become of the same order as $\sigma_G^2(k)$ only at those frequencies where the model is very flexible and depends only on a few data points (e.g., very sharp resonances). Because in this situation both terms in (9-7) cancel each other, the result becomes extremely sensitive to model errors. Expression (9-7) can even become negative! In that case the general expression (9-6) should be used. However, as the model errors are not accessible, this is impractical and leads us to the following conclusions: use $\sigma_G^2(k)$ as the uncertainty on the residuals. If $\sigma_G^2(\Omega_k, \hat{\theta}) \approx \sigma_G^2(k)$ the user should accept that in that region he cannot detect the presence of model errors because he has no reliable estimate of the residual uncertainty to decide whether or not they are significantly different from zero.

Example 9.1 (Calculation of the Uncertainty of Transfer Function Residuals): A very popular validation test is to compare the estimated transfer function $G(\Omega, \hat{\theta}_{\mathrm{ML}}(Z))$ with the measured transfer function, obtained directly from the measured input-output spectra: $\hat{G}(\Omega_k) = Y(k)/U(k)$. In order to decide whether the errors $\hat{G}(\Omega_k) - G(\Omega_k, \hat{\theta}_{\mathrm{ML}}(Z))$ are significantly different from zero, the variance of these residuals is calculated. Although the raw data were U and Y, we still use expression (9-7) with σ_G^2 calculated from the (co)variances σ_U^2, σ_Y^2, and σ_{YU}^2 with (2-25). A simulation is made on the system:

$$G(z^{-1}) = \frac{5.619{\times}10^{-3} + 2.248{\times}10^{-2}z^{-1} + 3.371{\times}10^{-2}z^{-2} + 2.248{\times}10^{-2}z^{-3} + 5.619{\times}10^{-3}z^{-4}}{1 - 1.585z^{-1} + 2.124z^{-2} - 1.544z^{-3} + 0.9034z^{-4}}$$

It is excited at the frequencies $kf_s/128$, $k = 1, 2, \dots, 44$, with $|U(k)| = 1$, and $\sigma_U(k) = 0.01\sqrt{2}$, $\sigma_Y(k) = 0.005\sqrt{2}$. The number of frequencies was kept very small in order to illustrate the effect of model errors on the uncertainty bounds. 1000 simulations are made and processed for a model order of $G(\Omega, \theta)$ equal to 4/4, 3/4, and 2/4. The results are given in Figure 9-1. The figures in the upper row compare the predicted uncertainty on the transfer function with the actual observed uncertainty. As can be seen, very good agreement is obtained as long as the model errors are small (models 3/4 and 4/4), while deviations become visible for model 2/4 at the second resonance peak. In this case, significant model errors are present. Observe that the uncertainty std($G(\Omega, \hat{\theta})$) on the estimated parametric model can become even larger than the measurement uncertainty $\sigma_G(k)$. The lower row shows the uncertainty on the residuals. For models 3/4 and 4/4 the theoretical values (9-7) and the observations are again in very good agreement. Note also that at most frequencies the standard deviation of the residual is almost equal to the measurement uncertainty (compare $\sigma_G(k)$ of the upper plot with std($G(\Omega_k) - G(\Omega_k, \hat{\theta})$) of the lower plot). Only at the second resonance peak (most important frequency band!) is there a significant drop. This is due to the fact that only two data points are put at the resonance peak so that the model can follow the raw data almost completely, leading to small residuals. Because of the errors for model 2/4, the predicted residual variance even becomes negative (the compensation in (9-7) fails), so that it loses all value. □

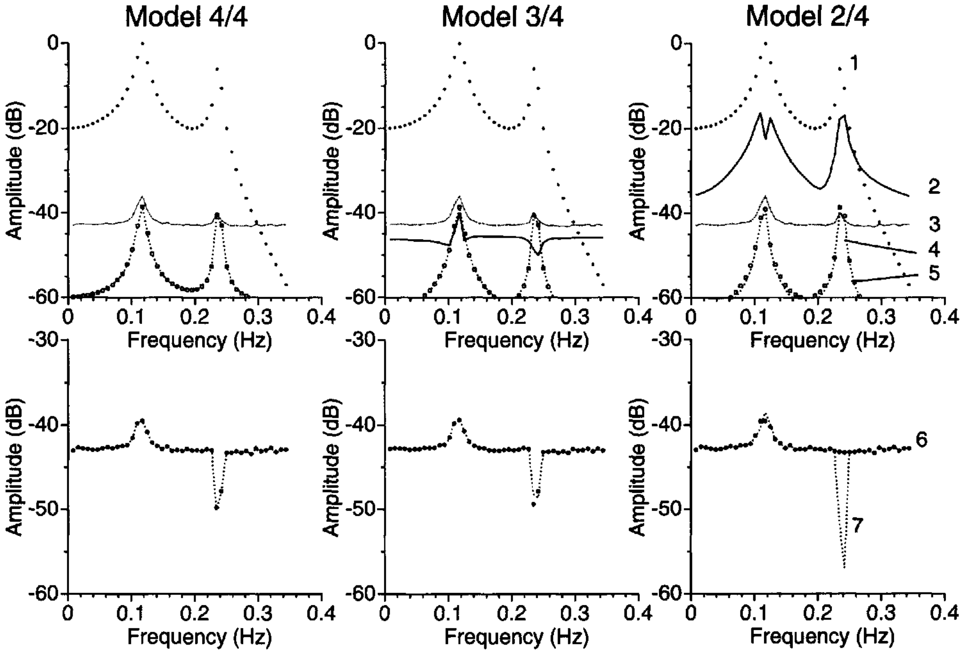


Figure 9-1. Study of the uncertainty bounds of the residuals for different levels of the model error. 1: $G_0(z_k^{-1})$, 2: model error, 3: measurement uncertainty $\sigma_G(k)$, 4: theoretic value of std($G(z_k^{-1}, \hat{\theta})$), 5: sample value of std($G(z_k^{-1}, \hat{\theta})$), 6: sample value of std($G(\Omega_k) - G(\Omega_k, \hat{\theta})$), 7: theoretic value of std($G(\Omega_k) - G(\Omega_k, \hat{\theta})$).

9.2.3 Uncertainty Bounds on the Poles/Zeros

The dispersion of the estimated parameters $\hat{\theta}$ around their mean value is given by the covariance matrix C_θ. Assuming that the estimates are normally distributed, the most compact uncertainty regions are ellipses. Practice has shown that this is a very usable description for realistic signal-to-noise ratios if θ are the coefficients of the numerator and denominator polynomials of the transfer function model. In the previous section it was shown how to calculate the covariance matrix of related system characteristics using linear approximations. However, if the user is interested in the uncertainty of the poles/zeros of the estimated system, it turns out that this linearization may fail. Even for high signal-to-noise ratios, the uncertainty ellipses calculated for the poles and zeros may not cover the true uncertainty regions. This is illustrated in the following simulation example. Consider the system $G(s)$ with zeros $-1.4355 \pm j4.0401$ and poles $-1.3010 \pm j4.8553, -3.5543 \pm j3.5543, -4.8553 \pm j1.3010$. The system has one dominating pole-zero pair and two pole pairs that have a smaller impact on the system. The transfer function is measured in 101 equidistant points between 0 and 1.25 rad/s with a signal-to-noise ratio of 40 dB $(\sigma_G(k) = |G(j\omega_k)|/100)$. Although we specified all characteristics in the frequency domain, the results are completely independent of the method that is used to identify the system (time or frequency domain identification). The only important information that is used are the model parameters and their covariance matrix. 10,000 realizations were generated and for each of them the poles/zeros were calculated and are shown in Figure 9-2. Also, the "classical" 95% confidence ellipsoids calculated using Eq. (9-1) are shown (see Guillaume et al., 1989). In this figure it is clearly seen that the shape of the uncertainty regions differs significantly from the ellipsoids (for the nondominating poles); and that even for the dominating pole/zeros the uncertainties are significantly underestimated. This is an unacceptable result because it is used as an input to many design procedures. Consequently, there is a need for more precise techniques to produce reliable uncertainty regions. The basic idea behind the improved technique is explained in the next section, and the precise mathematical description is given in Appendix 9.B.

9.2.3.1 Improved Method—Practical Calculation. The basic idea is to consider one pole (or zero) as a parameter and to move it away from its estimated position. The position of the remaining poles/zeros is shifted such that the total impact of the movement on the cost function is minimized. This step is the major difference from the method presented by Walter and Pronzato (1997). In Appendix 9.B it is shown that it is sufficient to observe the quadratic form

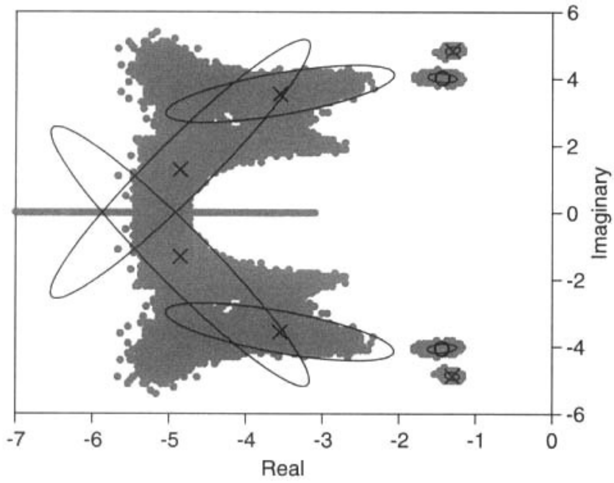


Figure 9-2. 95% confidence ellipsoids compared with the estimated poles and zeros of 10,000 simulations.

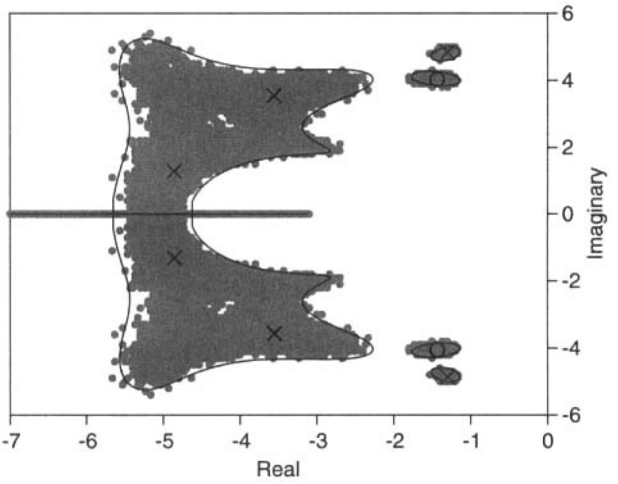


Figure 9-3. 95% confidence regions of the test system, calculated by perturbing the zeros and poles, using the coefficient covariance matrix.

$$\Delta\theta^T C_\theta^{-1}\Delta\theta \in As\chi^2(n_\theta) \tag{9-8}$$

Once this form reaches its maximum acceptable level given by the $p\%$ percentile $\chi_p^2(n_\theta)$, the border of the confidence region is found. The subsequent steps will follow that border so that the boundary is constructed (Vuerinckx et al., 1998).

9.2.3.2 Example. The improved method is applied to the previous example and the results are shown in Figure 9-3. It can be observed that there is now a very good match between the observed and the calculated uncertainty regions. In order to show how the confidence regions start to deviate from an ellipsoidal form, the 95% bounds are drawn in Figure 9-4 for increasing SNR. Starting from the same conditions as in the previous simulation (SNR=40 dB), the signal-to-noise ratio is increased in steps of 6 dB to 64 dB. Note that even for a high SNR the ellipsoidal form is not followed.

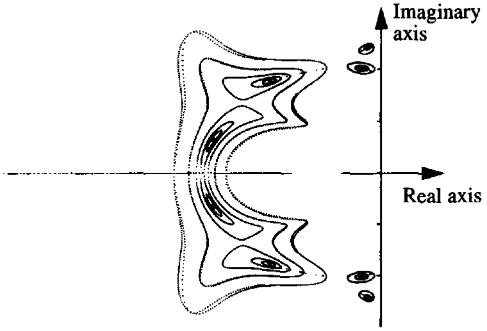


Figure 9-4. Evolution of the confidence regions as a function of the SNR (40 dB, 46 dB, 52 dB, 58 dB, 64 dB).

9.3 AVOIDING OVERMODELING

9.3.1 Introduction: Impact of an Increasing Number of Parameters on the Uncertainty

In this section we look into the dependence of the model variability on the model complexity. During the modeling process it is often quite difficult to decide whether or not the introduction of a new parameter is meaningful. A simple strategy would be to fit all the parameters that could be of possible interest, but this is not a good idea because the uncertainty on the estimates will then be increased. Consider a model with a partitioned set of parameters $\theta = (\theta_1, \theta_2)$. What is the impact on the model uncertainty if the simple model $G(\theta_1)$ is extended to the more complex one $G(\theta_1, \theta_2)$? In Example 1.5, it was illustrated that the uncertainty will increase. Here it is shown that this is a general result. Consider the information matrix of the full model:

$$Fi(\theta_1, \theta_2) = \begin{bmatrix} Fi_{11} & Fi_{12} \\ Fi_{21} & Fi_{22} \end{bmatrix} \tag{9-9}$$

The covariance matrix of the simple model is $C(\theta_1) = Fi_{11}^{-1}$, while the covariance matrix of the complete model is $C(\theta_1, \theta_2) = Fi^{-1}$. The covariance matrix C_{θ_1} of the subset θ_1 is related to the covariance matrix $C(\theta_1)$ of the complete set by

$$C_{\theta_1} = Fi_{11}^{-1} + Fi_{11}^{-1} Fi_{12} Fi_{22}^{-1} Fi_{21} Fi_{11}^{-1} = C(\theta_1) + \Delta \tag{9-10}$$

(see (13-8)). Because $\Delta \geq 0$ it is clear that adding additional parameters to a model increases its uncertainty. A similar result is available for transfer function estimation. Ljung (1985) has shown that in case of output noise only, the asymptotic expression (for the order increasing to ∞) for the variance $\sigma_G^2(\Omega, \hat{\theta}(Z))$ on the estimated transfer function becomes

$$\sigma_G^2(\Omega, \hat{\theta}(Z)) \sim \frac{n_\theta}{N} \mathrm{SNR}^{-1}(\omega) \tag{9-11}$$

This expression gives a great deal of insight: the uncertainty on the parametric model is proportional to that of the nonparametric estimate, but due to the averaging effect (over the frequency) of the parametric model an additional noise reduction of n_θ/N appears. The dependence on n_θ is illustrated in Figure 9-5. A 5th-order FIR system is identified, the first time using a 5th-order model ($n_\theta = 5$) and the second time with a 50th-order model ($n_\theta = 50$). From Eq. (9-11) it is expected that the standard deviation should increase about 9 dB, which is in agreement with the simulation results.

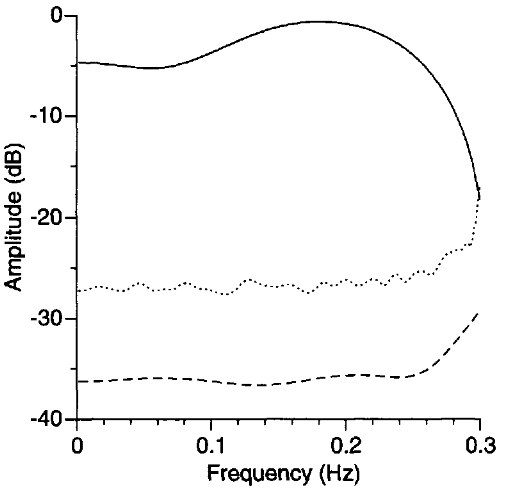


Figure 9-5. Dependence of $\sigma_G(\Omega, \hat{\theta}(Z))$ on the model order. ___ $G_0(z_k^{-1})$, --- $\sigma_G(\Omega, \hat{\theta}(Z))$ of 5th-order system, ... $\sigma_G(\Omega, \hat{\theta}(Z))$ of 50th-order system.

9.3.2 Balancing the Model Complexity versus the Model Variability

In the previous section it was illustrated that the systematic errors decrease with increasing model complexity. However, at the same time the model variability increases as shown in Eqs. (9-10) and (9-11). In practice, the optimal complexity should be selected from the available information. Usually this choice is based on the evolution of the cost function. As explained in Sections 17.6 and 17.7, it is not a good idea to select the model with the smallest cost function because it will continue to decrease if additional parameters are added. From a given complexity, the additional parameters no longer reduce the systematic errors but are used only to follow the actual noise realization on the data. As these vary from measurement to measurement, they increase only the model variability. Many techniques were proposed to avoid this unwanted behavior. These are based on extending the cost function with a model complexity term that estimates and compensates for the unwanted increasing model variability. Two popular methods are actually in use, the AIC (Akaike information criterion) and the MDL (minimum description length) (see Section 17.7):

$$\begin{array}{ll} \text{AIC:} & V_{\mathrm{ML}}(\hat{\theta}_{\mathrm{ML}}(Z), Z) + n_\theta \\ \text{MDL:} & V_{\mathrm{ML}}(\hat{\theta}_{\mathrm{ML}}(Z), Z) + \frac{n_\theta}{2}\ln(2\alpha F) \end{array} \qquad (9\text{-}12)$$

with n_θ the number of identifiable (free) model parameters (= total number of parameters minus the number of constraints), and F the number of frequencies. $\alpha = 1$ for output error problems, and $\alpha = 2$ for the errors-in-variables problem. MDL has a much better reputation than AIC. This is also illustrated in the following simulations.

9.3.2.1 Simulation of a Second-Order System. In the first simulation, a discrete time second-order system ($n_a = n_b = 2$) with independently uniformly distributed coefficients, $a_r, b_r \in [0, 1]$ for $r = 0, 1, 2$, was considered. Because the estimation is performed in the frequency domain, the stability of the system was not an issue, and we also kept the unstable systems in the simulation. The system is excited over the full frequency band $[0,\pi]$ with 200 equidistantly distributed frequencies. The measured output is disturbed with noise resulting in an average SNR of 37 dB. Then 83 random realizations of the random system were generated and identified. Each time all the models between 1/1 to 3/3 were tested and the best one was selected following the AIC and MDL rule (9-12). The results are shown in Table 9-1, giving how many times each model is selected. The MDL rule selects the correct model almost every time, while the AIC rule has a strong tendency to select models that are too complex.

TABLE 9-1 Selection of the Model Order Using the AIC and the MDL Rules

n_a n_b	1	2	3	n_a n_b	1	2	3
1	0	0	0	$n_b = 1$	0	0	0
2	1	53	5	$n_b = 2$	1	82	0
3	0	9	15	$n_b = 3$	0	0	0
		AIC				MDL	

9.3.2.2 Simulation of a Sixth-Order System. In the second simulation, a sixth-order system with well-separated resonances is identified. The user is actually interested only in a model for the middle resonance. Therefore, most power is focused in this band (see Figure 9-6). Only output noise was added (-40 dB). ($n_a = n_b = 6$). All models with $4 \leq n_a, n_b \leq 8$ are scanned, and the best one is selected following the AIC and MDL rule ($F = 200$). The results are given in Table 9-2. In this case none of the methods is able to select the correct 6/6 model. This is mainly due to the fact that only a part of the band is properly excited so that the concept "exact model" loses its value as is commonly experienced in practice. In order to get a better appreciation of the results, we focused on these realizations where MDL and AIC selected a different model (199 times over 385 realizations). The RMS error with respect to the exact model is plotted in Figure 9-6. The errors on the MDL models are smaller than those on the AIC models, but the uncertainties are very similar inside the frequency band of interest. This suggests that the additional model flexibility, used to model out-of-band effects, does not have a great impact on the in-band uncertainty. This is further analyzed in Section 9.4.1.

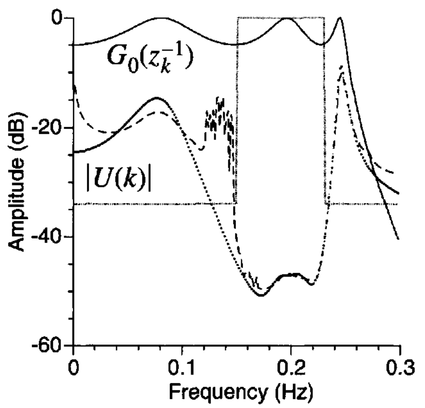


Figure 9-6. Comparison of the RMS error of AIC (---) and MDL (...) selected models for the realizations where a different selection is made.

Remarks

(i) In order to complete the view on the overmodeling problem, we also give the models as they would be selected on the cost function itself. Normally, the lowest value of the cost function should appear each time for the most complex model (8/8). However, due to the overmodeling, the numerical conditioning and the convergence properties of the search algorithm are strongly decreased. The search procedure was stopped after a maximum of 50 iterations, so that the algorithm could stop in suboptimal solutions. Nevertheless, it can still be observed that with-

TABLE 9-2 Selection of the Model Order Using the AIC and the MDL Rules

n_b \ n_a	4	5	6	7	8	n_b \ n_a	4	5	6	7	8
4	0	0	5	94	44	4	0	0	89	105	52
5	0	0	95	15	4	5	0	0	118	5	0
6	0	0	11	6	13	6	0	0	1	0	0
7	0	0	29	8	29	7	0	0	15	0	0
8	0	0	1	12	19	8	0	0	0	0	0
			AIC						MDL		

out an additional model complexity term, there is a strong tendency to select too complex models. Although this is not really a disaster from a model variability point of view (the uncertainty on the transfer function does not really explode), it still represents a lot of wasted work.

(ii) Although the selection of the model complexity is not that critical, it is important not to exaggerate the order (e.g., just doubling it), in order to avoid the appearance of coinciding pole-zero pairs that can create sharp resonances between two frequency points, resulting in a locally increased variance of the model.

TABLE 9-3 Cost Function–Based Selection

n_a n_b	4	5	6	7	8
4	0	0	0	1	4
5	0	0	0	0	1
6	0	0	1	3	14
7	0	0	3	26	58
8	0	0	3	95	176

Conclusions. Using the MDL rule, it is possible to balance the model complexity versus the model variability without requiring prior knowledge that would not be available in practice. The major contribution of the MDL rule is to give a data-based restriction on the maximum complexity of the models that need to be checked. As it is precisely these too complex models that require a lot of computation time, it can be concluded that the MDL rule essentially helps to save time.

Using the MDL rule, it is possible to go for the best model making an exhaustive scan over all possible models in a predefined set (e.g., $n_{\min} \le n_a, n_b \le n_{\max}$) and picking out the model with the smallest MDL cost function (9-12). The major disadvantage of this approach is that many models should be evaluated, and fine tuning is very time consuming. Consequently, a top-down approach will be presented in Section 9.5, where initially a model that is too complex is estimated and, next, the complexity is reduced by stripping off the superfluous parts, using the MDL rule again as a decision criterion. Special actions will be needed to avoid the numerical conditioning problems and to guarantee good convergence.

To solve the general model selection problem, it is also necessary to detect the presence of model errors so that unmodeled dynamics or nonlinear distortions can be detected. This will be discussed in the next section.

9.4 DETECTION OF UNDERMODELING

In the previous section we were mainly concerned to restrict the model complexity in order to avoid a noise sensitivity of the model that is too large. This might suggest that it is a good idea to select too simple models deliberately, hoping for a significant reduction of the noise sensitivity. This idea is checked by means of a simple simulation. Next we will analyze how can we detect, qualify, and quantify systematic errors. The detection step should indicate the presence of model errors. In the qualification step, it is checked if the model error is either due to too low a model order, so that there remain unmodeled dynamics, or due to nonlinear distortions. Finally, an idea is given about the average level of the model errors.

9.4.1 Undermodeling: A Good Idea?

In some applications, the user is not interested in a complete model covering the full frequency band. He wants only a good description in the frequency band of interest, which might be covered by a low-order model. Equation (9-11) suggests that a high-order model would suffer from a larger variance than the low-order model. A simulation is set up to analyze this problem. The same setup as in Section 9.3.2, simulation 9.3.2.2, is used. A hundred runs were made, and the results are shown in Figure 9-7. From the left side it is seen that in the frequency band of interest (where the input amplitude $|U(k)|$ is high) the standard deviations are about the same. However, on the right side it is seen that for the simple model, there remain significant model errors, even in the frequency band of interest. This shows that it is seemingly better to go for a sufficiently complex model that pushes the systematic model errors down to the noise level. This result conflicts with the previous asymptotic result where the model variability increased proportionally with n_θ. Note here also that the model variability is increased, but only outside the frequency band of interest, where the simple model is actually not applicable.

This brings the model complexity question to a mature level: how should the model complexity be chosen to balance the model variability versus its systematic errors? We advise the reader to increase the model complexity in the identification step until model errors can no longer be detected. If this model is too complex for his final goal, a model reduction step can be applied next. This offers the advantage that the user knows exactly what model errors he introduced himself.

9.4.2 Detecting Model Errors

In "classical" identification (Ljung, 1999), two tests are very popular to detect model errors. Both are based on the residuals that are given as the difference between the model-based predicted output and the actual measured output. In the most general case, a plant model and a noise model are estimated. If the "true models" are reached, it is shown that the residuals should be white. If one of both is wrong, correlation will be detected. In practice, the plant model is more important than the noise model. Many times the user is not really interested in modeling the noise characteristics. Some methods, such as the output-error

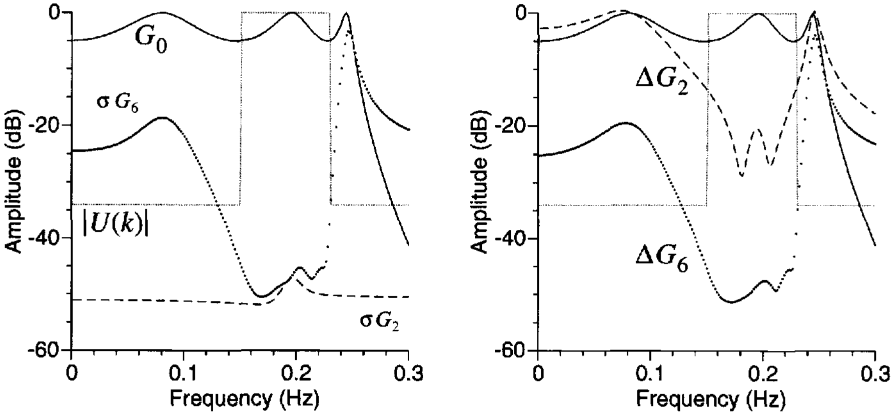


Figure 9-7. Impact of undermodeling on $\sigma_G(z_k^{-1}, \hat{\theta}(Z))$. Left: $\sigma_G(z_k^{-1}, \hat{\theta}(Z))$ for a second-order ($\sigma_{G_2}(z_k^{-1}, \hat{\theta}(Z))$) and a sixth-order ($\sigma_{G_6}(z_k^{-1}, \hat{\theta}(Z))$) system, and right: the mean model error of the transfer function estimate $\mathcal{E}\{|G(z_k^{-1}, \hat{\theta}(Z)) - G_0(z_k^{-1})|\}$.

method (Ljung, 1999), do not even estimate the noise model at the price of a poorer efficiency, but they are still consistent in open loop identification without input noise. In that case, the residuals will basically mimic the colored process noise, and, hence, they should not be white. Consequently, a whiteness test loses its applicability for these frequently occurring situations.

The other test checks for the presence of unmodeled dynamics by looking for cross-correlations between the residuals and the input. If all linear relations are modeled, the cross-correlation should not be significantly different from zero and this leads to a statistical validation test.

It is essential to note that in both tests no check of the absolute level of the residuals is made. This is intrinsically due to the fact that the variance of the disturbing noise is not estimated a priori from the raw data; it is estimated together with the plant and noise model. This is the major difference from the framework that is set up in this book. As explained before, periodic excitations make it possible to separate the signal and the noise before starting the identification process. So, an estimate of the noise model is obtained using the sample (co) variances $\hat{\sigma}_U^2(k)$, $\hat{\sigma}_Y^2(k)$, and $\hat{\sigma}_{YU}^2(k)$ as explained in Section 2.5.1. The knowledge of this noise model, which is obtained directly from the raw data independent of the identification process and the selected model, opens completely new possibilities. It will become possible to check for the amplitude of the residuals, which is a more direct measure of remaining model errors. Because the cost function is nothing other than the sum of the squared amplitudes of the normalized residuals, we will use its value as the primary check for the detection of model errors. The properties of the cost function are studied in detail in Section 17.6. It is shown that the expected value of the cost function (based on a nonparametric noise model) can be split in two parts, the first one accounting for the noise contributions and the second one being due to modeling errors (Theorem 17.12). Consider the cost function evaluated in the estimated parameters: $V_{\text{ML}}(\hat{\theta}_{\text{ML}}(Z), Z)$ (see Section 7.11); then the following result holds:

Theorem 9.2 (Properties Global Minimum ML Cost Function): The global minimum $V_{\text{ML}}(\hat{\theta}_{\text{ML}}(Z), Z)$ of the maximum likelihood cost function (7-82) has the following properties:

1. In the presence of model errors, deterministic inputs (Z_0 is deterministic), and circular complex normally distributed noise N_Z, $V_{\text{ML}}(\hat{\theta}_{\text{ML}}(Z), Z)$ is asymptotically ($F \to \infty$) normally distributed with mean and variance

$$\begin{aligned} \mathscr{E}\{ V_{\text{ML}}(\hat{\theta}_{\text{ML}}(Z), Z) \} &\approx V_{\text{noise}} + V_{\text{model}} \\ \text{var}(V_{\text{ML}}(\hat{\theta}_{\text{ML}}(Z), Z)) &\approx V_{\text{noise}} + 2V_{\text{model}} \end{aligned} \tag{9-13}$$

(assumptions of Section 7.6.4),

2. In the absence of model errors, deterministic or random inputs (Z_0 is deterministic or random), and circular complex normally distributed noise N_Z, $V_{\text{ML}}(\hat{\theta}_{\text{ML}}(Z), Z)$ is asymptotically ($F \to \infty$) normally distributed with mean and variance

$$\begin{aligned} \mathscr{E}\{ V_{\text{ML}}(\hat{\theta}_{\text{ML}}(Z), Z) \} &\approx V_{\text{noise}} \\ \text{var}(V_{\text{ML}}(\hat{\theta}_{\text{ML}}(Z), Z)) &\approx V_{\text{noise}} \end{aligned} \tag{9-14}$$

(assumptions of Sections 7.6.4 and 7.6.6),
with $V_{\text{noise}} = F - n_\theta/2$ and $V_{\text{model}} = \varepsilon^T(\tilde{\theta}_{\text{ML}}(Z_0), Z_0)\varepsilon(\tilde{\theta}_{\text{ML}}(Z_0), Z_0)$.

Proof. See Appendix 9.A. □

This result gives an extremely simple test to check for model errors. If the actual cost function is significantly larger than the expected value, model errors are present. Otherwise, it can be decided that no significant model errors are detected. This conclusion cannot be made within the "classical" identification schemes because those algorithms estimate the noise model and its variance. They cannot recognize the presence of white residuals that are too large. Here, these errors are detected because the noise (co)variances are known a priori. The impact of replacing the exact variance by its sampling value is studied in Theorem 8.5, where it is shown that similar rules still apply. If no model errors are present, then (9-14) becomes

$$\begin{aligned} &\mathcal{E}\{V_{\text{SML}}(\hat{\theta}_{\text{SML}}(Z), Z)\} \approx \frac{M-1}{M-2}V_{\text{noise}} \\ &\text{var}(V_{\text{SML}}(\hat{\theta}_{\text{SML}}(Z), Z)) \approx \frac{(M-1)^3}{(M-2)^2(M-3)}V_{\text{noise}} \end{aligned} \tag{9-15}$$

(see Theorem 8.5).

9.4.3 Qualifying and Quantifying the Model Errors

In this section we will develop the theory explicitly using the output error framework instead of the errors-in-variables viewpoint. The major reason for the choice is that we also want to include the nonlinear distortions, and this theory was set up assuming output noise only. In practice, we first do the identification in the errors-in-variables framework and next make the validation on the measured transfer function $G(\Omega_k)$ obtained from the raw data by Eq. (2-17) and the estimator $G(\Omega, \hat{\theta})$, assuming that the following assumptions are valid:

Assumption 9.3 (pdf FRF Measurement Errors): The noise $N_G(k)$ on the FRF measurement is independent (over k), circular complex normally distributed.

Assumption 9.4 (FRF Estimate): The estimate $\hat{\theta} \approx \hat{\theta}_{\text{ML}}(Z)$, with

$$\hat{\theta}(Z) = \arg\min_{\theta} \sum_{k=1}^{F} \frac{|G(\Omega_k) - G(\Omega_k, \theta)|^2}{\sigma_G^2(k)} \tag{9-16}$$

The residuals are the difference between the measured and the modeled transfer function, weighted by the standard deviation on the FRF measurement:

$$\varepsilon(\hat{\theta}, \Omega_k) = \frac{G(\Omega_k) - G(\Omega_k, \hat{\theta}(Z))}{\sigma_G(k)} \tag{9-17}$$

These residuals will be used to qualify the nature of the error, once model errors are detected (the cost function is too large). Because we also want to include nonlinear distortions in this analysis, we have to obey the assumptions and restrictions put forward in Chapter 3: normalized excitations $x_F \in \mathbb{E}_F$ (Definition 3.2); the class of systems from Definition 3.5; and the noise properties of a frequency domain experiment with $\sigma_U^2(k) = 0$ (see Section 7.6). The measured FRF can then be written as the sum of three parts:

$$G(\Omega_k) = G_R(\Omega_k) + G_S(\Omega_k) + N_G(k) \tag{9-18}$$

with $G_R(\Omega_k)$ the related dynamic system (the best linear approximation to the overall system), $G_S(\Omega_k)$ the stochastic nonlinear contributions, and $N_G(k)$ the errors due to the output noise. The related dynamic system $G_R(\Omega_k)$ consists of two parts:

$$G_R(\Omega_k) = G_0(\Omega_k) + G_B(\Omega_k) \tag{9-19}$$

with $G_0(\Omega_k)$ the underlying linear system and $G_B(\Omega_k)$ the bias or systematic errors due to the nonlinear distortions. $G_S(\Omega_k)$ is called a stochastic contribution; it behaves as uncorrelated (over the frequencies) noise, although the reader should be aware that it is not really a noise component. Its properties were explicitly stated in Theorem 3.9. Due to this noisy behavior, the presence of nonlinear distortions is often not recognized, although it is exactly this noisy behavior that will make it possible to detect their presence. The linear model will converge to the related linear dynamic system $G_R(\Omega_k)$ if the model complexity is high enough. So, depending on the nonlinear distortion and the nature of the excitation, but opposed to the classical validation techniques, the user gets a warning about their presence.

The model errors can be written as:

$$\begin{aligned} G(\Omega_k) - G(\Omega_k, \hat{\theta}(Z)) &= G_E(\Omega_k) - G_v(\Omega_k, \hat{\theta}(Z)) + q_k \\ G_E(\Omega_k) &= G_R(\Omega_k) - G(\Omega_k, \theta_*) \\ G_v(\Omega_k, \hat{\theta}(Z)) &= G(\Omega_k, \hat{\theta}(Z)) - G(\Omega_k, \theta_*) \end{aligned} \tag{9-20}$$

$G_E(\Omega_k)$ is the bias error due to undermodeling (unmodeled dynamics and approximation of the nonlinear system), $G_v(\Omega_k, \hat{\theta}(Z))$ is the model uncertainty contribution (the estimated parameters $\hat{\theta}(Z)$ are different from θ_* due to the noise), and $q_k = N_G(k) + G_S(\Omega_k)$ are the stochastic errors (see Chapter 3).

The basic idea to qualify the model errors is based on the sample correlation analysis of the transfer function residuals. Consider $\hat{R}_{\varepsilon\varepsilon}(m)$:

$$\hat{R}_{\varepsilon\varepsilon}(m) = \frac{1}{F-m}\sum_{k=1}^{F-m}\frac{(G(\Omega_k) - G(\Omega_k, \hat{\theta}(Z)))\overline{(G(\Omega_{k+m}) - G(\Omega_{k+m}, \hat{\theta}(Z)))}}{\sigma_G(k)\sigma_G(k+m)} \tag{9-21}$$

In the following theorem it is shown that the $\hat{R}_{\varepsilon\varepsilon}(m)$ converges to zero, except at the origin $(m = 0)$, if the selected model includes the RLDS model structure $(G_E(\Omega_k) = 0)$.

Theorem 9.5 (Properties Sample Correlation in the Absence of Model Errors): Consider a system belonging to the set $\mathbb{S}$ (see Definition 3.5), excited with a random multisine $u_F \in \mathbb{E}_F$. If no unmodeled dynamics are present $(G_E(j\omega_k) = 0)$, then under Assumption 9.3,

$$\begin{aligned}\hat{R}_{\varepsilon\varepsilon}(m) &= O_{\mathrm{p}}(F^{-1/2}) \qquad m \neq 0 \\ \hat{R}_{\varepsilon\varepsilon}(0) &= \frac{1}{F}\sum\nolimits_{k=1}^{F}\frac{\sigma_q^2(k)}{\sigma_G^2(k)} + O_{\mathrm{p}}(F^{-1/2})\end{aligned} \tag{9-22}$$

Proof. See Appendix 9.C. □

Note that this theorem gives an alternative interpretation of the cost function $(\mathcal{E}\{V_{\mathrm{ML}}(\hat{\theta}_{\mathrm{ML}}(Z), Z)\} \approx F\hat{R}_{\varepsilon\varepsilon}(0))$.

In order to separate the stochastic nonlinear distortions from the unmodeled linear dynamics, it is assumed that the latter have a smooth behavior. For a fixed bandwidth of the experiments, the density of the frequency grid increases in proportion to F. The neighboring model errors will be almost equal, for F large enough. Moreover, we assume that the model errors are bounded and that the derivative with respect to the parameters behaves well. This leads to the following formal assumptions:

Assumption 9.6 (Smooth Errors): For F sufficiently large

(i) Smooth unmodeled dynamics: $G_E(\Omega_k)\overline{G_E}(\Omega_{k+1}) = |G_E(\Omega_k)|^2 + O(F^{-1})$.

(ii) Smooth model variability: $\mathcal{E}\{G_v(\Omega_k)\overline{G_v}(\Omega_{k+1})\} = \mathcal{E}\{|G_v(\Omega_k)|^2] = \sigma_{Gv}^2(k)$.

(iii) Smooth disturbing noise spectrum: $\sigma_G(k) \approx \sigma_G(k+1)$.

(iv) The normalized model errors $G_E(\Omega_k)/\sigma_G(k)$ are bounded.

Under these assumptions, it can be shown that $\hat{R}_{\varepsilon\varepsilon}(m)$ is significantly different from zero for $m \neq 0$ in the presence of unmodeled linear dynamics, and a hypothesis test is set up to check this. Under Assumption 9.6 it is also possible to bound the unmodeled dynamics. A similar attempt has already been made, starting from the value of the cost function (Schoukens and Pintelon, 1991), but this idea cannot be applied directly in this nonlinear context because the cost function is too large not only due to the model errors of the related dynamic system but also due to the stochastic nonlinear contributions. In order to separate both effects, $\hat{R}_{\varepsilon\varepsilon}(1)$ is considered.

Theorem 9.7 (Properties Sample Correlation at Lag One): Consider a system belonging to the system set S, excited with a random multisine $x_F \in \mathbb{E}_F$. Under Assumptions 9.3 and 9.6, $\hat{R}_{\varepsilon\varepsilon}(1)$ depends only on the unmodeled dynamics:

$$\hat{R}_{\varepsilon\varepsilon}(1) = \frac{1}{F-1}\sum\nolimits_{k=1}^{F-1}\frac{|G_E(\Omega_k)|^2}{\sigma_G^2(k)} + O_{\mathrm{p}}(F^{-1/2}) \tag{9-23}$$

Proof. See Appendix 9.E. □

A stochastic bound on the bias error can be given starting from the sample value:

$$\mathrm{Prob}\left(\frac{1}{F-1}\sum_{k=1}^{F-1}\frac{|G_E(\Omega_k)|^2}{\sigma_G^2(k)} \leq |\hat{R}_{\varepsilon\varepsilon}(1)| + \alpha\,\mathrm{std}(\hat{R}_{\varepsilon\varepsilon}(1))\right) = P_\alpha \tag{9-24}$$

(see Appendix 9.F). In this expression the norm of $\hat{R}_{\varepsilon\varepsilon}(1)$ is considered because in general it is a complex value. Hence, chi-square tables should be used to determine the value P_α, because $\hat{R}_{\varepsilon\varepsilon}(1)$ is asymptotically circular complex normally distributed (see Section 17.5.2). For example, the 95% level is given by the bound $\sqrt{3}\text{std}(\hat{R}_{\varepsilon\varepsilon}(1))$ (see Appendix 9.F). The reader should be aware that a hypothesis test does not guarantee that no model errors are present. It only makes a statement on the probability that there are no significant (with respect to the noise level $\sigma_G(k)$) model errors left.

Conclusion. The following qualification rules can be used:

1. The cost function equals the noise value within the uncertainty bounds, for example, with 95% confidence

$$V_{\text{ML}}(\hat{\theta}_{\text{ML}}(Z), Z) \in [V_{\text{noise}} - 2V_{\text{noise}}^{1/2}, V_{\text{noise}} + 2V_{\text{noise}}^{1/2}] \tag{9-25}$$

with $V_{\text{noise}} = F - n_\theta/2$ (see (9-14)). If the sample (co)variances are used, then (9-25) should be adapted according to (9-15). No model errors are detected. This test should be confirmed by the fact that $\hat{R}_{\varepsilon\varepsilon}(m) \approx 0$. If this is not the case, another error source, not discussed in this section, should be present.

2. The cost function is significantly larger than the noise value, for example, with 95% confidence

$$V_{\text{ML}}(\hat{\theta}_{\text{ML}}(Z), Z) > V_{\text{noise}} + 2V_{\text{noise}}^{1/2} \tag{9-26}$$

(see (9-14)). If the sample (co)variances are used, then (9-26) should be adapted according to (9-15). Model errors are present. These can then be qualified by checking the correlation between the transfer function residuals:

2a. $\hat{R}_{\varepsilon\varepsilon}(m) \neq 0$ for $m \neq 0$: there are still unmodeled dynamics.

2b. $\hat{R}_{\varepsilon\varepsilon}(m) \approx 0$ for $m \neq 0$: no unmodeled dynamics can be detected. This behavior can be explained, assuming the presence of nonlinear distortions.

In order to test whether $\hat{R}_{\varepsilon\varepsilon}(m) \approx 0$, a percentile test can be used. In such a test it is checked if, e.g., $\alpha\%$ of the $\hat{R}_{\varepsilon\varepsilon}(m)$ samples have an amplitude below the predicted $\alpha\%$ level. This level can be calculated from the expressions given in Appendix 9.F. It turned out from our experience that this is a very sensitive test to indicate the presence of unmodeled dynamics.

3. If nonlinear distortions are detected, a new experiment can be set up to measure the variances of the nonlinear distortions and use them in a slightly modified estimation procedure (see Section 8.7 for a detailed description of the procedure). This will result in a model with a smaller variance and improved error bounds.

9.4.4 Illustration on a Mechanical System

The previous ideas are illustrated on a vibrating robot arm. Jan Swevers and Dirk Torfs (Department PMA of the Katholieke Universiteit Leuven, Belgium) have provided us with the experimental data (Torfs et al., 1998). As the input, the driving couple is measured, while the output is the acceleration at the tip of the robot arm. Ten periods are measured, each period consisting of 4096 points sampled at a frequency of 500 Hz. Only the odd harmonics (1, 3,..., 199) are excited. The identification results are shown in Figure 9-8 for models of order 4/4 and 6/6. Although the model 4/4 already gives quite a good fit, the correlation analysis clearly indicates that there are still significant (with respect to the noise level)

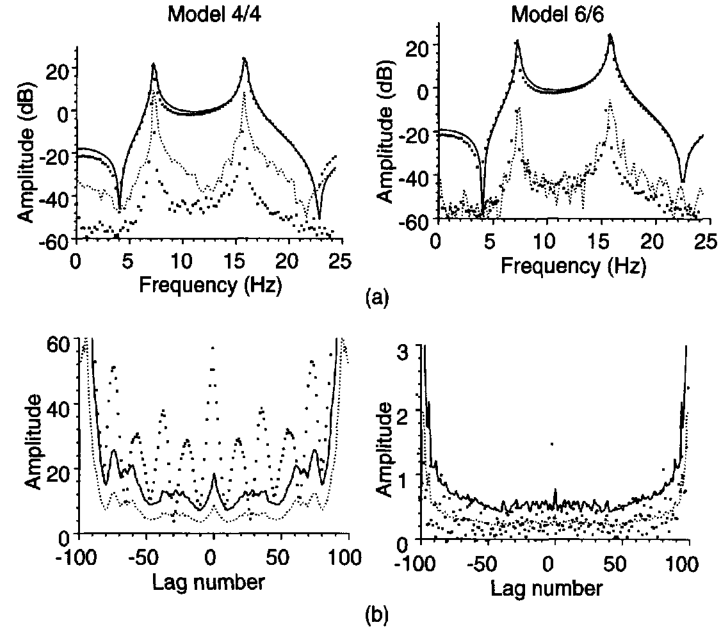


Figure 9-8. Illustration of model error detection and qualification on a vibrating robot arm. (a) The identified transfer function, dots: measurements, — model, ··· model errors, × measurement uncertainty σ_G; (b) $\hat{R}_{\varepsilon\varepsilon}(m)$, dots: measurement, ··· 50% bound, — 95% bound.

unmodeled dynamics. Hence, it makes sense to increase the model order. The cost function is 4964.8 while a value of 95.5 is expected. A closer inspection of Figure 9-8(a) shows that the errors are, indeed, larger than the measurement uncertainty. Increasing the model order to 6/6 reduces the cost function to 220.5 (compared with an expected value of 93.5), still pointing to significant model errors. However, the correlation analysis cannot detect any more unmodeled dynamics. 58% of the correlation results are above the 50% percentile (98% for the 4/4 model) and 4% above the 95% percentile (92% for the 4/4 model). From this, we conclude that it makes no sense to increase the model order further, and most probably there are nonlinear distortions present. This was confirmed by more detailed tests.

Remarks

(i) In practice, it is advisable not to use the correlation results at all the lags. The uncertainty on it increases very fast for the extreme lag numbers due to the small number of points that add to the sum. It is better to restrict the analysis to lag numbers that are smaller than half the number of frequencies.

(ii) The model validation is extremely simplified due to the presence of high-quality FRF measurements, so that even small model errors on the transfer function become visible.

9.5 MODEL SELECTION

In this section a "new" model selection procedure is proposed. In the classical approach (Stoica et al., 1986), a first guess of the model order is made directly from the raw data. A typical example is to plot the nonparametric frequency response function to get an initial idea about the required model complexity. A first trial is made and then the complexity is adapted to the results of the validation test. Usually, the model order is increased step-by-step until the point where acceptable validation results are obtained. The major reason for this cautious approach is the sensitivity of most algorithms to overmodeling. It usually results in very poor conditioning of the normal equations and convergence problems of the iterative schemes, so that the complexity can only be increased gradually. However, orthogonal parameterizations (see Section 7.16) guarantee good numerical conditioning, even in the case of extreme overmodeling, so that it becomes possible to reverse the previous sketched procedure (Rolain et al., 1997).

This is the approach that we will explain here. It consists basically of three steps:

1. Make an initial guess of the maximum order based on a rank decision of a raw data matrix. This choice should be conservative (biased toward a too high model order) so that the best order is below this selection.
2. The parameters of this high-order model are estimated.
3. A model reduction is performed by eliminating poles, zeros, or pole-zero pairs that do not significantly contribute to the model. The validity of each reduction is checked. If no further reduction is possible, a new estimate with the reduced order is made and the model reduction procedure is restarted.

A more detailed description of each step is given in the following. The whole procedure can be automatized, and from our experience it turns out that it results in reasonable and sometimes even better models than those selected by human operators.

9.5.1 Model Structure Selection Based on Preliminary Data Processing: Initial Guess

In order to start the model selection procedure automatically, an initial guess for the order is needed that is generated directly from the raw data. To do so, the user should specify an initial order $((n_a)_{\text{init}},\ (n_b)_{\text{init}})$ that is definitely too high, so that the best model order is guaranteed to be included. Next, the number of possible pole/zero cancellations is estimated. An identification method that is linear in the parameters is used. So, an improved order selection boils down to a rank detection problem on the raw data matrices. Consider the following equation error formulation (Section 7.8.2):

$$e(\Omega_k, \theta, Z(k)) = Y(k)A(\Omega_k, \theta) - U(k)B(\Omega_k, \theta) \tag{9-27}$$

where A and B are polynomials of order $(n_a)_{\text{init}}$ and $(n_b)_{\text{init}}$, respectively. The Jacobian matrix $J(Z) = \partial e(\theta, Z)/\partial\theta$ is parameter independent and its rank is, at most, $(n_a)_{\text{init}} + (n_b)_{\text{init}} + 1$ in the noiseless case and no model errors present. In case of model errors, the rank is at most $(n_a)_{\text{init}} + (n_b)_{\text{init}} + 1$. If there are common pole-zero pairs in the system, for the given model orders, degenerations will appear. Their number equals the dimension of the null space of $J(Z)$ minus 1 (to account for the structural degeneration of a transfer function model). The initial estimate of the model order is then given by

$$\hat{n}_a = (n_a)_{\text{init}} + 1 - \dim(\text{null}(J(Z))),\ \hat{n}_b = (n_b)_{\text{init}} + 1 - \dim(\text{null}(J(Z))) \tag{9-28}$$

In practice, only noisy data are available and this simple principle fails to work because the noise and model errors increase the rank of $J(Z)$. To reduce the noise sensitivity, the following extensions are made

- Add a frequency weighting to Eq. (9-27) to get as close as possible to the maximum likelihood weighting. This is exactly the problem that is solved in the starting values generating methods (see Section 7.12.4 on starting values) where $J(Z) \to WJ(Z)$ with the diagonal matrix (7-68).
- The column space of $W_{\text{Re}}J_{\text{re}}(Z)$ (see Section 7.10.3) is weighted with a square root of the column covariance matrix C_{WJ} (7-74) of $W_{\text{Re}}j_{\text{re}}(N_Z)$, with $j(N_Z) = J(Z) - J(Z_0)$ (see (7-67)). The whitened Jacobian is given by

$$W_{\text{Re}}J_{\text{re}}(Z)C_{WJ}^{-1/2} \tag{9-29}$$

For deterministic weighting matrices W it is shown in Rolain et al. (1997) that the "noise" singular values σ_k^2 converge strongly to those of the noiseless matrix + 1:

$$\underset{F \to \infty}{\text{a.s.lim}}\sigma_k^2 = \sigma_{k0}^2 + 1 \tag{9-30}$$

The variance of the noise singular values can be calculated (Rolain et al., 1997), but this requires unacceptably long computation time. Consequently, the dimension of the null space of $W_{\text{Re}}J_{\text{re}}(Z)C_{WJ}^{-1/2}$ is estimated by the number of singular values between zero and one, $\#\{\sigma_k | 0 < \sigma_k \leq 1\}$. This results in an overestimate of the rank of $W_{\text{Re}}J_{\text{re}}(Z)C_{WJ}^{-1/2}$. This is a desirable property because the initial estimate of the model order should be high enough so that the peeling process can be started.

Remarks

(i) In practice, the whitening (9-29) is not explicitly calculated as it is not guaranteed that C_{WJ} is of full rank (see Assumption 7.20(i) or (ii) and Appendix 7.K). Instead, the generalized singular value decomposition (GSVD) is used (Section 13.4.2; Paige, 1986; Bai and Demmel, 1993) to calculate the singular values of the whitened matrix directly from the matrix pair ($W_{\text{Re}}J_{\text{re}}(Z)$, $C_{WJ}^{1/2}$) without calculating the inverse $C_{WJ}^{-1/2}$.

(ii) As we are dealing here with extremely high orders, the numerical conditioning can be cumbersome. In order to avoid these problems, an orthogonal parameterization is used.

9.5.2 "Postidentification" Model Structure Updating

The input to this second step consists of an initial guess of both the model parameters and the model order as obtained, for example, in the coarse step. The methods discussed below are in principle applicable to any estimator, as long as the cost function is absolutely interpretable (this means that it should be possible to predict and calculate its expected value in case there are no model errors). This facilitates validating the intermediate models that are obtained when we reduce the model complexity.

The model reduction:

- Perform a full identification starting from the initial parameters and check whether the resulting model passes the validation test. After a positive validation, the reduction step can be started, otherwise the order should be increased until the validation test is successful.
- Rank the poles, zeros, and pole-zero pairs with respect to their impact on the transfer function in the frequency band of interest. Possible candidates for elimination are poles or zeros that are far away from the modeled frequency band or almost coinciding pole-zero pairs. Next, these roots are eliminated one after another without changing the remaining poles and zeros. Each time it is checked whether the remaining model is still acceptable, using a simplified validation test. For example, the MDL test is a good method to compensate for the reduced order.
- Once it is no longer possible to continue the peeling process without violating the validation test, a new estimate is calculated, starting from the last accepted model. This optimizes the positions of the remaining poles and zeros, again reducing the cost function. Before starting a new peeling step, a full validation is performed (for example, a cost function test and a whiteness test of the residuals).
- Repeating these steps a few times results eventually in a "simple" model that still passes the validation tests. Often, this model is too complex for practical use. If the user can specify an acceptable level of model errors, a further reduction can be made until the user-imposed restrictions are violated.

Remarks

(i) This procedure does not guarantee that the optimal model is found. However, from our experience, it turned out that the in most cases the resulting model is very reasonable.

(ii) Because the procedure is controlled by a series of mechanical rules, it is very suited for a fully automatized model selection. The only required user interaction is the definition of the maximum, acceptable level of model errors.

(iii) In practice, it is never guaranteed that the global minimum of the nonlinear cost function is reached, especially when dealing with more complex systems. Mostly a "good" local minimum is reached. We observed that with the top-down approach we sometimes ended in a better local minimum than the one obtained by a bottom-up approach on the same model order.

9.6 GUIDELINES FOR THE USER

In Table 9-4, we summarize the actions to be followed during the model selection process. This will help the less experienced reader to select a good model. We strongly advise comparing the estimated model with the nonparametric FRF at the end of the process to look for undetected anomalies (e.g., large errors in some frequency bands), strange behavior of the residuals (e.g., strong correlation in a subband), or undesired behavior of the model in frequency bands that were not well excited.

TABLE 9-4 Recommendations for the Model Selection Process

	White Residuals	Colored Residuals
The cost function is too large	■ Best linear approximation ■ Nonlinear distortions present ■ It makes no sense to increase the model order	■ There are still unmodelled dynamics (model errors). Increase the model order to reduce them
The cost function is not significantly different from the expected value	■ This is the ideal situation ■ Best linear model ■ No model errors detectable	■ Good linear approximation ■ Check the noise analysis
The cost function is too small	■ Good linear approximation ■ Check the noise analysis or reduce the model order	■ Good linear approximation ■ Check the noise analysis

9.7 EXERCISES

9.1. Consider a polynomial model:

$$y_0(k) = \sum\nolimits_{p=1}^{5} a_p u^p(k) \tag{9-31}$$

that is identified from a set of measurements $y(k) = y_0(k) + n_y(k)$, with $u(k) = [-N{:}N]/N$ and $n_y(k)$ zero mean iid distributed noise with variance σ_y^2. Set up the least squares estimator for this problem, and calculate the covariance matrix C_a, and the uncertainty of the model output $\sigma_y(u)$.

9.2. Check the previous results on simulations.

9.3. Set up a weighted least squares estimator to identify a parametric transfer function mode $G(\Omega, \theta)$ from measured values $G(\Omega_k) = G_0(\Omega_k) + n_G(k)$ with $n_G(k)$ zero mean iid distributed noise with variance σ_G^2.

9.4. Apply the estimator of Exercise 9.3 to simulation data. Select a second-order system to generate the data. Repeat the simulation 100 times, and compare the mean value with the exact system $G_0(\Omega_k)$.

9.5. Use the results of Exercise 9.4 to calculate the covariance matrix of the transfer function parameter C_θ, and predict from these results the uncertainty on the transfer function $\text{var}(G(\Omega, \theta))$. Compare the predicted variance with the simulation results.

9.6. Calculate $\text{var}(G(\Omega, \theta))$ after putting the nondiagonal terms in the covariance matrix to zero, and discuss your results. Start from the results of Exercise 9.5.

9.7. Make a residual analysis on the results of Exercise 9.4 (use only one simulation), and discuss the results as a function of the selected model order.

9.8. Select a fourth-order system with two well-separated resonances, and identify this system (use, for example, the estimator of Exercise 9.3). Apply the AIC and MDL model selection rules. Do this first for a simulation using a broadband excitation that covers the complete passband of the system, and then repeat the exercise with an excitation that concentrates most of its power on one of both resonances. Analyze the results.

9.9. Set up a simulation, using a Wiener-Hammerstein system as plant. Use an FIR structure for the linear dynamic parts and a third-order polynomial for the static nonlinearities. Scale the excitation signals so that the energy of the second-degree nonlinearity is 10% of that of the linear part, and the third-degree nonlinearity contributes about 1% to the output energy. For this structure determine the underlying linear system and the best linear approximation (see also Chapter 3).

9.10. Identify the best linear approximation of Exercise 9.9, and make a full model validation using the tools developed in this chapter.

9.11. Study the impact of the excitation signal on the quality of the identified model of Exercise 9.10 (cost function, residue analysis, model uncertainty). Analyze 10 realizations in each run.

9.8 APPENDIXES

Appendix 9.A: Properties of the Global Minimum of the Maximum Likelihood Cost Function (Theorem 9.2)

Theorem 9.2 would follow directly from Theorem 17.12 if the frequency domain errors of the time and frequency domain experiment (see Section 7.6) were mixing of order four (infinity). For the frequency domain experiment the frequency domain errors are mixing of order four (Assumption 7.4) but not of order infinity (moments of order higher than $4+\varepsilon$ do not necessarily exist, see Assumption 7.13). Hence, Lemma 17.11 is valid and only the asymptotic normality of $V_{\text{ML}}(\hat{\theta}_{\text{ML}}(Z), Z)$ in Theorem 17.12 remains to be proved. Because after a DFT the noise is not mixing of order four (infinity) (see Section 14.16), all the properties of $V_{\text{ML}}(\hat{\theta}_{\text{ML}}(Z), Z)$ in Theorem 17.12 remain to be proved for the time domain experiment. Fortunately, following the same lines as in the proof of Theorem 7.21 (see Appendix 7.E and Appendix 7.D), the resulting technical difficulties in the proofs can easily be solved using the results of Section 14.16. □

Appendix 9.B: Calculation of Improved Uncertainty Bounds for the Estimated Poles and Zeros

This appendix gives the theoretical foundation of the method explained in Section 9.2.3. In order to keep the application field as general as possible, we emphasize that this result is independent of the specific identification scheme that is used as long as it meets some minimum requirements.

Consider an identification scheme that extracts the model parameters $\theta \in \mathbb{R}^{n_\theta}$ from the measurements $z \in \mathbb{R}^N$ (note that we no longer specify that it is time or frequency domain measurements),

$$\hat{\theta}(z) = \arg\min_{\theta \in \boldsymbol{\theta}_{\text{r}}} V_N(\theta, z) \tag{9-32}$$

$V_N(\theta, z)$ is a well-designed cost function, such that

1. $V''(\tilde{\theta}(z_0), z_0) = C_{\bar{\theta}}^{-1}$, with $\tilde{\theta}(z_0) = \arg\min_{\theta \in \boldsymbol{\theta}_{\text{r}}} \mathcal{E}\{V_N(\theta, z)\}$ and z_0 the noiseless data.
2. $\text{a.s.}\lim_{N \to \infty} \hat{\theta}(z) = \tilde{\theta}(z_0)$, and $\Delta\theta = \hat{\theta}(z) - \tilde{\theta}(z_0) \in AsN(0, C_\theta)$.

Note that these assumptions are met for maximum likelihood and Markov estimators. Even linear least squares estimators can be used if, in a second step, a correct estimation of C_θ is made.

Note that

$$\Delta\theta^T C_{\theta}^{-1}\Delta\theta \in As\chi^2(n_\theta) \tag{9-33}$$

This result is also valid in the presence of model errors. At that moment (9-33) describes, under well-known conditions, the behavior of $\hat{\theta}(z)$ around the parameters that would be obtained in the noiseless case. The p -percentages uncertainty ellipsoids on $\hat{\theta}(z)$ are then given by

$$S_\theta = \{\theta|\Delta\theta^T C_{\theta}^{-1}\Delta\theta \le \chi_p^2(n_\theta)\} \tag{9-34}$$

with $\chi_p^2(n_\theta)$ the p -percentile of a χ^2 distribution with n_θ degrees of freedom. Starting from (9-34), an uncertainty set for the poles/zeros $\rho = \rho(\tilde{\theta}) + \Delta\rho$ around $\rho(\tilde{\theta})$ is defined, where $\rho(\tilde{\theta})$ are the poles and zeros corresponding to the parameters $\tilde{\theta}$.

$$S_\rho = \{\rho|([\Delta\theta(\Delta\rho)]^T C_{\theta}^{-1}[\Delta\theta(\Delta\rho)] \le \chi_p^2(n_\theta))\} \tag{9-35}$$

where $\Delta\theta(\Delta\rho)$ is the parameter variation due to the variation of the poles/zeros. Because the transformation $\theta \to \rho$ is highly nonlinear, linear approximations fail and an alternative method is formulated. In the following, we discuss the individual steps of the method that was explained in Section 9.2.3 in more detail. The basic idea was to move a pole (or zero) and to minimize the impact of this movement on the quality of the fit. First, a reparameterization is described to introduce a pole or zero as a parameter. For generality, a complex pair is considered but the results can be reduced without any problem to the situation of a real pole or zero. Next it is shown how the impact of a movement is minimized. Finally, the criterion $[\Delta\theta(\Delta\rho)]^T C_{\theta}^{-1}[\Delta\theta(\Delta\rho)] \le \chi_p^2(n_\theta)$ in (9-35) is used to accept or reject the movement. Remark that the poles and zeros do not fully determine θ because variation of the gain of the transfer function does not change the pole/zero positions. This additional free parameter will be set such that the impact of a pole or zero movement on the criterion is minimized.

9.B.1 Reparameterization. To focus the ideas, we consider a complex pole pair $\pi_1, \pi_2 = \bar{\pi}_1$ as parameter. The original transfer function $G(\Omega, \theta)$ is partitioned into two subsystems:

$$\underset{\sim}{G}(\Omega, \underset{\sim}{\theta}) = \frac{\pi_1\pi_2}{(\Omega-\pi_1)(\Omega-\pi_2)} = \frac{1}{\underset{\sim}{a}_2\Omega^2+\underset{\sim}{a}_1\Omega+1} \text{ and } \tilde{G}(\Omega, \tilde{\theta}) = \frac{\sum_{i=0}^{n_b} b_i\Omega^i}{\sum_{i=0}^{n_a-2}\tilde{a}_i\Omega^i} \tag{9-36}$$

such that $G(\Omega, \theta) = \underset{\sim}{G}(\Omega, \underset{\sim}{\theta})\tilde{G}(\Omega, \tilde{\theta})$.

Remarks

$\tilde{G}(\Omega, \tilde{\theta})$ includes the gain variations that were mentioned before.

Note that there exists a bilinear relationship between the old and the new parameterization:

$$\theta = T(\underset{\sim}{\theta})\tilde{\theta} \tag{9-37}$$

with

$$T(\underset{\sim}{\theta}) = \begin{array}{c} \left[\begin{array}{c:cccccc} I_{n_b+1} & 0 & 0 & \dots & 0 & 0 & 0 \\ \hdashline 0 & 1 & 0 & \dots & 0 & 0 & 0 \\ 0 & \underset{\sim}{a}_1 & 1 & \dots & 0 & 0 & 0 \\ 0 & \underset{\sim}{a}_2 & \underset{\sim}{a}_1 & \dots & 0 & 0 & 0 \\ 0 & 0 & \underset{\sim}{a}_2 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & \underset{\sim}{a}_2 & \underset{\sim}{a}_1 & 1 \\ 0 & 0 & 0 & \dots & 0 & \underset{\sim}{a}_2 & \underset{\sim}{a}_1 \\ 0 & 0 & 0 & \dots & 0 & 0 & \underset{\sim}{a}_2 \end{array}\right] \\ \begin{array}{cc} \longleftrightarrow & \longleftrightarrow \\ n_b+1 & n_a-1 \end{array} \end{array} \begin{array}{l} \updownarrow n_b+1 \\ \\ \\ \\ \\ \Big\updownarrow n_a+1 \\ \\ \\ \\ \\ \\ \end{array}$$

(9-38)

A similar transformation can be set up when a complex zero pair or a real pole or zero is selected as parameter.

9.B.2 Minimizing the Impact of a Movement, Accepting or Rejecting a Move. The impact of pole (zero) movement on the cost function is minimized by changing the remaining parameters $\tilde{\theta}$. This is done by minimizing

$$[\Delta\theta(\Delta\rho)]^T C_{\tilde{\theta}}^{-1}[\Delta\theta(\Delta\rho)] = [T(\underset{\sim}{\theta})\tilde{\theta} - \tilde{\theta}(z_0)]^T C_{\tilde{\theta}}^{-1}[T(\underset{\sim}{\theta})\tilde{\theta} - \tilde{\theta}(z_0)] \tag{9-39}$$

with respect to $\tilde{\theta}$. The solution is found by solving

$$T^T(\underset{\sim}{\theta})C_{\tilde{\theta}}^{-1}T(\underset{\sim}{\theta})\tilde{\theta} = T^T(\underset{\sim}{\theta})C_{\tilde{\theta}}^{-1}\tilde{\theta}(z_0) \tag{9-40}$$

Plugging this solution back into (9-35) allows us to verify whether $\rho \in S_\rho$, using relation (9-35). The corresponding pole/zero positions can be calculated from this "improved" parameter set and be used to construct all uncertainty regions at once, instead of looking for the extreme positions for all pole-zero pairs. In practice, it might be necessary to repeat the whole process for a few poles/zeros in order to get a precise description of the uncertainty regions as there is no guarantee that all pole/zeros reach their extreme positions at the same time.

Remark. In practice, the following approximations are made to evaluate the solutions: $\tilde{\theta}(z_0) \to \hat{\theta}(z)$ and $C_{\tilde{\theta}}^{-1} \to V''(\hat{\theta}(z), z)$, as the exact values are unknown.

Appendix 9.C: Proof of Theorem 9.5

Consider a system belonging to the set S (see Definition 3.5), excited with a random multisine $x_F \in \mathbb{E}_F$. If no unmodeled dynamics are present $(G_E(\Omega_k) = 0)$, then under the assumptions of Section 7.6.5 (frequency domain experiment with $\sigma_U^2(k) = 0$)

$$\begin{aligned} \hat{R}_{\varepsilon\varepsilon}(m) &= O_p(F^{-1/2}) \qquad m \neq 0 \\ \hat{R}_{\varepsilon\varepsilon}(0) &= \frac{1}{F}\sum\nolimits_{k=1}^{F} \frac{\sigma_q^2(k)}{\sigma_G^2(k)} + O_p(F^{-1/2}) \end{aligned} \tag{9-41}$$

Proof. In this proof we use the more compact notation $G_{vk} = G_v(\Omega_k, \hat{\theta}(Z))$. From (9-20) it follows that in the absence of model errors $(G_E(\Omega_k) = 0)$

$$\hat{R}_{\varepsilon\varepsilon}(m) = \frac{1}{F-m}\sum_{k=1}^{F-m} \frac{(q_k - G_{vk})(\bar{q}_{k+m} - \bar{G}_{v(k+m)})}{\sigma_G(k)\sigma_G(k+m)} \tag{9-42}$$

or

$$\begin{aligned} \hat{R}_{\varepsilon\varepsilon}(m) = {} & \frac{1}{F-m}\sum_{k=1}^{F-m} \frac{q_k\bar{q}_{k+m}}{\sigma_G(k)\sigma_G(k+m)} + \frac{1}{F-m}\sum_{k=1}^{F-m} \frac{G_{vk}\bar{G}_{v(k+m)}}{\sigma_G(k)\sigma_G(k+m)} \\ & - \frac{1}{F-m}\sum_{k=1}^{F-m} \frac{q_k\bar{G}_{v(k+m)}}{\sigma_G(k)\sigma_G(k+m)} - \frac{1}{F-m}\sum_{k=1}^{F-m} \frac{q_{k+m}\bar{G}_{vk}}{\sigma_G(k)\sigma_G(k+m)} \end{aligned} \tag{9-43}$$

Each of these terms converges for $m \neq 0$ to zero, at least, as an $O_p(F^{-1/2})$. Essential in the proof is that m does not tend to F: we require that $F - m = O(F)$ so that the results are also valid for a constant fraction $m = \alpha F$ with $\alpha < 1$.

(i) $s_1 = \frac{1}{F-m}\sum\nolimits_{k=1}^{F-m} \frac{q_k\bar{q}_{k+m}}{\sigma_G(k)\sigma_G(k+m)}$ is an $O_{\text{m.s.}}((F-m)^{-1/2})$

Using $\text{var}(s_1) \leq \mathcal{E}\{s_1^2\}$, we show the mean square convergence of s_1

$$\mathcal{E}\{|s_1|^2\} = \frac{1}{(F-m)^2}\sum_{k=1}^{F-m}\sum_{l=1}^{F-m} \mathcal{E}\left\{\frac{q_k\bar{q}_{k+m}}{\sigma_G(k)\sigma_G(k+m)} \frac{\overline{q_l\bar{q}_{l+m}}}{\sigma_G(l)\sigma_G(l+m)}\right\} \tag{9-44}$$

By careful examination of the right side and using the results of Theorem 3.9 and the noise assumptions of a frequency domain experiment (see Section 7.6.5), it can be shown that the double sum contains the following contributions:

- $k \neq l$: $O((F-m)^2)$ contributions of $O(F^{-2})$
- $k = l$: $O((F-m))$ contributions of $O(F^0)$

Hence, $\mathcal{E}\{|s_1|^2\} = O((F-m)^{-1})$ and $s_1 = O_{\text{m.s.}}((F-m)^{-1/2})$.

(ii) $s_2 = \frac{1}{F-m}\sum\nolimits_{k=1}^{F-m} \frac{[G_{vk}\bar{G}_{v(k+m)}]}{\sigma_G(k)\sigma_G(k+m)}$ is an $O_p(F^{-1})$.

$G(\Omega_k, \hat{\theta}(Z))$ is a consistent estimate obtained under the standard conditions for output error estimates, which is a special case of the errors-in-variables formulation. For this class of estimators it is known that $\hat{\theta}(Z) - \theta_*$ is an $O_p(F^{-1/2})$ (Theorem 7.21, properties 2 and 5). Applying the mean value theorem to $G(\Omega_k, \hat{\theta}(Z))$ gives

$$G_{vk} = \frac{\partial G(\Omega_k, \theta)}{\partial \widehat{\theta}}(\hat{\theta}(Z) - \theta_*) \tag{9-45}$$

with $\widehat{\theta} = (1-t)\theta_* + t\hat{\theta}(Z)$, and $t \in [0, 1]$. Under Assumption 7.8, the derivatives of G are uniformly bounded, so that G_{vk} is $O_p(F^{-1/2})$. Hence, s_2 is an $O_p(F^{-1})$.

(iii) $s_3 = \frac{1}{F-m}\sum_{k=1}^{F-m}\frac{q_k\bar{G}_{v(k+m)}}{\sigma_G(k)\sigma_G(k+m)}$ or $\frac{1}{F-m}\sum_{k=1}^{F-m}\frac{q_{k+m}\bar{G}_{vk}}{\sigma_G(k)\sigma_G(k+m)}$ are an $O_p(F^{-1/2})$.

We prove the result for the first sum; that of the second sum follows exactly the same lines. Taking the absolute value of s_3 gives

$$|s_3| \le \frac{1}{F-m}\sum\nolimits_{k=1}^{F-m}\frac{|q_k||G_{v(k+m)}|}{\sigma_G(k)\sigma_G(k+m)} \le \frac{O_p(F^{-1/2})}{F-m}\sum\nolimits_{k=1}^{F-m}|q_k| \tag{9-46}$$

Because q_k is an $O_{\text{m.s.}}(F^0)$, the conclusion follows directly. □

Appendix 9.D: Calculation of the Sample Correlation at Lag Zero (Proof of Theorem 9.5)

$$\hat{R}_{\varepsilon\varepsilon}(0) = \frac{1}{F}\sum_{k=1}^{F}\frac{|q_k|^2}{\sigma_G^2(k)} + O_p(F^{-1/2})$$

Proof. Putting $m = 0$ in (9-43) gives

$$\hat{R}_{\varepsilon\varepsilon}(0) = \frac{1}{F}\sum_{k=1}^{F}\frac{|q_k|^2}{\sigma_G^2(k)} + \frac{1}{F}\sum_{k=1}^{F}\frac{|G_{vk}|^2}{\sigma_G^2(k)} - \frac{2}{F}\text{Re}\left(\sum_{k=1}^{F}\frac{q_k\bar{G}_{vk}}{\sigma_G^2(k)}\right) \tag{9-47}$$

The proof of convergence of the second and third terms at the right-hand side in (9-47) is similar to that of the previous appendix. The first sum can be written as:

$$s_1 = \frac{1}{F}\sum_{k=1}^{F}\frac{|q_k|^2}{\sigma_G^2(k)} = \frac{1}{F}\sum_{k=1}^{F}\frac{\sigma_q^2(k)}{\sigma_G^2(k)} + \frac{1}{F}\sum_{k=1}^{F}\frac{|q_k|^2 - \sigma_q^2(k)}{\sigma_G^2(k)} \tag{9-48}$$

From Theorem 3.9(iv) and the noise assumptions of a frequency domain experiment (see Section 7.6), it follows directly that the last sum in (9-48) is of order $O_{\text{m.s.}}(F^{-1/2})$, which concludes the proof. □

Appendix 9.E: Study of the Sample Correlation at Lag One (Proof of Theorem 9.7)

$$\hat{R}_{\varepsilon\varepsilon}(1) = \frac{1}{F-1}\sum_{k=1}^{F-1}\frac{G_E(\Omega_k)\bar{G}_E(\Omega_{k+1})}{\sigma_G(k)\sigma_G(k+1)} + O_p(F^{-1/2})$$

Proof. In this proof we use the more compact notation $G_{Ek} = G_E(\Omega_k)$.

$$\begin{aligned}\hat{R}_{\varepsilon\varepsilon}(1) &= \frac{1}{F-1}\sum_{k=1}^{F-1}\frac{[G_{Ek}+q_k-G_{vk}][\bar{G}_{E(k+1)}+\bar{q}_{k+1}-\bar{G}_{v(k+1)}]}{\sigma_G(k)\sigma_G(k+1)} \\ &= \frac{1}{F-1}\sum_{k=1}^{F-1}\frac{[G_{Ek}-q^{\dagger}{}_k][G_{E(k+1)}-\overline{q^{\dagger}}{}_{k+1}]}{\sigma_G(k)\sigma_G(k+1)}\end{aligned} \qquad (9\text{-}49)$$

with $q^{\dagger}{}_k = G_{vk}-q(k)$. Compared with Appendix 9.C, a new term G_{Ek} appeared, raising new contributions of the type

$$\frac{1}{F-1}\sum_{k=1}^{F-1}\frac{G_{Ek}\overline{q^{\dagger}}{}_{k+1}}{\sigma_G(k)\sigma_G(k+1)} = \frac{1}{F-1}\sum_{k=1}^{F-1}\frac{G_E\bar{q}_{k+1}}{\sigma_G(k)\sigma_G(k+1)} - \frac{1}{F-1}\sum_{k=1}^{F-1}\frac{G_{Ek}\bar{G}_{v(k+1)}}{\sigma_G(k)\sigma_G(k+1)}$$

The last sum at the right-hand side is an $O_{\mathrm{p}}(F^{-1/2})$ because $\bar{G}_{v(k+1)}$ is an $O_{\mathrm{p}}(F^{-1/2})$. Using $\mathrm{var}(x) \le \mathcal{E}\{x^2\}$, the mean square convergence of the first sum is shown

$$\frac{1}{(F-1)^2}\sum_{k=1}^{F-1}\sum_{l=1}^{F-1}\frac{G_{Ek}\bar{G}_{El}\mathcal{E}\{\bar{q}_{k+1}q_{l+1}\}}{\sigma_G(k)\sigma_G(k+1)\sigma_G(l)\sigma_G(l+1)} \le O(F^{-1}) \qquad (9\text{-}50)$$

The last inequality follows from the fact that $G_{Ek}/\sigma_G(k)$ is uniformly bounded by assumption and because $\mathcal{E}\{\bar{q}_{k+1}q_{l+1}\} = O(F^{-1})$ for $k \ne l$, and $\mathcal{E}\{|q_{k+1}|^2\} = \sigma_q^2(k)$ for $k = l$. □

Appendix 9.F: Calculation of the Variance of the Sample Correlation

In this section the variance of $\hat{R}_{\varepsilon\varepsilon}(m)$ is calculated under Assumptions 1 and 2 and putting $G_v = 0$. The variance of $\hat{R}_{\varepsilon\varepsilon}(m)$ will be calculated assuming that there are no unmodeled dynamics left ($G_E = 0$). Hence, this result can be used to check whether or not this hypothesis is valid. Because $\mathcal{E}\{\hat{R}_{\varepsilon\varepsilon}(m)\} = O((F-m)^{-1})$ for $m \ne 0$, we have $\mathrm{var}(\hat{R}_{\varepsilon\varepsilon}(m)) \approx \mathcal{E}\{|\hat{R}_{\varepsilon\varepsilon}(m)|^2\}$ so that

$$\mathrm{var}(\hat{R}_{\varepsilon\varepsilon}(m)) \approx \frac{1}{(F-m)^2}\mathcal{E}\left\{\sum_{k=1}^{F-m}\sum_{l=1}^{F-m}\frac{q_k\bar{q}_{k+m}\bar{q}_l q_{l+m}}{\sigma_G(k)\sigma_G(k+m)\sigma_G(l)\sigma_G(l+m)}\right\} \qquad (9\text{-}51)$$

for $m \ne 0$ and $F \to \infty$. Replacing $q_k = N_G(k) + G_S(j\omega_k)$ and using the properties of $G_S(j\omega_k)$, the following asymptotic expression for $F-m \to \infty$ is found:

$$\begin{aligned}\mathrm{var}(\hat{R}_{\varepsilon\varepsilon}(m)) \approx \frac{1}{F-m} + \frac{1}{(F-m)^2}\Bigg[&\sum_{k=1}^{F-m}\frac{|G_S(j\omega_k)|^2|G_S(j\omega_{k+m})|^2}{\sigma_G^2(k)\sigma_G^2(k+m)} + \\ &\sum_{k=1}^{F-m}\frac{|G_S(j\omega_k)|^2}{\sigma_G^2(k)} + \sum_{k=1}^{F-m}\frac{|G_S(j\omega_{k+m})|^2}{\sigma_G^2(k+m)}\Bigg]\end{aligned} \qquad (9\text{-}52)$$

A detailed analysis shows that this expression is also valid for $m = 0$.

Remarks

(i) Expression (9-52) cannot be calculated directly because only q is available and not G_S. Replacing $G_S(j\omega_k)$ by q results in an overestimate of the uncertainty bounds. This can be compensated by substituting $|G_S(j\omega_k)|^2/\sigma_G^2(k)$ in the variance expressions by $|q_k|^2/\sigma_G^2(k) - 1$.

(ii) During the validation tests, graphical representations of the amplitude of $\hat{R}_{\varepsilon\varepsilon}(m)$ are used. Hence, the complex variance should be transferred into a bound on the amplitude. Because the variance dominates the bias error, it follows that $\hat{R}_{\varepsilon\varepsilon}(m)$ is asymptotically zero mean complex normally distributed (The real and imaginary parts of the individual contributions to the $\hat{R}_{\varepsilon\varepsilon}(m)$ are uncorrelated and have equal variance). So the amplitude is chi-squared distributed with two degrees of freedom. For example, the 95% level is given by the bound $95\% = \sqrt{3}\,\mathrm{std}(\hat{R}_{\varepsilon\varepsilon}(m))$.

(iii) If G_v is not zero in the previous calculations, the variance expression is still valid but only a weaker statement about convergence in distribution can be made because the expected value $\mathcal{E}\{G_{vk}G_{vl}\}$ is not guaranteed to exist. The sample correlation $\hat{R}_{\varepsilon\varepsilon}(m)$ converges in distribution to a random variable with zero mean and variance $\mathrm{var}(\hat{R}_{\varepsilon\varepsilon}(m))$.

(iv) In practice, the sample correlation is calculated using not the exact variances but the sample variances (see (8-10), Section 8.3.1). This also changes $\mathrm{var}(\hat{R}_{\varepsilon\varepsilon}(m))$ to $\mathrm{var}(\hat{R}_{\hat{\varepsilon}\hat{\varepsilon}}(m))$. So an additional factor $c_k(\theta)$ (8-17) appears in the calculations, and the variances are scaled: $\mathrm{var}(\hat{R}_{\hat{\varepsilon}\hat{\varepsilon}}(m)) = ((M-1)/(M-2))^2\mathrm{var}(\hat{R}_{\varepsilon\varepsilon}(m))$ if $m \neq 0$ (for $m = 0$, the cost function expression can be used).

(v) In Section 8.7 it is shown how to identify the best linear approximation for systems that are disturbed by nonlinear distortions. In that case not only the measurement noise but also the stochastic nonlinearities are considered as disturbing noise. The proposed procedure accounts for both effects during the extraction of the (co)variances from the raw data. As a consequence, the presence of nonlinear distortions will not be detected during the validation tests because under these conditions, it just acts as an additional noise source. In this case $\mathrm{var}(\hat{R}_{\varepsilon\varepsilon}(m))$ (9-52) reduces to $1/(F-m)$.

10

Basic Choices in System Identification

Abstract: In this chapter we discuss three fundamental questions regarding very basic aspects of the identification process. The first one deals with the signal assumption that is made to reconstruct the intersample behavior. Two possibilities are considered: zero-order-hold reconstruction and band-limited reconstruction. The second question looks into the selection of the excitation signal, dealing mainly with the choice between periodic or nonperiodic excitations. The choice between time domain or frequency domain identification is the topic of the third part of this chapter.

10.1 INTRODUCTION

At the beginning of the 1970s, the identification field gave a quite disordered impression. Many methods were proposed to identify linear dynamic systems. However, due to the lack of integration, the whole field looked more like a "bag of tricks" than a consistent scientific discipline. In the 1980s the field became well ordered, pointing out the relations and the differences between the widely scattered methods (Eykhoff, 1974; Ljung, 1999; Norton, 1986; Söderström and Stoica, 1989). The major part of this work was done in the time domain, leading to a complete dominance of these methods over frequency domain identification techniques. Since then, new methods have popped up, some of them being applicable to time domain or frequency domain identification (for example, subspace methods: McKelvey et al., 1996; Van Overschee and De Moor, 1994; Verhaegen, 1994; Viberg et al., 1997), others being completely focused on frequency domain identification. This does not lead us back to chaos, because the clear insight of the 1980s still applies to the new situations. Rather, these reviving approaches just complete the puzzle, making the picture better balanced. As explained in Chapter 1, the identification process is mainly determined by the answers to three basic questions: (i) what data will be used (experiment design), (ii) what model will be used (model selection), and (iii) how will the model be matched to the data (choice of a cost function)? These questions have to be answered, independent of the user's intention to work in the time domain or in the frequency domain. Perpendicular to these questions, two other important choices have to be made. (i) Choice of the intersample behavior. As identification starts mostly from discrete data, an assumption is needed to make precise what is going on between

System Identification: A Frequency Domain Approach. Rik Pintelon and Johan Schoukens
ISBN: 0-780-36000-1

the samples. This choice has a major impact on the experimental setup, the model choice, and the selection of the cost function. (ii) Periodic versus arbitrary excitation. This question is not linked to the choice between time or frequency domain identification. Periodic excitations offer significant advantages whenever they can be applied, and this is almost independent of the domain (time or frequency) that is selected to process the measurements.

In this chapter we analyze the consequences of these two basic choices that have to be made. Too often, no conscious selection is made, although the consequences maintain their full impact on the users' result. Therefore, it is important to make a well-considered selection, at the beginning of the process, to avoid undesired surprises at the end. The discussions are made without any prejudices to a specific application, so that for some fields the risk exists that some parts of this chapter are not relevant. For that reason we separated the objective facts, which are true without any discussion, from the interpretation of these facts, where we look for their (un)importance for specific fields. The latter is much more subjective as it is strongly influenced by our personal experiences. Consequently, we strongly advise the readers to test these sections according to their own experiences and to draw their own conclusions.

Finally we also deal with the choice between time and frequency domain methods. Too often this selection is presented as conflicting options. To address this, we first point to the (sometimes even unexpected) equivalences between time and frequency domain methods. Eventually, we zoom in on the differences so that at the end of the chapter the user should be able to select the most dedicated method for his problem.

10.2 INTERSAMPLE ASSUMPTIONS: FACTS

Nowadays, almost every identification scheme is applied to sampled data. In a first step, the continuous-time signals are sampled in time and stored in the computer. The discrete samples do not carry all information contained in the original signals, unless additional assumptions are made on the intersample behavior. What is going on between the samples? As we did not measure this information, we do not know. We can only make a guess, formalized as an assumption, and hope that in practice the real behavior is close to the assumed one. Two assumptions are very popular. The zero-order-hold (ZOH) assumption considers the signal to be constant between consecutive samples, while under the band-limited (BL) assumption we suppose that the power spectrum of the signal is zero above half the sampling frequency $f_{\max} < f_s/2$. We discuss, in detail, the impact of this choice on the experimental setup, the model, and the identification process. We also analyze what happens if the wrong assumption is applied to a given set of data, for example, band-limited data are processed under the ZOH assumption.

10.2.1 Formal Description of the ZOH and BL Assumptions

Consider a discrete-time signal $u_d(kT_s)$. Notice that in this section we will sometimes, explicitly, mention the sampling period T_s. This is to indicate that these samples are generated at the time instances kT_s and that the spectrum of this signal is periodic with period $f_s = 1/T_s$ as shown in Figure 10-1.

Assumption 10.1 (Zero-Order-Hold Assumption): The ZOH reconstruction of a discrete-time signal $u_d(kT_s)$ is

$$u_{\text{ZOH}}(t) = \sum_{k=-\infty}^{\infty} u_d(kT_s)\text{zoh}(t-kT_s) \text{ with } \text{zoh}(t) = \begin{matrix} 1 & 0 \le t < T_s \\ 0 & \text{elsewhere} \end{matrix} \quad (10\text{-}1)$$

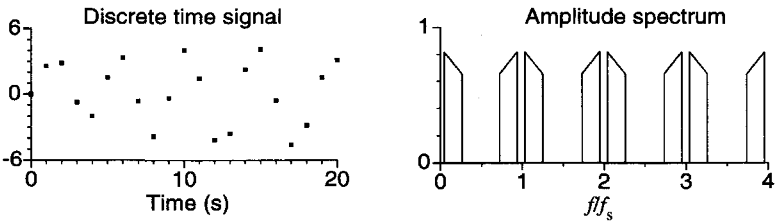


Figure 10-1. A discrete-time signal and part of its periodic spectrum.

The spectrum, after a ZOH reconstruction, is (see Exercise 10.1)

$$X_{\mathrm{ZOH}}(j\omega) = U_{\mathrm{d}}(j\omega)\mathrm{ZOH}(\omega/\omega_{\mathrm{s}}) \text{ with } \mathrm{ZOH}(x) = T_{\mathrm{s}}\frac{\sin \pi x}{\pi x}e^{-j\pi x} \tag{10-2}$$

Assumption 10.2 (Band-Limited Assumption): A signal $u(t)$ with power spectrum $\Phi(\omega)$ is called band-limited (BL) if there exists a value ω_{max} such that $\Phi(\omega) = 0$ for $\forall\ |\omega| > \omega_{\mathrm{max}}$.

In Figure 10-2(a), both reconstructions are illustrated. The reconstructed signals and their spectra differ considerably. The steps in a ZOH reconstruction create high-frequency components far above the sampling frequency, but this is not the case for the BL reconstruction.

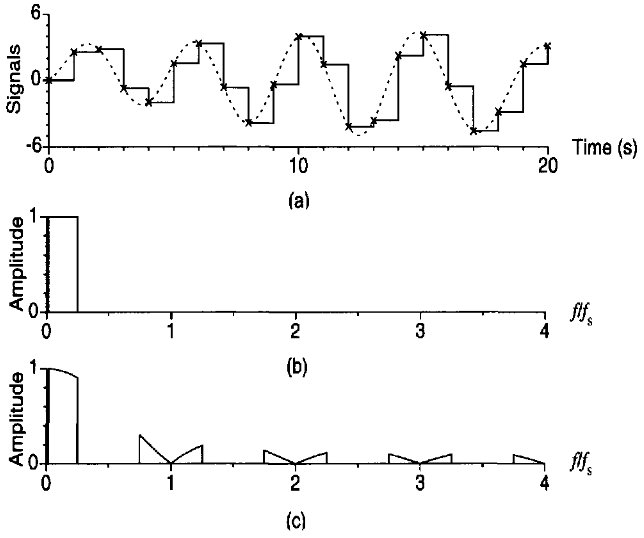


Figure 10-2. Reconstruction of a discrete-time signal under the BL and ZOH assumptions. (a) The time signals: x samples, --- BL, —— ZOH; (b) spectrum of the BL reconstruction; and (c) part of the spectrum of the ZOH reconstruction.

10.2.2 Relation between the Intersample Behavior and the Model

10.2.2.1 ZOH Assumption. In this setup, the transfer function between the discrete-time signal $u_d(k)$ and the output $y(k)$ is measured (Figure 10-3). This means that besides the linear system itself, the actuator and measurement channel are also modeled as indicated by the gray area in Figure 10-3. Assume, for simplicity, that the disturbing noise sources are zero. The overall continuous-time system transfer function $G_c(j\omega)$ comprises the actuator $A(j\omega)$, the process $G(j\omega)$, and the data acquisition $G_y(j\omega)$ transfer function. Under the ZOH setup, it is modeled as a discrete-time system with impulse response $g_d(k)$ that links the discrete-time input $u_d(k)$ to the discrete-time output $y(k)$ (Ljung, 1999):

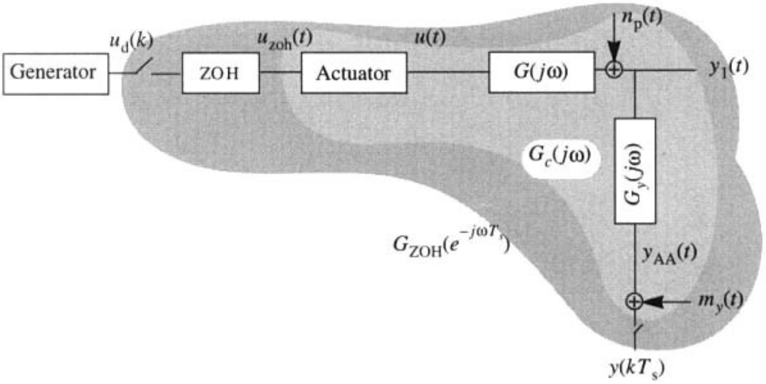


Figure 10-3. Basic setup for the ZOH assumption, interpretation of the continuous-time (G_c) and the equivalent discrete-time (G_{ZOH}) system.

$$y(m) = \sum\nolimits_{k=1}^{\infty} g_d(k)u_d(m-k) \text{ with } g_{ZOH}(k) = \int_{(k-1)T_s}^{kT_s} g_c(\tau)d\tau \qquad (10\text{-}3)$$

where $g_c(\tau)$ is the impulse response of the continuous-time system, between $u_{ZOH}(t)$ and the continuous-time output $y_{AA}(t)$.

$G_{ZOH}(z^{-1})$ and $G_c(s)$ are linked by the step invariant transformation for ZOH excitations (Middleton and Goodwin, 1990; see also Example 5.2):

$$G_{ZOH}(z^{-1}) = (1-z^{-1})Z\{G_c(s)/s\} \qquad (10\text{-}4)$$

If $f_s/2 > \mathrm{Im(poles)}/(2\pi)$, then the original continuous-time parameters can be retrieved from the discrete-time model, using an inverse transformation (Ljung, 1999). The poles are found using the impulse invariant transformation, but the transformation of the zeros is much more complex (Åström et al., 1984).

The final relation between the discrete input spectrum and the spectrum of the sampled output signal is (see Exercise 10.2)

$$Y(e^{j\omega T_s}) = U_d(e^{j\omega T_s}) \sum_{k=-\infty}^{\infty} G(j\Omega_k)A(j\Omega_k)G_y(j\Omega_k)\mathrm{ZOH}(\Omega_k/\omega_s)\Big|_{\Omega_k = \omega - k\omega_s} \qquad (10\text{-}5)$$

for $|\omega| < \omega_s/2$. The sum in (10-5) is due to the repeated spectra, as they appear in the spectrum of $u_{\text{ZOH}}(k)$ (see Figure 10-2).

10.2.2.2 BL Assumption. The BL setup is given in Figure 10-4. Only the gray box is directly involved in the identification process. Starting from the spectra of the continuous-time signals $U_1(j\omega) = F\{u_1(t)\}$ and $Y_1(j\omega) = F\{y_1(t)\}$, it can easily be shown that the following relations exist between the spectra of the sampled signals $U(e^{j\omega T_s}) = F\{u(k)\}$, $Y(e^{j\omega T_s}) = F\{u(k)\}$:

$$G_{\text{BL}}(j\omega) = \frac{Y(e^{j\omega T_s})}{U(e^{j\omega T_s})} = \frac{G_y(j\omega)Y_1(j\omega) + \sum_{k=-\infty, k\neq 0}^{\infty} G_y(j\Omega_k)Y_1(j\Omega_k)\big|_{\Omega_k = \omega - k\omega_s}}{G_u(j\omega)U_1(j\omega) + \sum_{k=-\infty, k\neq 0}^{\infty} G_u(j\Omega_k)U_1(j\Omega_k)\big|_{\Omega_k = \omega - k\omega_s}} \tag{10-6}$$

(see Exercise 10.3).

The sum terms in this expression are due to the alias effect of the sampling process (see Section 2.2.1). If the measurement channels are provided with good antialias filters with a cutoff frequency below $\omega_s/2$, the band-limited assumption holds and Eq. (10-6) becomes

$$G_{\text{BL}}(j\omega) = \frac{G_y(j\omega)Y_1(\omega)}{G_u(j\omega)U_1(\omega)} = G(j\omega)\frac{G_y(j\omega)}{G_u(j\omega)} \text{ for } |\omega| < \omega_s/2 \tag{10-7}$$

which shows that

$$G_{\text{BL}}(j\omega) = G(j\omega) \text{ for } |\omega| < \omega_s/2 \tag{10-8}$$

if $G_y = G_u$ in this frequency band.

Note that the model in Eq. (10-8) corresponds to the continuous-time representation of the process $G(j\omega)$. So the model equations are given by differential equations in the time domain and algebraic equations in the frequency domain. The latter are given by the transfer function model formulated in the Laplace domain ($s = j\omega$ is used as frequency variable in the transfer function).

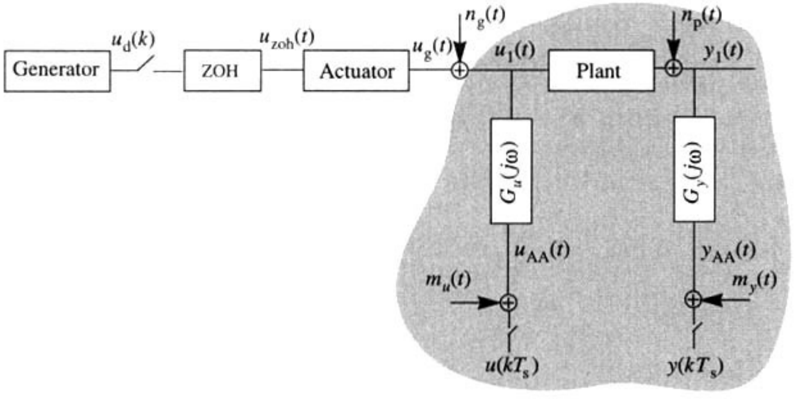


Figure 10-4. BL measurement setup.

10.2.2.3 Conclusion. The ZOH assumption imposes an experimental condition on the excitation signal; it is generated from a discrete-time sequence using a piecewise constant interpolation. Under these conditions, a discrete-time model is obtained between the discrete-time input and the sampled output.

The BL assumption is a condition on the observation of the signals and does not impose constraints on the applied excitation (e.g., BL observations of ZOH signals can be made). It results in a continuous-time model of the plant in the observed frequency band.

10.2.3 Mixing the Intersample Behavior and the Model

In the previous section it was found that the ZOH assumption leads, naturally, to a discrete-time model, and the BL assumption results in a continuous-time model. If a discrete-time model is combined with non-ZOH inputs, or continuous-time models are identified under the ZOH assumption (without applying antialias filters so that the BL condition is violated), systematic errors appear. In practice, these wrong combinations are often made (consciously or unconsciously). Hence, it is important to understand the impact of violating the basic assumption.

10.2.3.1 Violation of the ZOH Assumption. Consider the generalized setup of Figure 10-5 (the disturbing noise sources are not shown for simplicity). Instead of using the known input $u_d(k)$, a discrete-time model is built between the measured input $u(k)$ and output $y(k)$. Because the excitation signal passed through a first subsystem $L(j\omega)$, the signal $u(t)$ is no longer ZOH. The modeled transfer function is found directly, applying Eq. (10-4) twice

$$G_L(z^{-1}) = \frac{(1-z^{-1})Z\{L(s)G(s)/s\}}{(1-z^{-1})Z\{L(s)/s\}} \tag{10-9}$$

(see Example 5.3). Assuming that the sums converge, $G_L(e^{-j\omega T_s})$ is also given by

$$G_L(e^{-j\omega T_s}) = \frac{\sum_{k=-\infty}^{\infty} L(j\Omega_k)G(j\Omega_k)\text{ZOH}(\Omega_k/\omega_s)\big|_{\Omega_k = \omega - k\omega_s}}{\sum_{k=-\infty}^{\infty} L(j\Omega_k)\text{ZOH}(\Omega_k/\omega_s)\big|_{\Omega_k = \omega - k\omega_s}} \tag{10-10}$$

for $|\omega| < \omega_s/2$. Note that the result is still independent of the input $U_d(e^{j\omega T_s})$, but it depends on the preceding system $L(s)$. If the same subsystem $G(j\omega)$ is measured in another environment $(L(j\omega) \rightarrow \hat{L}(j\omega))$, the resulting model will change. Under these conditions, the model is no longer independent of the measurement environment, and the results cannot be transferred from one setup to the other. However, as long as the setup is not changed, a good description of the measurements is given. If $L(j\omega) = 1$, $\forall\omega$, the original ZOH setup is retrieved (see Exercise 10.4). If L is chosen as a perfect reconstruction filter, $L(j\omega) = 1$ for

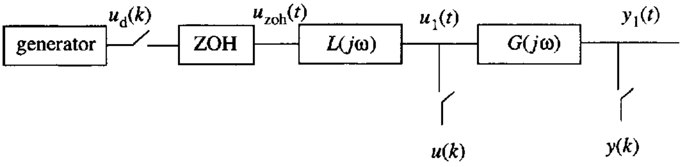


Figure 10-5. Violation of the ZOH assumption.

$|\omega| < \omega_s/2$ and $L(j\omega) = 0$ elsewhere, the BL setup is retrieved and $G_L(e^{-j\omega T_s})$ (10-10) equals $G(j\omega)$ instead of $G_{\text{ZOH}}(e^{-j\omega T_s})$. Assume now that a discrete-time model $G(z^{-1}, \theta)$ is fitted to these BL measurements such that $G(e^{-j\omega T_s}, \theta) = G(j\omega)$ for $|\omega| < \omega_s/2$. If the model $G(z^{-1}, \theta)$, based on BL measurements, is used later under ZOH conditions, the ratio $\gamma(\omega)$ between the predicted and the actual output is

$$\begin{aligned}\gamma(\omega) &= \frac{G(e^{-j\omega T_s}, \theta)U_1(e^{j\omega T_s})}{G_{\text{ZOH}}(e^{-j\omega T_s})U_1(e^{j\omega T_s})}\\ &= \frac{G(j\omega)}{G_{\text{ZOH}}(e^{-j\omega T_s})}\\ &= \frac{1}{\text{ZOH}(\omega/\omega_s) + \sum_{k=-\infty, k\neq 0}^{\infty}\frac{G(j\omega - kj\omega_s)}{G(j\omega)}\text{ZOH}(\omega/\omega_s - k)}\end{aligned} \tag{10-11}$$

for $|\omega| < \omega_s/2$. $\gamma(\omega)$ will be close to 1 only if in the frequency band of interest $|G(j\omega - kj\omega_s)| \ll |G(j\omega)|$, $k \neq 0$ and $|\omega/\omega_s| \ll 1$.

10.2.3.2 Applying Continuous Models to ZOH Measurements. The second possibility is to fit a continuous-time model $G(s, \theta)$ to the ZOH measurements such that $G(j\omega, \theta) = G_{\text{ZOH}}(e^{-j\omega T_s})$ for $|\omega| < \omega_s/2$. If the model $G(s, \theta)$, based on ZOH measurements, is used later under BL conditions, the ratio between the predicted and the actual output is

$$\frac{G(s, \theta)U_1(j\omega)}{G(j\omega)U_1(j\omega)} = \frac{G_{\text{ZOH}}(e^{-j\omega T_s})}{G(j\omega)} = \frac{1}{\gamma(\omega)} \tag{10-12}$$

Again, the same conclusions can be drawn.

10.2.3.3 Conclusion. In the previous sections it was shown that a continuous-time system can be modeled, without systematic errors, by selecting the proper experimental conditions so that the assumptions that describe the intersample behavior are met. If they are violated, it is still possible to get a good model for the observations, but this model is no longer independent of the measurement environment. The intersample behavior becomes an intrinsic part of the model.

10.3 THE INTERSAMPLE ASSUMPTION: APPRECIATION OF THE FACTS

In Sections 10.2.1 to 10.2.3 we have given objective facts. These are true without any discussion. However, the importance of these facts can be very different from one application field to another. As mentioned before, it is the responsibility of the reader to judge their impact on his application. Here, we give some thoughts on possible implications, but it should be clear that these are influenced by personal experiences. As such, we advise the reader to consider them critically.

10.3.1 Intended Use of the Model

Although from an information point of view there is no fundamental difference between the BL and the ZOH assumption (besides the fact that the ZOH setup provides more high-frequency information), it is still advisable to match the choice for the basic assumption with the intended use of the model. In theory, it is possible to relate a discrete-time and a continuous-time model, using (10-4), if the basic assumption is met, but in practice, additional errors are added due to the nonideal experimental conditions. For some applications, the signal choice is not really critical, but for others, the application leads to a natural choice. The following applications are discussed in more detail next: controller design, physical interpretation, simulation and modeling of subsystems.

10.3.1.1 Controller Design. In model-based control design, a mathematical model of the device under test is required. For discrete controllers, it is clear that a discrete-time model is the best choice. The digital controller generates a ZOH excitation that is exactly known (it is available in the memory of the controller). Everything between the controller output and the observed system output (including the actuator and the noise reduction filters) should be modeled because it is part of the control problem. This is the standard setup for ZOH modeling; in fact, the whole ZOH theory originated from this problem. Even if the ZOH reconstruction is poor, the behavior can still be included in the characteristic by increasing the model order. The drawback of this approach is that the models are not portable from one setup to the other because plant model and signal reconstruction are mixed up in one single model. Another nonideal ZOH characteristic results in another model.

Conclusion: for digital control design the ZOH assumption is well suited. Even if the ideal ZOH reconstruction is not closely matched, the nonideal ZOH characteristic can be included in the model.

10.3.1.2 Physical Interpretation. In some applications, the user is not really interested in the identified model but uses it only as an intermediate step to get deeper physical insight into a problem. He wants to measure model parameters that are not directly accessible with classical instruments, for example, time constants, diffusion constants, or the values of some components in an electrical circuit. Mostly, it is not a good idea to try to identify these parameters directly as they are linked to the measurements through highly nonlinear relations, complicating the identification significantly, and they may not even be uniquely identifiable. It is easier to identify, first, an intermediate model, and to extract the physical parameters from this result. Usually, the coefficients of the differential equations are closer linked to these parameters than those of the approximating difference equations, so that continuous-time models are preferably used.

Although it is possible to identify continuous-time models under the ZOH conditions using a direct continuous-time parameterization (Ljung, 1999) or using dedicated preprocessing methods based on block-pulse functions (Sinha and Rao, 1991) or delta operators (Ninness and Goodwin, 1991), we advise starting from BL measurements. These techniques can be applied in the time domain (Van hamme et al., 1991) or in the frequency domain, even for arbitrary excitations, using the extended models as explained in Section 5.3.2. The preference for the BL setup might be surprising because there is a formal relation (10-4) between both approaches. However, this relation is valid only if the experimental conditions were in perfect agreement with the underlying assumptions (perfect ZOH, perfect BL), and this can be very hard to realize, especially for broadband ZOH excitations. In many cases the ZOH reconstruction is disturbed due to the load of the output impedance $Z_{\text{out}}(j\omega)$ of the ZOH reconstructor by the input impedance of the actuator $Z_{\text{in}}(j\omega)$, so that the actual generated spectrum differs from the theoretical one by $Z_{\text{in}}(j\omega)/(Z_{\text{in}}(j\omega)+Z_{\text{out}}(j\omega))$. On the other hand, it is quite easy to get a set of two identical, good antialiasing filters, so that the BL assumption is matched well.

10.3.1.3 Simulation and Modeling of Subsystems. Building a model for a very complex system is a tedious task. Instead of catching the system in one extreme complex model, it is much more feasible to split the problem into a series of subtasks, each modeling a subsystem. In principle, for each of these subsystems we can build a discrete-time model under the ZOH assumption. However, even if these submodels are perfect, they will not describe the actual signals in the cascaded system because the subsystems are not excited by ZOH excitations. This is very similar to the setup given in Figure 10-5. Assume that perfect ZOH models $L_{\text{ZOH}}(z^{-1}) = (1-z^{-1})Z\{L(s)/s\}$ and $G_{\text{ZOH}}(z^{-1}) = (1-z^{-1})Z\{G(s)/s\}$ are available for $L(j\omega)$ and $G(j\omega)$; then the cascaded system $L(j\omega)G(j\omega)$ is described not by $L_{\text{ZOH}}(z^{-1})G_{\text{ZOH}}(z^{-1})$ but by $(1-z^{-1})Z\{L(s)G(s)/s\}$, so that an error appears:

$$\frac{(1-z^{-1})Z\{L(s)/s\}(1-z^{-1})Z\{G(s)/s\}}{(1-z^{-1})Z\{L(s)G(s)/s\}} \tag{10-13}$$

Note that this error is independent of the order of cascading. In Figure 10-6, the error due to this wrong combination is shown for the cascade of two first-order systems:

$$L(s) = \frac{1}{1+s/(0.6\pi)} \quad \text{and} \quad G(s) = \frac{1}{1+s/(0.8\pi)} \tag{10-14}$$

In this case severe errors appear because we considered systems with a bandwidth that is large compared with the sampling frequency, so that the repeated spectra of the ZOH excitation (see Figure 10-2) are not filtered out by the plant. If this were the case, the errors would be much lower. However, for a general approach this is an undesired restriction.

A first possibility would be to transform the ZOH models to continuous-time models, using the inverse relation, and next apply (10-4) again to the cascaded continuous-time models (assuming that $f_s/2 > \text{Im(poles)}/(2\pi)$). As mentioned before, this approach relies heavily on the ideal ZOH behavior, which may be difficult to obtain. The sound approach is

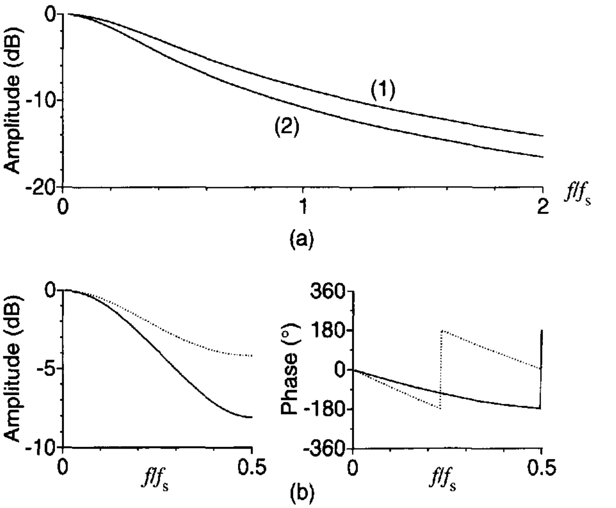


Figure 10-6. Illustration of the cascading error of ZOH models. (a) Transfer function of the original systems $L(j\omega)$ (1) and $G(j\omega)$ (2) and (b) comparison of the ZOH model of the cascaded system (solid line) with the cascade of the ZOH models (dots).

to select the BL assumption and combine it with discrete-time models. Although the resulting models lose their physical interpretation, they are perfectly suited for simulation. By increasing the complexity, an arbitrary precision can be obtained. Moreover, cascading of these models is allowed as long as the signals in the simulator obey the BL assumption.

Note that by using the same arguments, it is also possible to identify continuous or discrete-time models of an arbitrary subsystem of a complex system. Hooking the probes of the measurement device at the input and output of the subsystem makes it possible to zoom in on each accessible part of the overall system, as shown in Figure 10-4. The BL assumption is realized using good antialias filters. Because it is almost impossible to impose a ZOH excitation in the middle of a complex process, the ZOH assumption is not well suited to solve this kind of problem.

Conclusion: under the BL assumption it is possible to build continuous or discrete-time models of subsystems of a complex plant, even if they are preceded by nonlinear systems. These can be used, for example, as portable building blocks for simulators.

10.3.2 Impact of the Intersample Assumption on the Setup

The intersample assumption has a significant impact on the experimental aspects and the actual quality of the measurements. In each measurement setup it is important to reduce the errors. The identification methods take care of the stochastic errors but cannot cope with systematic errors. These should be removed in an appropriate calibration procedure. This is relevant only if accuracy is important, but why bother about consistency and efficiency if the systematic instrumentation errors dominate? Therefore, it is always necessary to check the quality of the measurement setup and to verify its impact on the quality of the final models. Typical errors that appear in many data acquisition channels are DC offsets and dynamic distortions due to the measurement channel characteristics $G_u(j\omega)$ and $G_y(j\omega)$. The offset errors can often be eliminated by excluding the DC information, while the compensation of the channel characteristics requires a calibration.

10.3.2.1 Perfect BL Setup. Under the BL assumption, two channels measuring the input and the output are needed. Because in Eq. (10-7) only the ratio $G_y(j\omega)/G_u(j\omega)$ appears, a relative calibration that measures this ratio will do, and this for $|\omega| < \omega_s/2$. In order to guarantee that the BL assumption is met, the acquisition channels should be equipped with antialias filters. An alternative is to filter the excitation signal so that no power is injected above $\omega_s/2$. If the plant is guaranteed to be linear, the measured input and output also obey the BL assumption. The advantage of this approach is that only one filter is required, so it is easier to get two identical measurement channels.

10.3.2.2 Perfect ZOH Setup. Under the ZOH assumption, the situation changes drastically. In this case, only the output is measured. From Eq. (10-5) it is seen that the acquisition channel should have a transfer function $G_y(j\omega) = 1$ in a frequency band that covers ω_s many times in order to pass the high-frequency components created in the ZOH reconstruction. This is a difficult constraint. An absolute calibration is required in this case to measure and compensate the channel characteristics. This will be a tedious task, especially if arbitrary excitations are used, since in that case the compensation should be done in the time domain using inverse filtering techniques (Pintelon et al., 1990; Kollár et al., 1991). The alternative is to select an instrument with a very large bandwidth compared with the sampling frequency and hope that the roll off and phase distortion will be small in the frequency band of interest. Notice that in this setup it is NOT allowed to use an antialias filter as this would eliminate all the repeated spectral contributions of the ZOH. These results are grouped in Table 10-1. From this table we conclude that it is easier to approach the ideal measurement setup for the BL assumption compared with the ZOH requirements. This suggests that due to the experimental constraints, the BL setup is best suited for accurate measurements of the system.

TABLE 10-1 Implications of the BL Assumption and the ZOH Assumption for the Ideal Measurement Setup

BL	ZOH
Two-channel measurement	Single-channel measurement
Relative calibration	Absolute calibration
No flat amplitude/linear phase required	Flat amplitude/linear phase required
Instrument bandwidth ≥ $\omega_s/2$	Instrument bandwidth ≥ many times ω_s
Antialias filters required	Antialias filtering not allowed

10.3.2.3 ZOH Setup for Control. For many control applications the situation is, luckily, not that bad. Often, the bandwidth of the plant is not very large (for example a few kHz or lower) and the sample frequency is typically chosen 10 times larger. This makes it possible to filter the output before feeding it back to the controller in order to reduce the out-of-band process noise without adding too much delay to the system. Under these conditions, the previously mentioned problems with the ZOH setup become less pronounced. Moreover, in this field, the desired accuracy is also much lower than the accuracy that is typically required in many measurement applications. This leads to the conclusion that the ZOH setup is the natural choice for digital prediction/control design, where the actuator is an intrinsic part of the modeling problem and high accuracy is not the first requirement.

10.3.3 Impact of the Intersample Behavior Assumption on the Identification Methods

In the stochastic approach to system identification, the cost function is completely set by the noise model. From Figures 10-3 and 10-4, it is seen that under the ZOH setup, the input is assumed to be known, whereas under the BL assumption the input is measured. As each measurement is disturbed by noise, two (correlated) noise sources are needed for the stochastic model under the BL assumption, whereas only one noise source on the output measurements is needed under the ZOH assumption. This has a significant impact on the general structure of the identification scheme: the BL assumption leads to the errors-in-variables approach (see Chapter 7), while the ZOH assumption is the basis for the prediction error methods (see Ljung, 1999 and Sections 8.9 and 8.10).

10.4 PERIODIC EXCITATIONS: FACTS

In the next sections, we study the impact of selecting periodic or arbitrary excitations. The impact of this choice is less dependent on the application field than the selection of the intersample assumption. Here, we give an enumeration of the facts connected to periodic excitations. In the next section, we give a more detailed discussion, including the appreciation, of these facts. Finally, we look into some user aspects of periodic excitations.

The most important facts about periodic excitations are:

1. Data reduction linked to an improved signal-to-noise ratio of the raw data
2. Separation of the signal from the noise disturbances
3. Elimination of nonexcited frequencies
4. Independent estimation of nonparametric noise models
5. Improved frequency response measurements
6. Detection, qualification, and quantification of nonlinear distortions
7. Improved model validation
8. Detection of trends

10.5 PERIODIC EXCITATIONS: DETAILED DISCUSSION AND APPRECIATION OF THE FACTS

The choice between periodic and nonperiodic excitations is one of the most important selections to be made during the experiment design. Many times, it is incorrectly linked to the selection between time and frequency domain identification. Periodic excitations open up a number of possibilities that are not accessible with arbitrary excitations and this for time and frequency domain identification.

10.5.1 Data Reduction Linked to an Improved Signal-to-Noise Ratio of the Raw Data

When periodic excitations are applied, it is possible to collect M successive periods (with length N_p) and to average the measurements in the time domain over these repeated periods, for example, for the output measurement (Figure 10-7):

$$\hat{y}(k) = \frac{1}{M}\sum_{l=1}^{M} y(k+(l-1)N_p) = \frac{1}{M}\sum_{l=1}^{M} y^{[l]}(k) \tag{10-15}$$

with $y^{[l]}(k) = y(k+(l-1)N_p)$. It is clear that due to the averaging process, the noise is reduced in $1/\sqrt{M}$ under very weak conditions (the total measurement time should be much larger than the correlation length of the noise), and $\lim_{M\to\infty} \hat{y}(k) = y_0(k)$ w.p. 1. Many dynamic signal analyzers offer this measurement option; for example, $M = 128$ averages are made over $N_p = 2048$ data points. As this reduces the record length at a very low computational cost, it is strongly advised to make full use of this option. Why should we restrict ourselves to 2048 data samples if we can get 128 × 2048 data samples almost for free? In practice, M is determined by the maximum measurement time T and the minimum required frequency resolution f_0: $M = f_0 T$. Note that for a fixed experiment time, the frequency resolution dropped by a factor M. Another interpretation of using consecutive periods is given in Section 10.5.3.

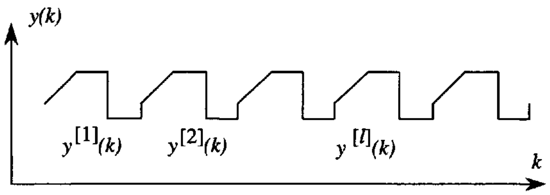


Figure 10-7. Making use of the periodic nature to improve the SNR.

10.5.2 Separation of the Signal from the Noise Disturbances

With periodic excitations, all nonperiodic variations are assigned to the disturbing noise (Schoukens et al., 1997b) because the signal repeats itself from period to period. This is a very general technique that can be applied even in the presence of nonlinear distortions. It fails in two situations:

- In the presence of periodic noise that is synchronous with the periodic excitation
- In the presence of bifurcations, where a periodic input does not necessarily result in a periodic output

10.5.3 Elimination of Nonexcited Frequencies

When a periodic excitation is applied, the user knows mostly what spectral lines are present. Often, not all lines are excited and it is possible to eliminate the "zero lines." This operation offers many new possibilities: generation of improved starting values and data reduction.

10.5.3.1 Improved Starting Values. Eliminating nonexcited lines does not change the asymptotic properties of well-designed estimators because the information matrix is not affected by removing zero lines. However, during the generation of starting values, simplified schemes such as ARX (linear least squares) methods are used. These are sensitive to the noise on the zero lines, so their elimination results in improved starting values (Schoukens et al., 1994). The risk of getting stuck in a local minimum is much larger if all spectral lines are retained, including the nonexcited lines.

10.5.3.2 Data Reduction. If a very wide frequency band has to be covered, fine resolution is needed at the low frequencies, whereas in the higher frequency bands the resolution can be reduced. For this reason, many systems are shown on a logarithmic frequency axis in a Bode plot. In these situations, it is advisable to excite the system with a semilogarithmic multisine, exciting the system at logarithmically spaced frequencies (on an equidistant grid of a DFT), so that a constant relative frequency resolution is obtained. This results in a sparsely filled spectrum, where only a small fraction of all lines is excited, for example, 200 out of 8192 lines. Only these frequencies should be retained so that the amount of raw data to be stored can be reduced further.

10.5.3.3 Special Case: Measuring M Consecutive Periods. In Section 2.2.3, it was shown that the spectrum of a signal consisting of M measured periods is sparse with nonzero lines at the multiples of M. The other lines are different from zero only due to the presence of noise. By putting them to zero, the spectrum of the averaged signal (e.g., $\hat{y}(k)$ in (10-15)) repeated over M periods is obtained.

10.5.3.4 Examples

Example 10.3. In Figure 10-8 the measured impedance of an electrochemical reaction $Fe^{3+} + e \rightarrow Fe^{2+}$ is shown in a frequency band from 0.123 Hz to 64 kHz (see also Section 12.7). Using a semilogarithmic multisine, this very wide frequency range is covered with a small number of points. □

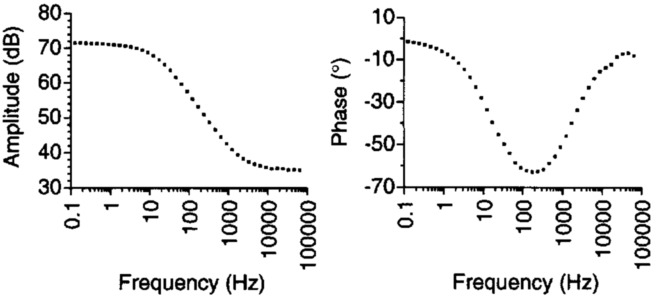


Figure 10-8. Measurement of the impedance of an electrochemical reaction in a wide frequency range on a semilogarithmic frequency grid.

Example 10.4. The effect of removing the nonexcited lines on the noise level is illustrated in Figure 10-9, where the impact of averaging and filtering (removing the zero lines) is shown on a signal with a semilogarithmic spectrum. □

Remark. In some special cases such as ARX modeling, the disturbing noise also contributes to the plant knowledge through the common denominator of the noise model and the plant model. Some modifications are needed to restore the information lost during the averaging (Gustafsson and Schoukens, 1998).

10.5.4 Independent Estimation of Nonparametric Noise Models

Once the signal is separated from the noise, the noise properties can be extracted from the noise record, leading to a nonparametric noise model. In practice, this is done by calculating the DFT for each period. Next, the sample mean and the sample variance are calculated (see Section 2.5.1). The sample mean carries the information, while the (co)variance can be used as a nonparametric noise model. This analysis is done before starting the identification process; for example, no model is selected yet. Consequently, the nonparametric noise model is independent of the plant model errors. This noise model can be used as a weighting in the estimation step, even during the generation of starting values.

Example 10.5. In Figure 10-10, the results of a nonparametric noise analysis of electrical machine measurements are shown. It not only gives the noise levels but also makes it possible to make a quality check of the measurements (e.g., What is the SNR? Is the system well excited in the frequency band of interest?) before starting the identification procedure.

10.5.5 Improved Frequency Response Measurements

When an integer number of periods is measured in steady-state conditions, the spectra of the signals calculated using the DFT are free of leakage errors. High-quality FRF measurements are obtained by first averaging the output and input spectra and next making the division (see Chapter 2). From the nonparametric noise analysis, the uncertainty on this estimate is obtained directly. The availability of the FRF measurements not only simplifies the model validation significantly (compare the FRF of the estimated transfer function with the measured one), it also gives a prior view of the required model complexity so that the model selection process is speeded up.

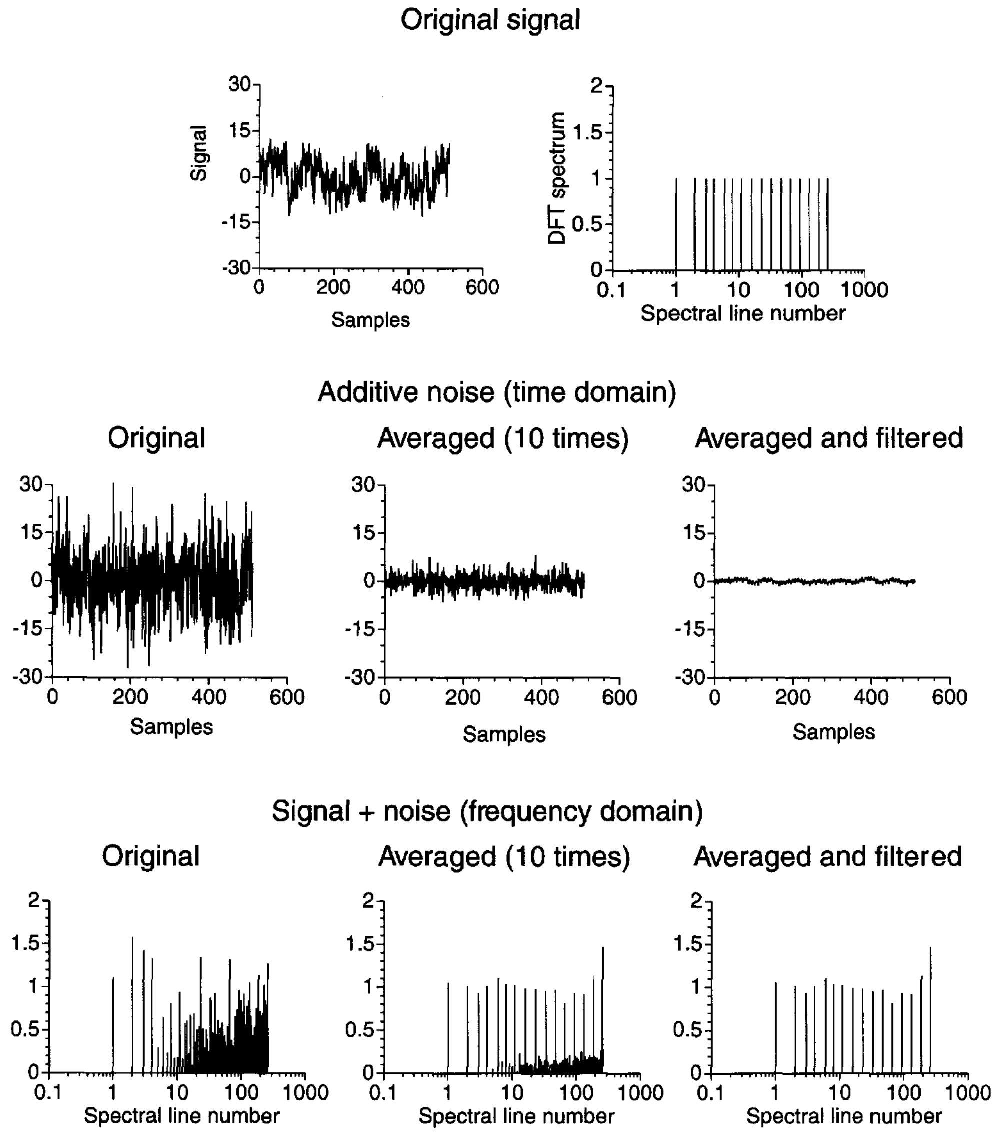


Figure 10-9. Impact of averaging and filtering of the noise.

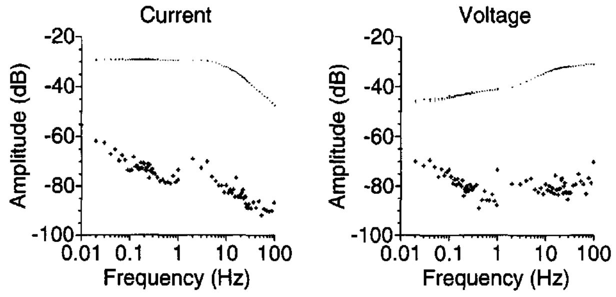


Figure 10-10. Separation of the signal and the noise using 10 repeated experiments on an electrical machine. — the raw measurements, + the nonparametric noise model.

10.5.6 Detection, Qualification, and Quantification of Nonlinear Distortions

Using, for example, an odd-odd periodic excitation consisting of nonzero components at frequencies $l_k f_0$, $l_k = 1, 5, 9, \ldots, l_{\max}$, gives a great deal of insight into the nonlinear behavior of the plant. The even nonlinearities become visible at the even frequencies $(2f_0, 4f_0, 6f_0, \ldots)$, while the odd nonexcited frequencies $(3f_0, 7f_0, 11f_0, \ldots)$ can be used to detect and quantify the level of the odd nonlinear distortions. Again, this is a nonparametric test that can be applied directly to the raw data before starting the identification process. These methods are extensively discussed in Chapter 3. Odd multisines also reduce the impact of nonlinear distortions, so that better measurements are obtained in a shorter time.

10.5.7 Improved Model Validation

As shown before, the direct noise characterization allows an absolute interpretation of the cost function, leading to significant advantages during the model selection and validation process. Not only is it possible to make an absolute detection of model errors, starting from the value of the cost function, but also the presence of unmodeled dynamics and nonlinear distortions can be checked (see Chapter 3).

Besides these global qualifications, the direct comparison of the measured FRF (see Section 10.5.5) with the modeled transfer function shows in which frequency bands the model fails, as illustrated next on a mechanical system (Figure 10-11). Using $M = 34$ measured periods, the mean value and the standard deviation (complex error) of the FRF were measured and compared with a parametric model. From this simple test we can conclude that the errors mainly appear at the resonance frequencies.

10.5.8 Detection of Trends

In this book, we consider time-invariant systems. Periodic excitations facilitate testing for this assumption by calculating the sample mean of each period $\mu_y(l) = \sum_{k=1}^{N} y^{[l]}(k)/N$, and checking that it does not vary systematically from one period to the other. This test makes it possible to detect very small variations of the mean value revealing the presence of a (weak) trend.

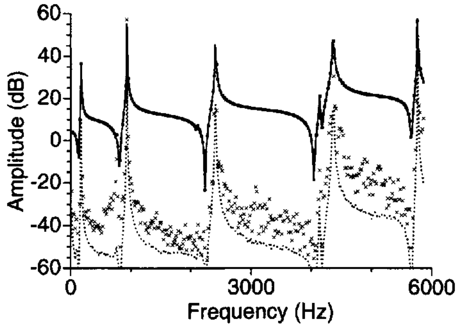


Figure 10-11. Measurement of a mechanical system (acceleration as a function of force): • measurement, ... measurement uncertainty σ_G, ____ model, × difference between model and measurement.

10.6 PERIODIC VERSUS RANDOM EXCITATIONS: USER ASPECTS

Three aspects are discussed in detail: how difficult it is to design a signal; what about the frequency resolution; and finally, how flexible the signal characteristics can be set.

10.6.1 Design Aspects: Required User Interaction

Although periodic excitations offer a number of extended possibilities, they are still not very popular. One of the basic reasons for the unpopularity of periodic signals might be the user interaction required to design them. The user should specify the power spectrum (what frequencies are excited) and the period T (determining the frequency resolution $f_0 = 1/T_0$ that is obtained in the frequency domain) before the periodic signal can be generated. For arbitrary signals, this information is seemingly not required. For example, the user can decide to generate a random sequence of length $T_0 = NT_s$ without even bothering about all these boring questions. However, this is a misleading impression. Also, in the latter case the user should realize that even if he does not select a power spectrum and a frequency resolution consciously, his choices fix (unconsciously) these parameters, as shown below.

10.6.1.1 Frequency Resolution. A periodic signal, with period T_0, probes the plant only at frequencies kf_0 with $f_0 = 1/T_0$, and it is completely insensitive to what happens between these frequencies, so that very narrow (compared with f_0) resonance peaks can be missed. However, although arbitrary excitations have a continuous spectrum, they do not offer an unrestricted resolution. As the signals are measured only in a finite time interval T_0, the windowing (or leakage) effect smears the spectra so that details below a resolution f_0 are also lost.

10.6.1.2 Imposing the Power Spectrum. A good excitation signal should excite the process in the frequency band of interest. For periodic excitations, the user has full control over the power spectrum, so it is easy to inject all the available power in this band. Moreover, because these are deterministic signals, it is guaranteed that after one period this power spectrum is realized exactly. The situation is completely different for arbitrary excitations. Only indirect control, using digital filters, is possible. Starting from a white noise generator, a colored noise process is generated. Because this is a stochastic signal, its power spectrum is reached only asymptotically, and for short data records significant differences between the actual power spectrum and the desired one can appear (see Section 2.6).

Example 10.6. In Figure 10-12 a typical example of an arbitrary (white uniform noise) and a periodic (a flat multisine with 31 components, period 64 samples) ZOH excitation is shown. Both signals have a flat power spectrum filling the complete frequency band. They were generated in 64 samples. It can be seen that a perfect realization of the flat power spectrum is obtained for the multisine, but the spectrum of the considered noise realization is not flat at all. The spectrum drops at some frequencies more than 10 dB, resulting in a poor SNR at those frequencies. □

10.6.1.3 Small Crest Factors. A second important aspect is the crest factor, measuring the ratio between the peak value and the rms value. Signals with a low crest factor make it possible to inject more power into the system for a restricted peak value of the excitation, resulting in a better SNR. Algorithms are available to minimize the crest factor of periodic sig-

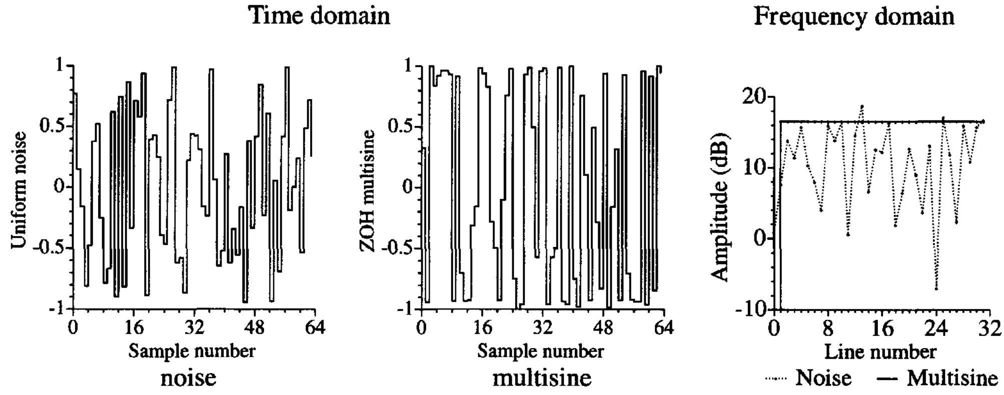


Figure 10-12. Comparison of an arbitrary (white noise) excitation with a periodic (flat multisine) excitation in the time and frequency domain.

nals for a specified power spectrum. Also for arbitrary excitations, the crest factor can be reduced, but at the cost of a distorted power spectrum: the user does not have full control over the power spectrum and the crest factor at the same time (see Chapter 4).

10.7 TIME AND FREQUENCY DOMAIN IDENTIFICATION

Time and frequency domain identification were considered as competing methods for a long time. However, in most cases, the frequency domain data are obtained by a DFT from the raw time domain data. Note that there is a one-to-one relation between the time and the frequency domain. The only difference is that some information is more easily accessible in one domain than in the other. This is also reflected in the methods themselves. Assuming that all transients have died out, it can be shown by Parseval's relation that the least squares cost functions are identical (Ljung, 1993). Moreover, it is shown in Section 5.3.2 that the leakage effect on transfer function modeling in the frequency domain data is described by exactly the same transient model structure as in the time domain, where it is used to include the initial conditions. So there is full equivalence between time and frequency domain identification. The only difference is how the available information is formulated, but even here mixed algorithms popped up that combine the time and frequency domain representation in one algorithm, for example, the use of a nonparametric noise model in the time domain (Gustafsson and Schoukens, 1998) or mixed implementations of ARX methods (Schoukens et al., 1998b). This results in the (surprising) conclusion that it is not possible to give a clear formal definition of what time and frequency domain identification schemes are. Nevertheless, we still like to use the term time domain identification for algorithms that mainly operate on time domain data and frequency domain identification for algorithms that work on frequency domain data.

Many advantages that are often claimed for frequency domain identification are intrinsically due to the periodic nature of the excitation. So, the prime question is not to choose time or frequency domain methods but to select periodic or nonperiodic excitations! As explained before, many of the advantages of periodic excitations can be used in both domains. Some of the equivalences between both domains are not directly visible and are discussed in more detail later. Next, we also deal with some important differences between time and frequency domain identification.

Again, the text is organized along the same lines. First, a series of facts is stated, in this case equivalences and differences. Next, a more personal interpretation of the methods is given in the section on the natural choices in identification.

10.8 TIME AND FREQUENCY DOMAIN IDENTIFICATION: EQUIVALENCES

10.8.1 Initial Conditions: Transient versus Leakage Errors

Two situations are considered: the nonparametric measurements (impulse response or frequency response) and the parametric transfer function model.

10.8.1.1 Nonparametric Measurement. Usually, the frequency domain is cursed because the time-frequency transform is prone to leakage errors, so that the frequency response function obtained by dividing these spectra will also be wrong. In the time domain, the standard nonparametric impulse response measurements are based on a correlation analysis; for example, for discrete-time systems we have

$$R_{yu}(t) = g(t)*R_{uu}(t) \tag{10-16}$$

with $R_{yu}(t)$ and $R_{uu}(t)$ the cross- and the autocorrelation (Bendat and Piersol, 1980). These have to be estimated from the finite set of available data, e.g., $\hat{R}_{yu}(k) = \sum_{l=0}^{N-1} y(l)u(l-k)/N$, where the data outside the window are put equal to zero. This shows that in the time domain windowing problems occur also, hence we can conclude that nonparametric measurements are prone to window errors in both domains. In the frequency domain these appear as leakage errors.

Remark. The Wiener-Hopf equations (10-16) are usually solved in the frequency domain (see (2-43)), emphasizing even more the time-frequency equivalence (Bendat and Piersol, 1980).

10.8.1.2 Parametric Models. For parametric system identification the time and frequency domain problem (initial conditions—leakage) is cured in exactly the same way for both domains: the model is extended with a transient term that is linear-in-the-parameters (see Chapter 5) so that the additional computational cost is low. This solves the problem completely.

10.8.1.3 Impact on Whiteness Tests. At the end of the identification process the model is validated. A very popular test is to check the whiteness of the residuals using a correlation analysis. It turns out that this test is very sensitive to unmodeled initial conditions because these appear as model errors. If they are not recognized as such, the test leads to a too complex model structure. For that reason it is strongly advised to add an initial conditions estimate (keeping the model parameters fixed) for the validation data at the beginning of the validation process or to wait until the transients have died out.

10.8.2 Windowing in the Frequency Domain, (Noncausal) Filtering in the Time Domain

Sometimes we want to emphasize or deemphasize some spectral bands, expressing our belief in the quality of these measurements. Eliminating frequency lines, as explained in Section 10.5.3, is an extreme example of this method. Weighting in the frequency domain (multiplication with a frequency weighting $W(\Omega)$) corresponds to a filter operation in the time domain (convolution of the measured input and output with the impulse response $w(t) = F^{-1}\{W(\Omega)\}$). Removing some frequencies from the data set corresponds to a rectangular window. This is a noncausal filter with an impulse response of the form $\sin(\alpha t)/(\alpha t)$, $t \in]-\infty, \infty[$. Its absolute value is not summable, as should be for a stable system, so there are no simple alternative formulations in the time domain. Moreover, the maximum likelihood interpretation of the classical prediction error scheme is also lost because the transformation matrix, as it is introduced in Söderström and Stoica (1989, pp. 251), is no longer triangular due to the noncausal filter operation. The filtered output depends not only on the past but also on the future data.

Remark. Prefiltering the raw data is somehow cheating because it also changes the noise model. For Box-Jenkins methods, prefiltering does not change anything because it is completely compensated by a similar change in the noise model. However, in practice, the identification process is continued with the simple noise model, so that the efficiency can be affected, or a bias can even appear in closed loop identification.

10.8.3 Cost Function Interpretation

The cost function that measures the goodness of the fit can be expressed in the time domain or in the frequency domain. It is clear that there should again be a full equivalence. However, a detailed study reveals some small differences depending on the practical implementation.

10.8.3.1 Extra Term in the Frequency Domain. If not only the plant model $G(z^{-1}, \theta)$ but also the noise model $H(z^{-1}, \theta)$ is identified, an additional term appears in the frequency domain interpretation of the cost function that was seemingly missing in the time domain expressions. Consider the maximum likelihood formulation for the generalized output error situation, assuming normally distributed noise and neglecting the plant and noise transient terms (see Ljung, 1993 and Appendix 10.A),

$$\sum_{k=0}^{N-1} \ln(|H(z_k^{-1}, \theta)|^2) + \left(N\lambda + \frac{1}{\lambda}\sum_{k=0}^{N-1} |\varepsilon(z_k^{-1}, \theta)|^2\right) \tag{10-17}$$

with $\varepsilon(z_k^{-1}, \theta) = H^{-1}(z_k^{-1}, \theta)(Y(k) - G(z_k^{-1}, \theta)U(k))$ and $\lambda = \mathrm{var}(e(t)) = \mathrm{var}(E(k))$. The first term in (10-17) does not appear explicitly in the frequency domain interpretation (8-50) of the classical time domain expressions. However, if the frequencies z_k are equidistantly distributed on the unit circle, then this extra term is exactly zero (see Appendix 10.A), and the difference between the time and frequency domain cost function disappears. This also reveals an additional condition on time domain identification: to get consistent noise models, it is not allowed to eliminate some frequency lines, nor is it allowed to restrict the identification to a subband on the unit circle.

The previous discussion is irrelevant if a prior known noise model is used; for example, the nonparametric model obtained from independent repeated measurements ($\lambda|H(z_k^{-1}, \theta)|^2$ in the third term of (10-17) is then replaced by the sample variance $\hat{\sigma}_Y^2(k)$ (2-31)). At that moment the noise model is fixed, and the additional term only adds a parameter-independent constant to the cost.

10.8.3.2 Optimization Aspects. When time and frequency domain identification lead to the same cost function, the only remaining difference is the optimization technique that is used to minimize the cost function. This can sometimes lead to tricky situations. Consider, for example, the generalized output error problem in the frequency domain formulation as discussed in Chapter 7 (see (7-82) with $\Omega = z^{-1}$, $\sigma_U^2(k) = 0$ and $\sigma_{YU}^2(k) = 0$):

$$\sum_{k=0}^{N-1} |\varepsilon(z_k^{-1}, \theta, Z(k))|^2 \tag{10-18}$$

where $\varepsilon(z_k^{-1}, \theta, Z(k))$ can be written as

$$\varepsilon(z_k^{-1}, \theta, Z(k)) = \frac{e(z_k^{-1}, \theta, Z(k))}{\sigma_Y(k)|A(z_k^{-1}, \theta)|} \quad \text{or} \quad \varepsilon(z_k^{-1}, \theta, Z(k)) = \frac{e(z_k^{-1}, \theta, Z(k))}{\sigma_Y(k)A(z_k^{-1}, \theta)} \tag{10-19}$$

with $e(z_k^{-1}, \theta, Z(k)) = A(z_k^{-1}, \theta)Y(k) - B(z_k^{-1}, \theta)U(k)$ (see (7-83)). Although both expressions in (10-19) lead to the same cost function (10-18), it turns out that the first form creates less problems with local minima. It also has a wider convergence region if a Gauss-Newton optimization method is used. In this method, the second-order derivatives are approximated from the first-order derivatives, and, seemingly, this approximation is better for the first expression (where some phase dependence is eliminated) than for the second. The disadvantage is that more calculations are needed to deal with the derivative of the absolute value, and slower convergence is obtained in the close neighborhood of the solution.

10.9 TIME AND FREQUENCY DOMAIN IDENTIFICATION: DIFFERENCES

10.9.1 Choice of the Model

Discrete-time models are the natural model class to be used in combination with time domain methods. Generalizing to other classes such as continuous-time models is not completely excluded, but it turns out from the literature that it is quite a complicated task (Sinha and Rao, 1991), and unexpected problems can appear (Söderström et al., 1997a, 1997b; Söderström and Carlsson, 2000).

In the frequency domain, the choice is more general. This is basically due to the fact that the differential (or difference) equations are replaced by algebraic equations in the related frequency variable. For continuous-time systems, the Laplace representation (transfer function) is used and evaluated on the imaginary axis ($s = j\omega$). For discrete-time systems, z-domain models are used and evaluated along the unit circle ($z = e^{j\omega T_s}$). Also, other frequency variables can be chosen; for example, $\sqrt{j\omega}$ is the natural representation for diffusion phenomena (e.g., used to model electrochemical processes) and $\tanh(\tau_R j\omega)$ is the logical choice to model commensurate microwave structures.

10.9.2 Unstable Plants

Prediction error techniques are mostly used to identify discrete-time models. These, typically, consider the following model structure:

$$y(t) = G(q, \theta)u(t) + H(q, \theta)e(t) \tag{10-20}$$

with q, the unit delay operator ($qu(t) = u(t-1)$). The plant $G(z^{-1}, \theta)$ and the noise $H(z^{-1}, \theta)$ models are rational functions of z^{-1}, parameterized in θ. $e(t)$ is white noise, and

$|H(z^{-1}, \theta)|^2$ models the power spectrum of the disturbing noise. The parameters θ are estimated by minimizing the prediction errors

$$\varepsilon(t, \theta) = H^{-1}(q, \theta)(y(t) - G(q, \theta)u(t)) \tag{10-21}$$

in least squares sense. It is clear that $H^{-1}(z^{-1}, \theta)$ and $H^{-1}(z^{-1}, \theta)G(z^{-1}, \theta)$ and their derivatives with respect to θ should be stable in order to be able to calculate (10-21). A stable plant and noise model is a sufficient condition to guarantee stability. Recently, a less restrictive solution was proposed for this problem (Forssell and Ljung, 2000a) by adding an all-pass section to the noise filter that cancels the unstable plant poles (see Section 8.10).

In the frequency domain, there is no problem to model unstable plants because these methods calculate the transfer function only at a discrete grid on the unit circle (or imaginary axis). As long as a pole does not coincide with one of these grid points, the cost function remains well defined; otherwise, regularization procedures can be used.

10.9.3 Noise Models: Parametric or Nonparametric Noise Models

The efficiency of the estimates is improved using a well-chosen weighting function. The best option is to choose it inversely proportional to the power spectrum of the disturbing noise. Time domain methods apply filtering techniques to realize this weighting. Without these, the full covariance matrix of the noise should be inverted and next used in each iteration step, leading to more calculation work. For prediction error methods (time domain), these noise filters are an intrinsic part of the method (see Eq. (10-21)), and the noise model $H(z^{-1}, \theta)$ is estimated together with the plant model $G(z^{-1}, \theta)$. The obvious advantage is that no constraints are imposed on the excitation at a cost of a second model selection problem. Moreover, the convergence is significantly slowed down.

For periodic excitations a nonparametric noise model is generated automatically without user interaction, leaving the complexity of the methods unaffected (see Section 10.5.2.)

10.9.4 Extended Frequency Range: Combination of Different Experiments

The number of measured data points N in an experiment is directly linked to the record length T and the sample frequency f_s, as $N = Tf_s$. The minimum record length is mainly imposed by the lowest frequency of interest (or the spectral resolution) $T = 1/f_0$. The sample frequency is imposed by the highest frequencies, which have to be chosen so that the frequency band of interest of the plant is covered $f_s \geq 2f_{\max}$. Hence, the minimum number of samples that should be measured and processed is $N > 2f_{\max}/f_0$. It is obvious that the number of measurements increases drastically if a large frequency range should be covered. This leads to impractical situations if N becomes very large, for example, 1 million points.

If periodic excitations are applied and combined with frequency domain identification, two significant simplifications can be made, the first being the data compression, as explained in Section 10.5.1, and the second consisting of a simplification of the experimental conditions. The latter is obtained by splitting the experiment into a number of subexperiments, each covering another frequency range. For each of these subexperiments a much shorter record length can be used while it is still possible to measure all the required Fourier coefficients. A similar approach cannot be applied to the ZOH models because they strongly depend on the sample frequency and, hence, combination of the different records is much more complicated. An alternative might be to use multirate systems (Crochiere and Rabiner, 1983).

Example 10.7. In the measurement of the electrochemical process (Figure 10-8) a wide frequency band [0.123 Hz, 64 kHz] had to be covered. To do this in one experiment, at least 1 million points are needed. The actual measurements were obtained in two experiments covering [0.123 Hz, 100 Hz] and [100 Hz, 64 kHz], using 4096 points each time. □

10.9.5 The Errors-in-Variables Problem

The errors-in-variables concept is a more general approach than the classical feedback situation as shown in Figure 10-13. The basic structure is captured in the gray area and it can be part of a larger structure, for example, a feedback system. However, this additional information is not used in the algorithms developed in this book. Starting from the measured input $u(t)$ and output $y(t)$, the plant model $G(z^{-1}, \theta)$ is identified. The noise sources can be correlated with each other but are assumed to be independent of the driving signal $r(t)$. In general, this is an unidentifiable problem (Anderson and Deistler, 1984; Bohlin, 1971) unless additional information is available (see Section 10.3.3). The following three situations have been considered: (i) the noise model structure is known to belong to a given class (for feedback) (Ljung, 1999), or (ii) the signals are known to be periodic, or (iii) an exactly known external reference signal is available. The first situation is the classical setup for time domain identification; however, some generalizations have also been studied in the time domain (Söderström and Stoica, 1981). These solutions impose additional restrictions on the noise behavior, such as the assumption that the input noise is white. Solution (ii) is a very attractive in combination with frequency domain identification and nonparametric noise models. Actually, this is the standard setup we consider in this book. The last possibility (iii) can be used to generate solutions in the time or in the frequency domain (Forssell and Ljung, 2000b; Schoukens et al., 1999b).

Note that the major difference between this setup and the ZOH setup is the information that is used to get the input signal $u_1(t)$. In the ZOH setup, the user relies on the validity of the ZOH assumption and the exact knowledge of the generator signal $u_d(k)$, whereas in this framework the assumption is replaced by a BL measurement.

10.10 CONCLUSIONS

In this chapter we have refined the order in the identification field by putting forward three basic questions that should be answered before starting the identification process: (i) What signal assumption should be used, zero-order-hold (ZOH) or band-limited (BL)?; (ii) What excitation should be preferred, arbitrary or periodic excitations?; and (iii) Finally, a last choice

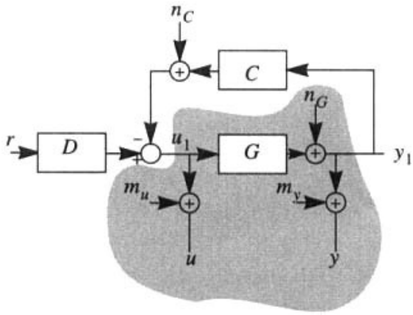


Figure 10-13. The errors-in-variables concept.

that should be made is the criterion to match the model and the data, a generalized output error or an errors-in-variables cost. This leads to the following major steps in the design of the identification process:

1. Experiment design

   1a. Select the ZOH or BL signal assumption

   1b. Choose between arbitrary or periodic excitations

2. Model: discrete- or continuous-time model?

   2a. ZOH $\rightarrow$ discrete-time models

   2b. BL $\rightarrow$ continuous-time models

   If the signal assumption is violated, the choice is free but a more complex model is needed to describe the measurements.

3. Cost function

   3a. Noise on input and output: errors-in-variables method

   3b. Noise on the input negligible: generalized output error method

And what about time or frequency domain identification? For some selections among the preceding choices, frequency domain identification methods seem to be preferred, for example, nonparametric noise models, very wide frequency ranges, continuous-time models, modeling subsystems for simulation, errors-in-variables with periodic excitations. For on line identification (Ljung and Söderström, 1983), or identification in the presence of nonstationary noise, time domain identification is the natural choice. For the other situations, both domains are equivalent, and the user can make the choice by using other criteria such as familiarity with one domain.

10.11 EXERCISES

10.1. The ZOH reconstruction of a discrete-time signal $u_d(kT_s)$ is

$$u_{\text{ZOH}}(t) = \sum_{k=-\infty}^{\infty} u_d(kT_s)\text{zoh}(t-kT_s)$$

with $\text{zoh}(t) = 1$ for $0 \leq t < T_s$ and $\text{zoh}(t) = 0$ elsewhere. Show that the spectrum after a ZOH reconstruction is $X_{\text{ZOH}}(\omega) = U_d(e^{j\omega T_s})\text{ZOH}(\omega/\omega_s)$ with $\text{ZOH}(x) = T_s(\sin\pi x/\pi x)e^{-j\pi x}$.

10.2. Consider the setup in Figure 10-3 on page 354 and prove relation (10-5) (hint: first calculate the spectrum of $u_{\text{ZOH}}(t)$, and $y(t)$. Next, apply the sampling theorem).

10.3. Consider the setup in Figure 10-4 on page 355 and prove relation (10-6) (hint: first calculate the spectrum of $u_{\text{ZOH}}(t)$, and $u(t), y(t)$. Next, apply the sampling theorem).

10.4. Consider the setup in Figure 10-5 on page 356 and show that Eq. (10-9) reduces to (10-5) for $L(j\omega) = 1$. Note that this result is different from what would be found starting directly from Eq. (10-10) using a Taylor series expansion of $\text{cosec}(x)$ (hint: check the convergence of the series expansions carefully).

10.5. Reproduce the transfer characteristics of Figure 10-6 on page 359 for the systems given in (10-14).

10.12 APPENDIX

Appendix 10.A: Frequency Domain Maximum Likelihood Solution of the Prediction Error Problem

Neglecting the plant and noise transient terms in the generalized output error model (5-65), the Gaussian likelihood function of the prediction error problem is given by

$$\frac{1}{\pi^N \lambda^N \prod_{k=0}^{N-1} |H(z_k^{-1}, \theta)|^2} \exp\left(-\sum_{k=0}^{N-1} \frac{|Y(k) - G(z_k^{-1}, \theta)U(k)|^2}{\lambda |H(z_k^{-1}, \theta)|^2}\right) \tag{10-22}$$

with $\lambda = \mathrm{var}(E(k)) = \mathrm{var}(e(t))$ (apply (14-14) to the circular complex normally distributed noise $E(k)$ in (5-65)). Taking the negative natural logarithm of (10-22) gives (10-17), within the constant term $N\ln(\pi)$.

The first term in (10-17) can be written as

$$\sum_{k=0}^{N-1} \ln(|H(z_k^{-1}, \theta)|^2) = 2\mathrm{Re}\left(\sum_{k=0}^{N-1} \ln(H(z_k^{-1}, \theta))\right) \tag{10-23}$$

For stable and inversely stable monic noise models $H(z^{-1}, \theta)$ we have

$$\ln(H(z^{-1}, \theta)) = \sum_{r=0}^{\infty} f_r z^{-r} \text{ for any } |z| > 1 \text{ with } f_0 = 0 \tag{10-24}$$

Because $\sum_{k=0}^{N-1} z_k^{-r} = 0$ for $r \neq 0$ it follows that

$$\sum_{k=0}^{N-1} \ln(H(z_k^{-1}, \theta)) = \sum_{r=0}^{\infty} f_r \sum_{k=0}^{N-1} z_k^{-r} = N f_0 = 0 \tag{10-25}$$

which shows the equivalence between the time and frequency domain cost functions. This result is also valid for unstable $H(z^{-1}, \theta)$ in identification in feedback (see Section 8.10). To prove this, it is sufficient to note that $|H(z_k^{-1}, \theta)|^2$ in (10-23) is not changed by reflecting the unstable poles of $H(z^{-1}, \theta)$ within the unit circle.

11

Guidelines for the User

Abstract: Before illustrating the previously developed methods on dedicated experiments and practical applications, a guideline for the user is set up in this chapter. Once more, it gives an overview of the complete identification process, but this time we discuss the decisions that should be made at each stage. So, the inexperienced user has a road map that reduces the risk of getting trapped and increases his chances of arriving at a good model for his problem.

11.1 INTRODUCTION

From the previous chapters, it became clear that identification is a complex task, bringing together many different skills. It is not enough to know the specific application field well (e.g., automotive, acoustics, electrochemistry), but the user is also expected to be familiar with measurement techniques, statistical theories, and numerical methods. As it is quite unlikely that all these skills are found in one person, the risk of making a serious mistake during one of the identification steps is always present. The aim of this chapter is not to turn all readers into absolute specialists but to offer some guidance to inexperienced users in order to increase their chances of a successful identification. To do so, we present two tools for the reader to select the best solution for his problem. First, we provide a table that will guide the reader to a good identification scheme (experiment setup, excitation design, estimator) for his problem, starting from a few simple questions. Second, we provide some rules of thumb that may help the reader to avoid frequently appearing problems and pitfalls.

11.2 SELECTION OF AN IDENTIFICATION SCHEME

The aim of this section is to make a proper selection among many possible identification schemes. By answering a few questions, we will guide the user to a good candidate method that can solve his problem. It should be clear that this does not guarantee that every problem will be solved, but at least the chances for success will be maximized. Of course, these guidelines are strongly influenced by our personal background and experiences. For these reasons, we strongly urge the reader to judge them critically and combine our advice with

Copyright © 2001 Institute of Electrical and Electronics Engineers. ISBN: 0-780-36000-1

his own experience. We first discuss the questions; then we look through the suggested solutions for the different situations.

TABLE 11-1 Selection of the Advised Identification Approach

Application?	Digital Controller Design			Other							
Domain?	Discrete-time			Discrete-time				Continuous-time			
Excitation?	Periodic	Nonperiodic		Periodic	Non-periodic			Periodic	Non-periodic		
Noise?	Don't care	Open loop	Closed loop	Don't care	Output	Input	Output	Don't care	Output	Input	Output
Advice	A	B	C	D	E		F	G	H		I

11.2.1 Questions

11.2.1.1 Application? What is the problem to be solved? For what reason do I need a model? We consider two possible answers. The first possibility is that a model-based digital controller will be designed on the basis of the identified model. The other possibility consists of all other goals, for example, designing simulation models, or more generally, models of substructures that can be cascaded; models for physical interpretation; models to identify physical parameters such as damping, resonance frequency.

11.2.1.2 Domain? Would you like to build a discrete-time model or a continuous-time model? It is clear that for digital controllers, a discrete-time model is needed. For the other applications, both domains are possible. For example, discrete-time models are very suitable to do time domain simulations on a digital computer, whereas continuous-time models are more suitable for physical interpretation.

11.2.1.3 Excitation? A crucial choice is the use of periodic or nonperiodic excitations. We extensively motivated the use of periodic excitations in this book. We strongly advise making this choice whenever it is possible, because it simplifies life significantly. We consider the use of nonperiodic excitations as a last resort.

11.2.1.4 Noise? How does the disturbing noise sneak into the process: As process noise? On the output measurements? Or on the input and output measurements? For some methods, this is not important at all, as indicated by the "don't care" answer, but for other methods it is a critical issue. Identification in feedback is a more tedious problem in control design than solving the same problem in open loop. The most general situation is that we consider (correlated) noise on the input and output measurements. This also includes the feedback problem.

11.2.2 Advice

For each situation we give advice, indicated by a letter. It is this advice that we formulate next.

A. Digital Control Design, Periodic Excitation

Method

Frequency domain ML (7-82), nonparametric noise model, ZOH setup.

Remarks

The "classical" time domain methods (see Ljung, 1999 and (8-50)) can be used if the input is observed without errors. The disadvantage, with respect to the advised method, is the need to estimate a parametric noise model. Also, the advantages connected to periodic excitations are not exploited (see Chapter 10).

For the advised method it is also not important if disturbing (jointly correlated) noise is present on the input-output measurements, or only output (process) noise, because the nonparametric noise model automatically takes care of it without user interaction.

B. Digital Control Design, Nonperiodic Excitation, Open Loop

Method

Prediction error (see Ljung, 1999 and (8-50)), parametric noise model, ZOH setup.

Remarks

In this case, the choice of the method does not affect the consistency of the model, but the efficiency will be influenced. The generation of uncertainty bounds is possible only if a full noise model is estimated.

Frequency domain alternatives, using nonparametric noise models, are available. However, in this case it would be an artificial choice, because these methods are quite involved for this situation (see Schoukens et al., 1999b).

C. Digital Control Design, Nonperiodic Excitation, Closed Loop

Method

Prediction error (see Ljung, 1999 and (8-50)), parametric noise model, ZOH setup.

Remarks

The noise model should be identified with care because the consistency of the estimate depends critically on it. A wrong noise model will lead to biased estimates. Alternatively, more complicated methods exist (see Forssell and Ljung, 1999, and the references in this chapter for an overview). Most of these methods require exact knowledge of an external reference signal.

D. Other Applications; Discrete-Time Model; Periodic Excitation

Method

Frequency domain ML (7-82), nonparametric noise model, BL setup.

Remarks

In this case, the "classical" prediction error methods such as ARX, OE, and Box-Jenkins (see Ljung, 1999 and also (8-50)) could be used as an alternative. However, in that case it is again necessary to take care of the noise conditions. These methods will fail if the input measurements are disturbed with noise. In that case, the methods referred to under C can be used again.

E. Other Applications; Discrete-Time Model; Nonperiodic Excitation, Output Noise Only

Method

Prediction error (see Ljung, 1999 and (8-50)), parametric noise model, BL setup.

Remarks See B.

F. Other Applications; Discrete-Time Model; Nonperiodic Excitation, Input-Output Noise

Methods

Two possibilities, both combined with the BL setup.

Time domain: Instrumental variables or projection methods (see Forssell and Ljung, 1999 and the references in this chapter) can be used. Exact knowledge of an external reference signal is needed. The efficiency can be increased by using parametric noise models.

Frequency domain: An extended ML estimator method can be used. A nonparametric noise model is extracted during the identification process (no longer before starting the identification). The method is much more involved because additional sets of initial conditions have to be estimated (Schoukens et al., 1999b). Also in this case, exact knowledge of an external reference method is needed.

G. Other Applications; Continuous-Time Model; Periodic Excitation

Method

Frequency domain ML estimator (7-82), nonparametric noise model, BL setup.

Remarks

For this method, it makes no difference at all that a continuous-time model is identified instead of a discrete-time model. It is also not important if disturbing (jointly correlated) noise is present on the input-output measurements, or only output (process) noise, because the nonparametric noise model automatically takes care of it without user interaction.

H. Other Applications; Continuous-Time Model; Nonperiodic Excitation, Output Noise Only

Advice

Apply periodic excitations if possible, and move to G.

Methods

Three possibilities,

Mixed Box-Jenkins method (see Section 8.9), BL setup.
In this case, a continuous-time plant model is identified, using a discrete-time parametric noise model as weighting to increase the efficiency. The method will be inconsistent under feedback.

Prediction error (see Ljung, 1999 and (8-50)), ZOH setup.
In this case, the continuous-time model is obtained by transforming the discrete-time model back to the continuous-time domain, assuming a perfect ZOH setup. As discussed before, this is a critical issue in practical applications.

An extended ML estimator method can be used, BL setup.
A nonparametric noise model is extracted during the identification process (no longer before starting the identification). The method is much more involved because additional sets of initial conditions have to be estimated (Schoukens et al., 1999b).

I. Other Applications; Continuous-Time Model; Nonperiodic Excitation, Input-Output Noise

Advice

Apply periodic excitations if possible, and move to G.

Method

An extended ML estimator method can be used, BL setup.
A nonparametric noise model is extracted during the identification process (no longer before starting the identification). The method is much more involved because additional sets of initial conditions have to be estimated (Schoukens et al., 1999b).

11.2.3 Special Case

If periodic excitations have been used and no noise information is available or can be extracted from the measurements, then the NLS-FRF (7-46) or LOG (7-54) estimators should be used. These estimators are "practically" consistent if the worst case input and output SNR per frequency is larger than 10 dB (see also Table 7-5 on page 238).

11.3 IDENTIFICATION STEP-BY-STEP

In this section, a series of general advice is formulated covering the different phases of the identification process as listed:

Check of the experimental setup
Design of an experiment
Preprocessing of the data
The identification step
Validation of the results

Each of these topics will be visited shortly, resulting sometimes in an overlap with earlier or later material in this book. However, we preferred to bring it all together here in order to optimize the global overview and insight of the reader, to minimize the risk of making extremely bad decisions.

11.3.1 Check and Selection of the Experimental Setup

In many cases, an identification run starts from data that were made available at some place. From our experience, it is definitely no loss of time to inspect the experimental setup and to check how the data were collected. Quite often, significant improvements can be obtained by very simple changes in the setup. Are the amplifiers properly set? What preprocessing is done on the raw data? What are the properties of the sensors that are used to get the raw data? Is the process operating under stationary conditions? ... Each of these aspects can have a considerable impact on the overall quality of the data. A short visit to the experimental site is very informative in revealing unexpected complications that would be detected only after wasting a lot of time and effort. For example, the data can be collected with a specific goal in mind (e.g., quality control), paying no attention to disturbing effects or bad settings that eventually make the data useless for the intended modeling purposes.

When looking at a measurement setup, two levels can be distinguished. A typical instrumentation configuration consists of a signal generator, a data acquisition arrangement, and a data-processing part, which extracts the parameters of interest from the raw data. Understandably, the sensor and actuator technology of the setup are closely connected to the application, whereas the actual data acquisition (amplification, attenuation, filtering, sampling, and quantization) is only loosely coupled to a specific application.

It is not easy to give general rules on the sensor or actuator part although it is always worthwhile to check for the linearity, offsets, and drifts of these devices. These questions are closely linked to the calibration of the setup. A good identification scheme makes it possible to reduce the impact of stochastic errors, but systematic errors should be eliminated, either by a proper calibration procedure that minimizes these errors or by extending the model to include them as unknown parameters. What choice is optimal depends strongly on the effort that is needed to go for one of these solutions. In general, the quality of the model improves with the quality of the measurements. Identification should be no excuse to do sloppy measurements, although it can open new possibilities to extract the desired information under worse operational conditions.

Because the acquisition part is quite similar for many instruments, more general advice can be formulated. A first general choice is to select between the ZOH and the BL setup. Although it is not critical in every situation, it is better to match this choice with the application in mind (discrete-time versus continuous-time model, simulation versus physical model, control application, etc.). The BL setup is assured to be valid by putting proper antialias filters in place before sampling the signals (check for the cutoff frequency, the stopband rejection, the linearity).

A second, very important, aspect is the synchronization between the generator and the acquisition. If periodic signals will be applied and explicit advantage will be taken of the periodic nature (averaging, nonparametric noise model, etc.), it is extremely important that the generator is synchronized with the acquisition. Otherwise, it is more complicated or even impossible to use the redundancy induced by the periodic behavior.

For critical applications, it is also necessary to check the stability of the master clock and the triggering in order to assure the best quality. Jitter (see Section 2.5.2) decreases the SNR of the measurements, and clock instabilities (Schoukens et al., 1996a) can induce systematic errors.

Advice

Visit the site of the experimental setup and talk to the operators to learn from their experience.

Check the systematic errors of the complete data acquisition.

Check the validity of the signal assumptions (ZOH or BL, antialias filters).

Pay attention to the synchronization of the setup.

11.3.2 Experiment Design

The second phase of the identification process is the design of the excitation signals. Sometimes, the user cannot influence the process at all. But even then, it should be checked whether the natural fluctuations carry enough information to give, at least, a chance of a successful identification. In all the other cases an excitation signal should be selected. This raises a series of questions immediately: what excitation level should be applied? What frequency band should be excited? In the initial phase of the identification process, we can only use prior information to set these values. For operator-controlled processes, the operators should have good knowledge of acceptable values. For other devices the nominal values given in the user manual might give some indications. And if none of this information is available, initial tests should give the required information. In this case we can only hope that our experience will help us to protect the device under test against dangerous overloads.

A second question is the linearity of the device. In this book we deal with methods to model linear systems. So it is important to know whether or not the linearity assumption is met. If the user is very confident, it is not necessary to check for nonlinear distortions; otherwise it is better to use excitations that make it possible to detect their presence.

If significant nonlinearities are detected, it is important to reflect carefully on the goal of the modeling process. Do you intend to extract the underlying linear system or are you interested in a best linear approximation? An appropriate excitation design in agreement with the previous selection should be made as explained in Chapter 3. In the first case, the amplitude should be made as small as possible (although some nonlinearities such as stick slip are pronounced in that case). In the latter case, the excitation should be representative of the class of excitations that will be applied later on to the device (e.g., same amplitude distribution and same power spectrum).

A third general question to be answered is the selection between periodic and arbitrary excitations. Imposing periodicity can be too strong a restriction in some applications. However, whenever it can be realized, it offers so many advantages in the rest of the identification process that we strongly urge the user to go for it. The design rules for these signals are discussed in Chapters 4 and 10. It is clear that the specific experimental requirements (synchronization, measuring an integer number of periods,...) should be met in order to take full advantage of the periodicity.

Finally, it is necessary to check whether the experiments have to be done under feedback conditions. These can be explicitly visible (a controller is in place) or can be implicitly present. An example is a mechanical device that interacts with the output of the shaker. In these cases, additional care is needed because many identification methods fail when proper action is not taken (see Section 11.2).

Advice

Go for periodic excitations whenever it is possible.

Select the amplitude range and frequency band of the excitation signal to cover the frequency band of interest.

Check for the presence of nonlinear distortions.

Check whether the device is captured in a feedback loop.

Keep your application in mind.

11.3.3 Preprocessing

The raw data, collected during the experiment, need to be preprocessed before starting the more demanding identification step. This not only facilitates checking for anomalies in the data (e.g., outliers or missing data) and bad experiments (poor signal-to-noise ratio) in an early phase of the identification process but also provides more insight into the complexity of the problem (look to the FRF), and makes it possible to separate different side aspects (such as trends or sensor drift) from the main task, which is to extract a linear parametric model from the data.

11.3.3.1 Removal of Trends, Drifts, and Offsets. In many problems, a linear model is used as a local linearization of a nonlinear system, around a given operating point, that might be slowly varying as a function of an uncontrolled input (e.g., temperature). If the user is not interested in building a full-blown nonlinear model that accounts for all these effects, it is important to eliminate their impact on the data as much as possible. A whole bunch of methods, ranging from very simple to complex procedures, can be used to eliminate these undesired effects. The simplest technique is to eliminate the DC offset from the measurements. This can be done effectively, under periodic operating conditions, by putting the DC line to zero after the DFT (or just do not use the DC line during the identification). In that case it is also very simple to observe slow drifts of the offset signals: calculate the mean value for a series of successive periods and check for systematic variations as a function of the period number. Next, simple correction methods such as linear interpolation between the successive DC values can be used to remove the first-order effects of these variations. If the variations are large, then more sophisticated trend-removing algorithms are recommended (McCormack et al., 1994a; Peirlinckx et al., 1996). An alternative is to disregard the spectral contributions at the low frequencies that are well below the reverse of the dominating time constants of the system. These techniques are also applicable to reduce the impact of sensor drift.

Advice

Check for trends by calculating the mean value over the successive periods.

Do not use the DC information during the identification.

11.3.3.2 Dealing with Outliers and Missing Data. Sometimes the measurements are very disturbed during a short interval (e.g., the presence of spikes or loss of data in a transmission link). This results in a few data that are very unreliable or even completely missing. In such a case the first advice is to repeat the experiment at a reasonable cost, if possible. Only if this is excluded, we advise restoring the data by trying to remove the artifacts. In case of missing data in highly oversampled signals, simple interpolation methods can help a lot (Rolain et al., 1998). In more complex cases, where the oversampling is low, the missing or heavily disturbed data can be replaced, considering them as missing data that also need to be identified (Pintelon and Schoukens, 2000; see also Section 11.3.4.5). This increases the complexity of the algorithms considerably and should be regarded as a last resort.

Advice

Perform new experiments.

If this is not possible, use simple interpolation methods if $f_{\max} < 0.1 f_s$.

Last resort: estimate the missing data.

11.3.3.3 Estimate the Nonparametric FRF. We strongly advise calculating, always, the nonparametric FRF estimate before starting the parametric modeling step. This additional effort is negligible (see Chapter 2). Simple visual inspection of the FRF not only gives a first impression of the model complexity but also allows a first evaluation of the quality of the data and reveals, in a very early phase of the process, many problems. Sensor failure, saturated amplifiers, acquisition overloads, etc. all result in an unexpected but mostly conspicuous distortions of the FRF. Finally, the user can check whether the appropriate frequency band is excited.

Advice

Calculate the FRF and make a visual inspection.

Select the frequency band of interest.

11.3.3.4 Check Whether the System Is Time Invariant. Slow-varying trends and offsets not only disturb the measurements but also can change the linearized behavior of the system intrinsically. Under these conditions, the user should carefully reflect on the value of his models and the aim of the experiments. A useful idea for the variability is a necessary condition to make a ripe decision. In case of periodic excitations, the FRF can be calculated for each individual period and, again, a simple visual inspection will give good insight into the significance of the problem.

Advice

Check the time invariance of the system by calculating the FRF over successive periods after trend removal (see 11.3.3.1).

11.3.3.5 Extract the Nonparametric Noise Model. It is very easy to extract the nonparametric noise model from periodic records (Chapter 2). Again, this information is very revealing.

Observing the SNR of the input and output measurements not only gives a good impression of the overall quality of the data but also shows where the noise sneaks into the measurements. A low SNR at the input (or the output) points to high noise levels at the input (or output). Low SNR values at the input and the output in combination with a high input-output correlation indicate dominating generator noise or process noise that is turning around in a feedback loop.

The noise levels are known as a function of frequency. So, the user can check whether or not the frequency band of interest is affected too much by the noise. It also gives feedback in an early stage of the identification process for the design of improved experiments, such as putting more power in the frequency bands with a too low SNR. Of course, the noise information can also reveal problems in the measurement setup and alert the user to their presence. For example, bad grounding can be denoted by the presence of high disturbing components at the harmonics of the mains frequency.

Advice

Make a nonparametric noise analysis.

Check for anomalies.

Judge the quality of the experiment.

Improve the experiment if necessary and possible.

11.3.3.6 Check for the Presence of Nonlinear Distortions. A final, but important, check is to look for the presence of nonlinear distortions. From Chapter 4 it is known that such distortions can be masked completely as white noise in the case of random excitations. All classical validation tests at the end of the identification process will fail to indicate their presence. This may lead to dangerous situations in which the user erroneously believes he captured a good model. A significant change in the excitation signal, in a later phase of the design process, would completely fool the quality of the predicted output. Moreover, the noisy behavior of the measurements is actually not due to the noise but should be attributed to the stochastic nonlinear distortions. For this reason, early detection of the presence and the level of nonlinear distortions is very valuable. It gives the user, from the very beginning, an idea of the best quality that can be obtained through linear modeling. This makes it possible to make a conscious decision to go on or to stop with the modeling effort before wasting a lot of time in the identification step.

Advice

Use specially designed periodic excitation to check for the presence of nonlinear distortions.

11.3.4 Identification

Only at the fourth step do we arrive, finally, at the kernel of the identification procedure where a parametric model is extracted from the preprocessed data. Just as for the previous phases, a number of user decisions have to be made. Among them, we will discuss the choice of a model class; the selection of the model complexity; the impact of initial conditions or transients; dealing with time delays; and finally, we spend a few moments on the problem of local minima.

11.3.4.1 Choice of a Model Class. In a first step the desired model class should be selected: do you want to get a continuous-time model (e.g., physical interpretation of the results), a discrete-time model (e.g., for control design or to set up a simulator), or one of the special models such as $\sqrt{s}$ to model diffusion processes or distributed systems. Remember that this choice should be matched with the selected experimental setup. Otherwise, more complex models will be needed.

Advice

Select the model class that best fits your application.

11.3.4.2 Selection of the Model Complexity. During the identification process, not only do the parameters need to be estimated but also the model order should be selected. There exist a series of simplified estimators, with increased noise sensitivity (e.g., linear least squares), that make it possible to estimate a whole bench of models in one step. This result can be used to get an initial guess of the required complexity. An alternative is to calculate the FRF to get an initial idea (e.g., counting the number of resonances for vibrating mechanical structures). Next, this guess should be refined using more advanced estimators. Two strategies are possible. The first one is conservative, starts from a simple model, and searches gradually for more complex models. The alternative is to go for a very complex model and check next what poles and zeros can be eliminated. This method can be used only if the estimator is robust for numerical singularities (common pole-zero pairs). In both cases, the cost function and a residue analysis are very valuable tools to guide one in the selection process (see Chapter 9).

Advice

Calculate the FRF to get an initial idea of the complexity of the problem.

Check the phase of the FRF and the cross-correlation between the input and output signals to detect the presence of a delay.

11.3.4.3 Impact of Initial Conditions or Transients. During the identification and validation it is important to safeguard against the impact of initial conditions (time domain) or leakage effects (frequency domain) on the data and the model. Both effects can be included in the model by adding an additional transient term (see Chapter 5). Even if these effects have only a second-order impact on the quality of the identified model, they become dominant during the analysis of the residuals. This can lead the user to very complex models because the correlation test of the residuals is very sensitive to these effects. For this reason, the user is advised to add these additional terms to his model whenever arbitrary excitations are used. For periodic excitations, the transients should be considered only if no integer number of periods is measured (and cannot be extracted from the raw data) or the measurements were not made under steady-state conditions (the transients have not yet disappeared at the beginning of the measurement).

Advice

Use periodic excitations and measure under steady-state conditions.

If this is not possible, estimate a transient term in each identification or validation step.

11.3.4.4 Dealing with Time Delays. Some systems, such as transmission lines or transport phenomena, cannot be modeled as a lumped system. An additional delay term becomes explicitly visible and should be added to the model. The presence of such terms can be recognized from the impulse response (where a delay is explicitly visible) or from the FRF (by looking for a rapidly varying (linear) phase). When a delay is present, we advise that all information is used to get a good initial estimate. It reduces the risk of stumbling on a local minimum during the optimization. Delay systems have many local minima, and it is very hard to find the global minimum.

Advice

Add an explicit delay term to the model and use all prior information available to get an initial value.

Restart the search, using different starting values, to make sure that you are not trapped in a poor local minimum.

11.3.4.5 Combining Experiments. It may happen that data sets of different experiments on the same plant are available. These data sets may originate from time domain experiments, frequency domain experiments, or time and frequency domain experiments. The basic question that arises then is how to combine these data sets in an optimal way. The solution to this problem depends on the prior knowledge and the type of experiments performed. We distinguish between the following three cases.

1. It is known only that the time and/or frequency domain experiments are independent (the noise in one experiment has nothing to do with the noise in the other experiments).
2. The independent time or frequency domain experiments are synchronized.
3. The time domain experiments stem from one experiment where at several time instances a (large) number of consecutive input and output samples are missing.

The following solutions are recommended for each of these situations:

1. To solve case 1 we apply the Gaussian maximum likelihood (ML) principle to independent experiments. The only thing to do is to extend the frequency domain data vector Z as $Z^T = [Z^{[1]T}\, Z^{[2]T} \ldots Z^{[M]T}]$ with $Z^{[r]}$ the input-output DFT spectra of the rth experiment (see (7-3)), and similarly for the noise (co)variances. If no periodic excitations are used, the equivalent initial conditions are different for each experiment and they should be added to the model parameters.
2. For synchronized experiments, the sample mean and sample (co)variance should be calculated, and these data are then considered as the raw input data for the identification process. Note that the improved signal-to-noise ratio also relaxes the starting value generating problem, resulting in a wider convergence rate of the search algorithms. The single experiment software can be used without any modification to handle the synchronized experiments, even in case of arbitrary excitations (see Appendix 11.B).
3. A first possibility to tackle case 3 is to handle the complete input-output data sets as independent experiments. However, it is better to express that the data sets stem from the same experiment. The identification starts, then, from the DFT spectra of the concatenated data sets (the missing data points are just taken out), and an extended model is used: for example, consider two data sets that are separated by missing data. The first set starts at $t = 0$ and the second at $t = K_c T_s$. The modeled output becomes

   $$Y^c(\Omega_k, \theta) = G(\Omega_k, \theta)U^c(k) + I(\Omega_k, \theta) + z_k^{-K_c} I_c(\Omega_k, \theta) \tag{11-1}$$

   with $U^c(k)$ the input DFT spectrum of the concatenated data sets and $I(\Omega_k, \theta)$, $I_c(\Omega_k, \theta)$ polynomials of order $\max(n_a, n_b) - 1$ for $\Omega = z^{-1}$ and order $\geq \max(n_a, n_b) - 1$ for $\Omega = s$ (see Appendix 11.C). All the unknown parameters are then estimated in the identification step.

11.4 VALIDATION

At the end of the identification process, it should be checked whether the identified model is a valid one. Ideally, the estimated model should be close to the exact one, but the reader should realize that the "exact" model is only an idealized concept. Most real-life systems cannot be described exactly by a rational transfer function. Moreover, because the exact system will always be unknown, we will never be able to answer that question. For this reason, we should focus on more realistic questions such as: Does the model describe the data well? Does the model fit my needs? These questions can be properly answered.

A set of tools is available to check whether all information is extracted from the data. We briefly repeat them here; for more details, the reader is referred to Chapter 9.

The test is based on the value of the cost function and the autocorrelation of the residuals.

(i) The cost is too small: check the noise analysis, are the (co)variances correct?

(ii) The value equals the expected value within the uncertainty bands: no information is left in the data (this should be confirmed by the residual analysis).

(iii) The value is too large: there are still model errors present. Their nature can be determined from a residual analysis. Correlated residuals point to unmodeled dynamics and demand increasing the model complexity. White residuals point to nonlinear distortions, and increasing the model complexity will not help.

Sometimes it is not necessary to extract all information from the data, especially when this would lead to very complex models. At that stage the reader should specify an acceptable error level (e.g., no model errors larger than 1 dB), and once this level is reached the model complexity is no longer increased. The choice between these options depends on the intended use of the model completely.

In the "classical" identification approach (see, for example, Ljung, 1999), it is strongly advised to split the available data in two sets: an identification set, used to identify the model, and a validation set to check for the model. Although this is a very robust check of the quality of the model, we prefer to use all the available experimental time and data to identify the model. The availability of the nonparametric (high-quality) FRF and a nonparametric noise model turns out to be a good alternative for the validation set. Of course, it always makes sense to perform a second experiment with another excitation signal, but then we advise using this information also during the identification step. Before starting the parametric identification, the two nonparametric FRF's can be compared with each other to check whether one model can be used to describe both experiments.

Advice

Compare the parametric model with the nonparametric FRF.

Check the value of the cost function.

Analyze the residuals.

11.5 APPENDIXES

Appendix 11.A: Independent Experiments

Because the Gaussian negative log-likelihood function (7-77) of the union of independent experiments is the sum of the contributions of the individual experiments separately, it follows that the ML solution (7-82) equals the sum of the ML cost functions the data sets separately. The only thing to do is to extend the frequency domain data vector Z as $Z^T = [Z^{[1]T}\ Z^{[2]T} \ldots Z^{[M]T}]$ with $Z^{[r]}$ the input-output DFT spectra of the rth experiment (see (7-3)), and similarly for the noise (co)variances. For arbitrary excitations the equivalent initial conditions are different for each experiment and they should be added to the model parameters. To see this it is sufficient to note that the weighted residual $\varepsilon(\theta, Z)$ of the ML solution (7-82) of the combined experiments can be written as

$$\varepsilon^T(\theta, Z) = [\varepsilon^{[1]T}(\theta, Z^{[1]})\ \varepsilon^{[2]T}(\theta, Z^{[2]})\ \ldots \varepsilon^{[M]T}(\theta, Z^{[M]})] \tag{11-2}$$

with $\varepsilon^{[r]}(\theta, Z^{[k]})$ the weighted residual vector of the rth experiment.

Appendix 11.B: Relationship between Averaged DFT Spectra and Transfer Function for Arbitrary Excitations

The experiments are synchronized (case number 2) if the phases of the true input DFT spectra are the same

$$\angle U_0^{[1]}(k) = \angle U_0^{[2]}(k) = \cdots = \angle U_0^{[M]}(k),\ k = 1, 2, \ldots, F \tag{11-3}$$

The solution consists of averaging the data $Z = M^{-1}\sum_{r=1}^{M} Z^{[r]}$ and changing the noise (co)variances accordingly $\sigma^2 = M^{-2}\sum_{r=1}^{M}\sigma^{[r]2}$ with $\sigma = \sigma_U$, σ_Y and σ_{YU}. The input-output DFT spectra of each experiment satisfy

$$\begin{aligned} A(s_k, \theta^{[r]})Y^{[r]}(k) &= B(s_k, \theta^{[r]})U^{[r]}(k) + I(s_k, \theta^{[r]}) + \Delta^{[r]}(s_k) \\ A(z_k^{-1}, \theta^{[r]})Y^{[r]}(k) &= B(z_k^{-1}, \theta^{[r]})U^{[r]}(k) + I(z_k^{-1}, \theta^{[r]}) \end{aligned} \tag{11-4}$$

with $\theta^{[r]} = [a_0 a_1 \ldots a_{n_a} b_0 b_1 \ldots b_{n_b} i_0^{[r]} i_1^{[r]} \ldots i_{n_i}^{[r]}]^T$, $r = 1, 2, \ldots, M$ (see (5-33), (5-34)). Averaging (11-4) over all experiments gives

$$\begin{aligned} A(s_k, \theta)Y(k) &= B(s_k, \theta)U(k) + I(s_k, \theta) + \Delta(s_k) \\ A(z_k^{-1}, \theta)Y(k) &= B(z_k^{-1}, \theta)U(k) + I(z_k^{-1}, \theta) \end{aligned} \tag{11-5}$$

where $\theta = [a_0 a_1 \ldots a_{n_a} b_0 b_1 \ldots b_{n_b} i_0 i_1 \ldots i_{n_i}]^T$ with $i_s = M^{-1}\sum_{r=1}^{M} i_s^{[r]}$, $s = 0, 1, \ldots, n_i$; $X(k) = M^{-1}\sum_{r=1}^{M} X^{[r]}(k)$ with $X = Y$ and U; and $\Delta(s_k) = M^{-1}\sum_{r=1}^{M}\Delta^{[r]}(s_k)$. Because the experiments are synchronized, the averaged DFT spectra will not tend to zero as $M \to \infty$. □

Appendix 11.C: Relationship between DFT Spectra of Concatenated Data Sets and Transfer Function

The proof will be given for the discrete-time case ($\Omega = z^{-1}$). That of the continuous-time case ($\Omega = s$) is similar and the analogy with the discrete-time case is the same as in Appendix 5.B.

Suppose that M_c consecutive samples are missing in the input and output signals, starting from $t = K_c T_s$. Define, furthermore, the signals $x_1(t)$, $x_2(t)$ as the data blocks of the known samples

$$x_1(nT_s) = \begin{cases} x(nT_s) & n = 0, 1, \ldots, K_c - 1 \\ 0 & \text{elsewhere} \end{cases}$$

$$x_2(nT_s) = \begin{cases} x((K_c + M_c + n)T_s) & n = 0, 1, \ldots, N - K_c - M_c - 1 \\ 0 & \text{elsewhere} \end{cases}$$

and $x^c(t)$ as the concatenation of these data blocks

$$x^c(nT_s) = \begin{cases} x(nT_s) & n = 0, 1, \ldots, K_c - 1 \\ x((M_c + n)T_s) & n = K_c, K_c + 1, \ldots, N - M_c - 1 \\ 0 & \text{elsewhere} \end{cases}$$

$x = u,\ y$. The Z-transforms $X_1(z^{-1})$ and $X_2(z^{-1})$, $X = U,\ Y$, of the signals $x_1(t)$ and $x_2(t)$, $x = u,\ y$, satisfy respectively (see Appendix 5.B, equation (5-83))

$$A(z^{-1})Y_1(z^{-1}) = B(z^{-1})U_1(z^{-1}) + I_1(z^{-1}) - z^{-K_c}I_2(z^{-1}) \tag{11-6}$$

$$A(z^{-1})Y_2(z^{-1}) = B(z^{-1})U_2(z^{-1}) + I_3(z^{-1}) - z^{-(N-K_c-M_c)}I_4(z^{-1}) \tag{11-7}$$

where $I_i(z^{-1})$, $i = 1, 2, 3, 4$, are polynomials of order $n_i = \max(n_a, n_b) - 1$ which represent the initial or final state of the system at $t = 0$, $t = K_cT_s$, $t = (K_c + M_c)T_s$ and $t = NT_s$, respectively. Multiplying (11-7) by z^{-K_c} and adding the result to (11-6) gives

$$A(z^{-1})Y^c(z^{-1}) = B(z^{-1})U^c(z^{-1}) + I_1(z^{-1}) - z^{-(N-M_c)}I_4(z^{-1}) + z^{-K_c}(I_3(z^{-1}) - I_2(z^{-1})) \tag{11-8}$$

where $X^c(z^{-1}) = X_1(z^{-1}) + z^{-K_c}X_2(z^{-1})$, $X = U,\ Y$, is the Z-transform of $x^c(t)$, $x = u,\ y$. Evaluating (11-8) along the unit circle at the DFT frequencies $z_k = \exp(2\pi jk/(N - M_c))$, taking into account that $z_k^{-(N-M_c)} = 1$, gives, after division by $\sqrt{N - M_c}$,

$$A(z_k^{-1})Y^c(k) = B(z_k^{-1})U^c(k) + I(z_k^{-1}) + z_k^{-K_c}T_c(z_k^{-1}) \tag{11-9}$$

with $X^c(k) = X^c(z_k^{-1})/(\sqrt{N - M_c})$, $X = U,\ Y$, the scaled DFT spectra of the concatenated data sets $x^c(nT_s)$, $n = 0, 1, \ldots, N - M_c - 1$, $x = u,\ y$, and where $I(z_k^{-1}) = (I_1(z_k^{-1}) - I_4(z_k^{-1}))/\sqrt{N - M_c}$ and $I_c(z_k^{-1}) = (I_3(z_k^{-1}) - I_2(z_k^{-1}))/\sqrt{N - M_c}$ are polynomials in z_k^{-1} of order n_i. Equation (11-9) results in parametric model (11-1) with $\theta = [a_0 a_1 \ldots a_{n_a} b_0 b_1 \ldots b_{n_b} i_0 i_1 \ldots i_{n_i} i_{c0} i_{c1} \ldots i_{cn_i}]^T$. □

12

Applications

Abstract: To illustrate its usefulness, the theory developed in the previous chapters is applied to a wide variety of real-life problems. Special attention is paid to the measurement setup (hardware requirements, physical limitations, ...). It is shown that it is possible to extract highly accurate parametric models for linear systems from noisy observations.

12.1 INTRODUCTION

The goal of this chapter is to illustrate the following aspects using real measurement examples:

1. Identification in feedback: CD player (Section 12.2)
2. Detection of nonlinear distortions using special designed excitation signals: CD player (Section 12.2)
3. Use of simulation (discrete-time) models for continuous-time systems excited by band-limited signals: bandpass filter (Section 12.3)
4. Use (power and limitation) of the best linear approximation of a nonlinear plant: electrical nonlinear circuit (Section 12.4)
5. Identification of a parametric noise model: electrical nonlinear circuit (Section 12.5)
6. Identification of a multivariable system: synchronous machine (Section 12.6)
7. Identification of diffusion phenomena ($\sqrt{s}$ models): traction battery and reduction of iron (Section 12.7)
8. Identification of systems with a time delay: microwave device (Section 12.8)

For each experiment, the measurement setup, the experiment design, the identification results, and the (cross-)validation of the model are discussed in detail.

System Identification: A Frequency Domain Approach. Rik Pintelon and Johan Schoukens
Copyright © 2001 Institute of Electrical and Electronics Engineers. ISBN: 0-780-36000-1

12.2 COMPACT DISC PLAYER

12.2.1 Measurement Setup

Device under Test. A Philips CD320/00G compact disc player modified to get access to the control loop of the radial servo system. All measurements were done at the start of track 1. The experiment time was maximum 26.2 s.

Generator. A HPE1445A VXI card in the ZOH mode without reconstruction filter. The sampling frequency was set to $f_s = 10\text{ MHz}/2^{10} \approx 9765.6\text{ Hz}$.

Acquisition. Two HPE1430A VXI cards with antialias protection on and sampling frequency equal to f_s. The two acquisition cards are synchronized with the generator card (the clock frequencies stem from the same 10 MHz mother clock). Voltage buffers (high input impedance, low output impedance) with a gain of about 0.5 were used since the 50 Ω input impedance of the acquisition loads the plant too much.

Excitation Signals. A special odd multisine with F frequencies at $l_k f_0$, $l_k = 1, 3, 9, 11, 17, 19, \ldots$, $k = 1, 2, \ldots, F$, was used to excite the system with $F = 305$ or $F = 39$. The crest factor of the signals was compressed to about 1.55 starting from random initialized phases. The rms value of the applied excitation signals was about 1 V.

12.2.2 Introduction

In this section, we identify the open loop transfer function of the radial position servo system of a CD player. First, a short description of the setup is given. Next, the successive identification steps are discussed: experiment design, nonparametric analysis (check for nonlinear behavior), parametric modeling, and finally, model validation is addressed.

Figure 12-1 shows a simplified block diagram of the compact disc (CD) player measurement setup. The block G stands for the cascade of a power amplifier, a low-pass filter, the actuator system, and, finally, the optical position detection system. The actuator transfer function represents the dynamics of the arm moving over the compact disc and is, in a first approximation, proportional to $1/s^2$. In practice, due to the friction, the double pole at the origin moves into the left half-plane to a highly underdamped position. This explains why the characteristics of the position mechanism of a CD player are very hard to measure in open loop. Moreover, at high frequencies, flexible modes appear that complicate the dynamic behavior of the arm significantly. The block C stands for the parallel implementation of a lead-lag controller with some additional integrating action that stabilizes the unstable actuator characteristics and takes care of the position control. In order to excite and to measure the open loop transfer function, two operational amplifiers have been inserted in between the lead-lag controller C and the power amplifier at the input of the process G. An external reference signal r is injected, and the resulting signals u, y are measured. The loop is also disturbed by the process noise d, mainly induced by tracking irregularities due to potato-shaped spirals; noneccentric spinning of the disc; and dirt, stains, and scratches on the disc surface. The following relations exist between the Fourier spectra (assuming that they all exist):

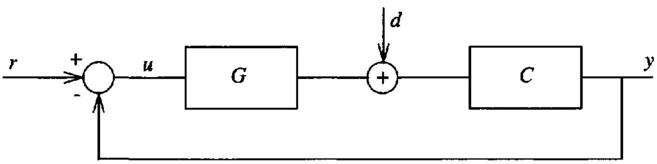


Figure 12-1. Setup of the CD measurements. The transfer function GC is modeled.

$$\begin{aligned} U(j\omega) &= \frac{1}{1+G(j\omega)C(j\omega)}R(j\omega) - \frac{C(j\omega)}{1+G(j\omega)C(j\omega)}D(j\omega) \\ Y(j\omega) &= \frac{G(j\omega)C(j\omega)}{1+G(j\omega)C(j\omega)}R(j\omega) + \frac{C(j\omega)}{1+G(j\omega)C(j\omega)}D(j\omega) \end{aligned} \qquad (12\text{-}1)$$

In the absence of disturbances, the open loop transfer function between u and y is $G(j\omega)C(j\omega)$, and it is the aim of this section to provide a parametric model for it.

12.2.3 Experiment Design, Preliminary Measurement

For control reasons, a 582.5 Hz sinusoidal wobble signal is internally injected in the feedback loop. It is measured at different points in the electronic circuit and serves as an input signal for an automatic gain controller (AGC), to compensate, among other things, for the effect that the displacement of the arm is not perpendicular to the track over the whole disc, resulting in a variable gain of the process. The wobble signal complicates the measurement process significantly, as it is more than 20 dB above the normal signal levels in the loop. For this reason, we had to make a careful experiment design to eliminate its impact on the measurements.

As external reference signal a multisine $r(t) = \sum_{k=1}^{F} A_k \sin(2\pi f_0 l_k t + \varphi_k)$ with $F = 305$, $f_0 = 2.3842$ Hz, $l_k = 1, 3, 9, 11, 17, 19, 25, 27, \ldots$, and $A_k = \text{constant}$ is used. The frequencies are selected to allow the detection of nonlinear distortions. The multisine is generated with a clock frequency of 10 MHz/2^{10} and $N = 4096$ points in one period. In the first experiment, $M = 64$ periods of this signal are measured (256 K points). The long record is broken in 16 blocks of four periods each. This is done in order to reduce the leakage effect of the wobble signal on the rest of the spectrum (no integer number of periods of the wobble signal is measured because its frequency is not synchronized to the measurement system; see Section 2.2.3). Starting from these 16 spectra, the amplitude spectrum of u and y is estimated and is shown in Figure 12-2. In this figure, it can be seen that the contribution of the reference signal is clearly above the disturbance level. Also, the wobble signal (with its leakage) is clearly visible and its amplitude is more than 20 dB above the signals of interest. That is the main reason for combining four periods in one FFT as this reduced the leakage considerably. Moreover, in the second phase it makes it possible to apply a Hanning window, reducing the leakage further, without generating systematic errors in this special case (see Section 2.2.3).

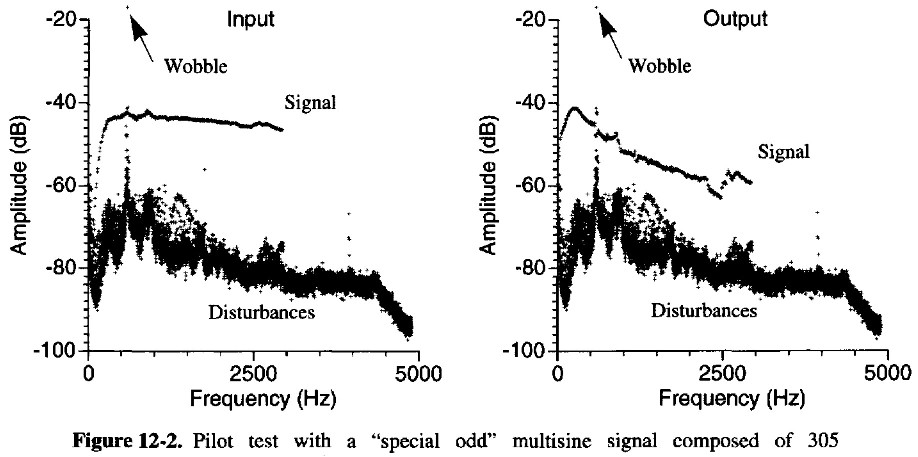


Figure 12-2. Pilot test with a "special odd" multisine signal composed of 305 components.

12.2.4 Quantifying the Nonlinear Distortions

Checking the nonexcited lines of the pilot test seems to indicate the presence of odd nonlinear distortions, but they are almost completely hidden under the noise level of the test. Hence, a second experiment with a reduced set of frequencies ($f_0 = 19.07$ Hz with $F = 39$) is made. From Chapter 3 it is known that for a normalized random multisine the ratio linear/nonlinear output power is an $O(F^0)$, so that the relative level of the nonlinearities is not affected by this modification. However, because the same power is concentrated in fewer components, the SNR of the measurements is increased, and a more sensitive test is obtained. The results are shown in Figure 12-3, where it is seen that the odd distortions are now well above the noise level of the test. Especially at the lower frequencies a very high distortion can be seen, indicating that the linearized models will have poor value in this frequency band.

12.2.5 Identification

After these tests, we already have a good idea about the limiting quality of the model. For the given input power, the nonlinearities are certainly larger than –30 dB. The parametric model will be identified using the data of Section 12.2.3. In order to safeguard against disturbing the AGC (based on the wobble), we left a gap of at least 15 Hz around the wobble frequency where the system is not excited. First, a nonparametric noise model is extracted from the data. A typical result for the output is shown in Figure 12-4. After processing the 256 K points, the mean value and the (co)variances are calculated. It turns out that there is an extremely high correlation ($\sigma^2_{YU}/\sqrt{\sigma^2_U\sigma^2_y} \approx -1$) between the noise on U and Y. From (12-1), it is seen that this indicates that the process noise dominates completely the measurement noise. The errors-in-variables approach followed in this book accounts, automatically, for this behavior. A 24th-order discrete-time model ($n_a = n_b = 24$) was identified. The measured FRF is compared with the parametric model in Figure 12-5. As can be seen, a very good fit is obtained. The residuals are below the noise level. Only at the low frequencies, where the nonlinearities detected in the nonlinearity test are very large, is the fit poor. Because we knew in advance that in this band the data are of poor quality, the frequencies below 230 Hz were not considered during the fit. The cost function of the fit is 842.6, while a theoretical value of 256.5 is expected. This points to model errors. However, the autocorrelation of the residuals is white as shown in Figure 12-6; therefore we can conclude that the best linear approximation is extracted (see Section 9.5.2 and 11.4). The remaining errors are due to the nonlinear behavior of the process. A stability analysis showed that two poles of the model were unsta-

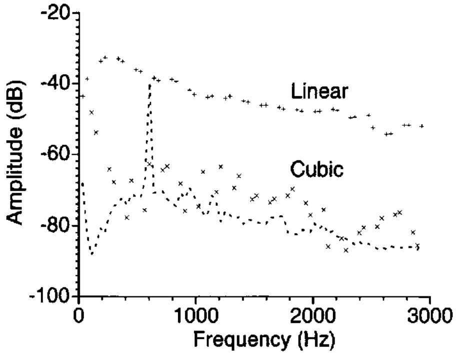


Figure 12-3. Nonlinearity test: + power on the FRF measurement frequencies, x: detected cubic distortions (after compensation for the linear feed through), --- σ_Y.

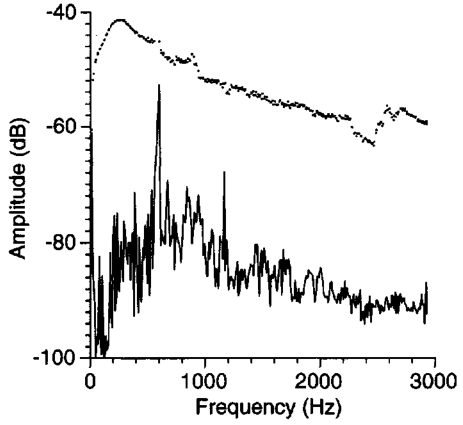


Figure 12-4. SNR of the output measurements after processing the raw data.

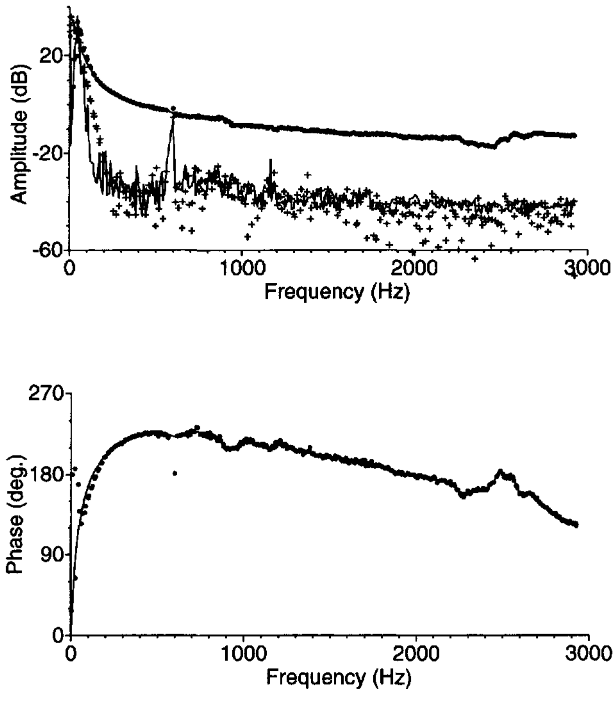


Figure 12-5. Comparison of the estimated transfer function (full line) with the measured FRF (dots). The residuals (+) are compared with the 95% noise level (thin full line).

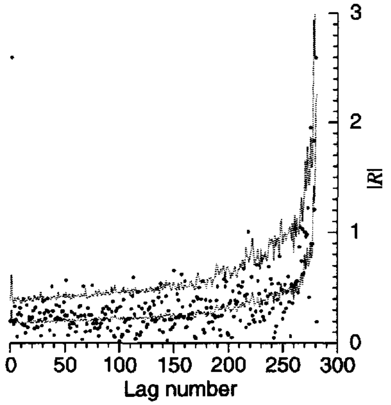


Figure 12-6. Whiteness test of the residuals with a correlation analysis (dots). The broken lines give the 50% and 95% uncertainty levels. Note that only at lag zero is a value significantly different from zero found.

ble ($z = 1.021 \pm j0.00344$). However, the corresponding closed loop is stable and, hence, the model is valuable for a closed loop analysis. This instability is due to the fact that the system has two poles, almost equal to one (double integration in z -domain), that are very difficult to estimate because of the presence of the nonlinearities in this band.

12.3 EXTRACTION OF A SIMULATION MODEL

In this section it is shown that it is possible to construct a model for the discrete-time simulation of a continuous-time system for band-limited excitations. For this purpose, we have selected a very linear and low-noise bandpass filter. We performed two experiments on this system under identical conditions. In the first experiment the filter is excited with a periodic excitation. The input and output are measured using good antialias filters. Under these conditions a continuous-time (s -domain) model should be theoretically used (BL assumption, see Sections 10.2 and 10.3). Instead we identify a discrete-time model. In the second experiment, the filter is driven with normally distributed noise with the same power and bandwidth as the periodic excitation of the first experiment. The measured output is simulated with the discrete-time model obtained from the first experiment, using the measured input, and the quality of the prediction is analyzed. These steps are discussed in more detail next.

12.3.1 Experimental Setup and Measurements

DEVICE UNDER TEST. A Brüel&Kjær passive filter, type 1630, with the weighting switch on off, and terminated with 179.5 kΩ. This is an octave bandpass filter, centered around 500 Hz.

GENERATOR. A HPE1445A VXI card in the ZOH mode without reconstruction filter. The sampling frequency was set to $f_s = 10 \text{ MHz}/2^{11} \approx 4882.8 \text{ Hz}$.

ACQUISITION. Two HPE1430A VXI cards with antialias protection on and sampling frequency equal to f_s. The two acquisition cards are synchronized with the generator card (the clock frequencies stem from the same 10 MHz mother clock). Voltage buffers (high input impedance, low output impedance) with a gain of about 0.5 were used because the 50 Ω input impedance of the acquisition loads the plant too much.

EXCITATION SIGNALS. A special odd multisine with $F = 185$ frequencies at kf_0, $k = 1, 3, 9, 11, 17, 19, ..., 737$ was used to excite the system in the first experiment. The frequency $f_0 = f_s/2^{11} \approx 2.38 \text{ Hz}$ such that one period of the signal contains $N = 2048$ points. In the second experiment, a normally distributed zero mean random excitation was used. Both signals have the same power (rms value of 90 mV) and the same bandwidth (1.8 kHz).

MEASUREMENTS. In the first experiment the input and the output were measured during eight periods ($M = 8$). The measurements were started once the transient disappeared. In the second experiment, 20,480 points of the random process were measured.

12.3.2 Processing

The averaged spectra and their variance (on the average) at the excitation lines are plotted in Figure 12-7. From this figure it is seen that the noise level is extremely low (SNR of about 90 dB, except at some harmonics of the mains frequency). Also, the linearity of the

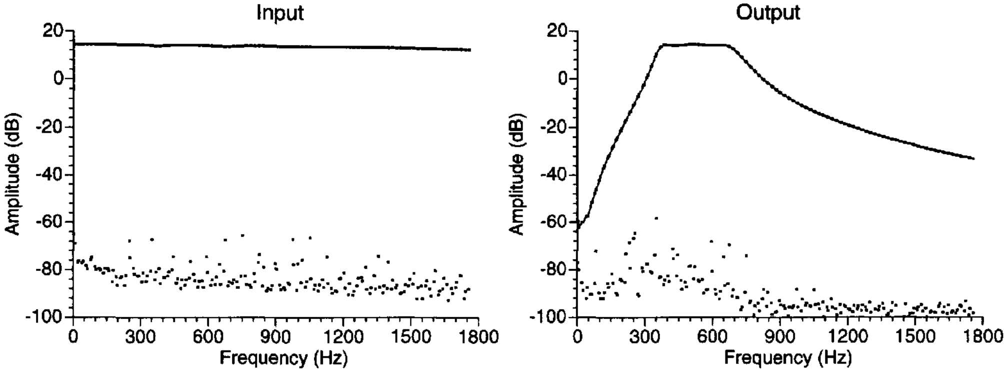


Figure 12-7. Measured input and output spectrum. Upper line: the amplitude spectrum. Lower dots: the standard deviation of the complex noise.

system is checked by verifying the spectra at the nonexcited frequencies. The even nonlinear contributions are 70 dB below the linear output in the passband, and the odd nonlinear contribution are even below 80 dB.

A sixth-order s-domain model $(n_a = n_b = 6)$ could describe the data (cost function 924, theoretical value 178.5). However, for the simulation purpose, a discrete-time model is needed. A seventh-order model $(n_a = n_b = 7)$ is needed to get an optimal fit (cost function 1055). The increase in complexity with one pole and zero is needed to compensate for the discrete-time approximation of a continuous-time system. For a sixth-order discrete-time model $(n_a = n_b = 6)$ the cost function is 1.2×10^5. The model is stable in both cases. The seventh-order model is shown in Figure 12-8.

This model is used in the second experiment to simulate the measured output using the measured input. We get

$$\hat{y}(t) = \sum_{n=0}^{N-1} \hat{g}(t-n)u(n) \tag{12-2}$$

with $u(n)$ the observed input, $\hat{g}(t)$ the impulse response of the identified model $G(z^{-1}, \hat{\theta}_{\mathrm{ML}}(Z))$, $\hat{y}(t)$ the simulated output, and N the number of observed time domain samples. The difference between the observed output $y(t)$ and the simulated output $\hat{y}(t)$ is called

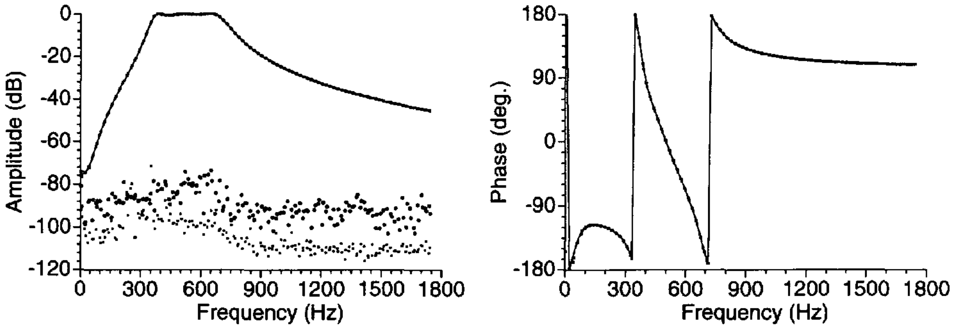


Figure 12-8. Comparison of the measured and modeled transfer function: upper dots: measurements; —: model; dots: amplitude complex error; x: standard deviation on the measurements σ_G.

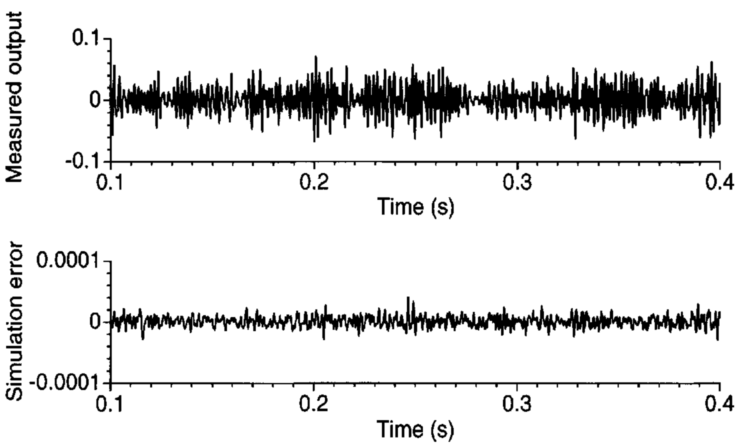


Figure 12-9. Comparison of the simulated and measured output for random excitation.

the simulation error. A part of the simulated output $\hat{y}(t)$ is shown in Figure 12-9, where it is seen that an extremely good result is obtained. The ratio $(y(t)-\hat{y}(t))_{\text{rms}}/(y(t))_{\text{rms}}$ was 0.4×10^{-3}, which is at the level of the detected nonlinearities.

12.4 IDENTIFICATION OF THE BEST LINEAR APPROXIMATION IN THE PRESENCE OF NONLINEAR DISTORTIONS

The aim of this section is to illustrate the use of the best linear approximation. This is done on the electrical nonlinear circuit described in Sections 3.4.6 and 3.5.2. We check whether the extracted linear models on the basis of (random) multisines can simulate the (nonlinear) plant output for normally distributed, random noise excitations. Three models are considered: the first one obtained using a Schroeder multisine (see Figure 3-11), the second one with a crest factor–minimized random multisine, and the last one with a random multisine without crest factor minimization. The models are obtained with a weighted output error method (Eq. 12-1), using respectively the averaged FRF and its sample variance $\hat{\sigma}^2_{\hat{G}}(k)$ as the raw data and the applied weighting. These are extracted from FRF measurements using 10 different realizations of the odd random phase multisine, with $F = 168$ as excitation. In Figure 12-10 the simulation errors are shown. The Schroeder multisine results are poor compared with the random phase results. The simulation errors (difference between the observed output $y(t)$ and the simulated output $\hat{y}(t)$ (12-2)) of the minimized crest factor model (c) have a smaller standard deviation but have more spikes than the random phase multisine (d). These spikes appear at those instances where the output makes a large excursion. It is clear that the application will have a strong impact on the final choice of the model.

Notes

(i) The best linear approximation makes it possible to predict, accurately, the response of a nonlinear plant to Gaussian random noise having the same power and bandwidth as the random phase multisine (see Definition 3.2) used for the identification. The simulation error is, however, bounded below by the stochastic nonlinear contributions ($y_s(t)$ in Figure 7-15 on page 244). If the error $y_s(t)$ is too large

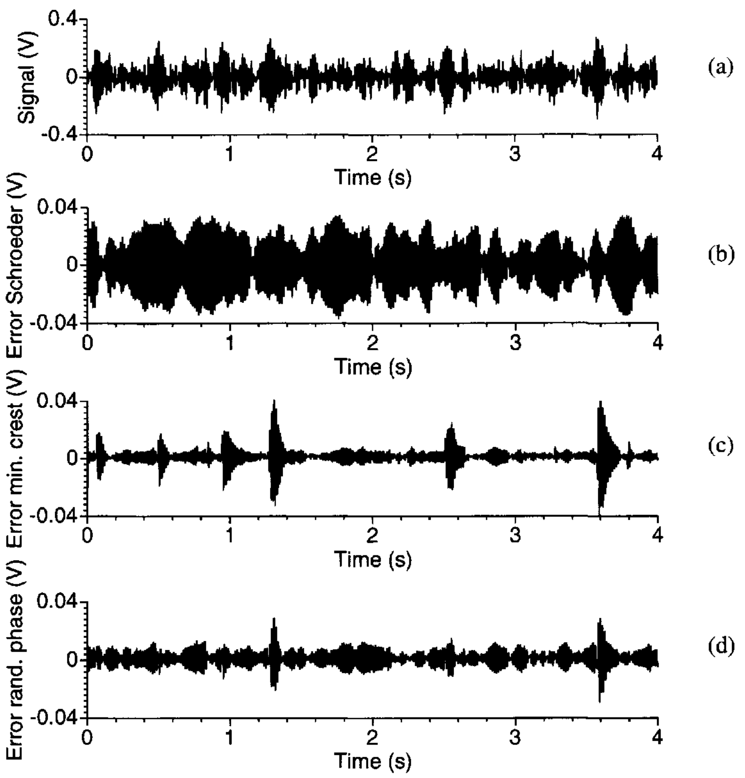


Figure 12-10. Simulation error for the linear model obtained with, respectively, a Schroeder multisine (b), random phase minimal crest factor multisine (c), and random phase multisine (d). (a) shows the output signal.

for the particular application one has in mind, then it can be reduced only by also modeling the nonlinear behavior of the plant. This is often a very difficult task.

(ii) Similar results are obtained if a Box-Jenkins model (see Section 5.7.3) is used to identify the system.

12.5 IDENTIFICATION OF A PARAMETRIC NOISE MODEL

12.5.1 Measurement Setup

Device under Test. A second-order electrical circuit with an internal feedback loop. The output signal is mainly disturbed by the process noise.

Generator. HPE1445A VXI card in the ZOH mode without reconstruction filter. The sampling frequency was set to $f_s = 10 \text{ MHz}/2^{14} \approx 610 \text{ Hz}$.

Acquisition. Two HPE1430A VXI cards with antialias protection on and sampling frequency equal to f_s. The two acquisition cards are synchronized with the generator card (the clock frequencies stem from the same 10 MHz mother clock). The measured signals pass through a voltage buffer (high input and low output impedance) with a gain of about 0.5 before being fed to the acquisition units, which have an input impedance of 50 Ω.

Excitation Signals. Two experiments have been performed, the first with a band-limited random signal, and the second with a random phase multisine (see Definition 3.2). Both signals have the same power (rms value of 16 mV) and the same bandwidth (about

200 Hz). The random phase multisine has a flat amplitude spectrum ($|U_k|$ in (3-7) is independent of $k = 1, 2, \ldots, F$) and harmonic frequencies $(2k+1)f_0$, $k = 0, 1, \ldots, F-1$, with $f_0 = f_s/2^{13}$ and $F = 1342$.

12.5.2 Identification

In this experiment, the main disturbing noise source is the process noise $n_p(t)$ and the generator noise $n_g(t)$ and the measurement errors $m_u(t)$, $m_y(t)$ can be neglected (see Figure 7-2 on page 185). Hence, the errors-in-variables problem of Figure 7-2 can be simplified to the general output error problem of Figure 8-5 on page 299. $N = 10{,}400$ input-output samples of the random excitation experiment are used to identify the hybrid Box-Jenkins model

$$Y(s_k, \theta) = G(s_k, \theta)U(k) + T(s_k, \theta) + H(z_k^{-1}, \theta)E(k) + T_H(z_k^{-1}, \theta) \tag{12-3}$$

with the prediction error method (8-50). It turns out that a model structure with a second-order plant ($n_a = 2$, $n_b = 0$ and $n_i = 2$) and an eight-order noise model ($n_c = 8$, $n_d = 8$ and $n_j = 7$) explains the data very well (see Figures 12-11 and 12-12). Figure 12-11 compares the estimated plant model $G(s_k, \hat{\theta}_{\mathrm{BJ}}(Z))$ with the frequency response function $\hat{G}_{\mathrm{r}}(s_k)$ obtained from the random excitation experiment

$$\hat{G}_{\mathrm{r}}(s_k) = \frac{Y(k) - T(s_k, \hat{\theta}_{\mathrm{BJ}}(Z)) - T_H(z_k^{-1}, \hat{\theta}_{\mathrm{BJ}}(Z))}{U(k)} \tag{12-4}$$

and Figure 12-12 compares the estimated noise model $\hat{\sigma}H(z_k^{-1}, \hat{\theta}_{\mathrm{BJ}}(Z))$, where $\hat{\sigma}$ is estimated as

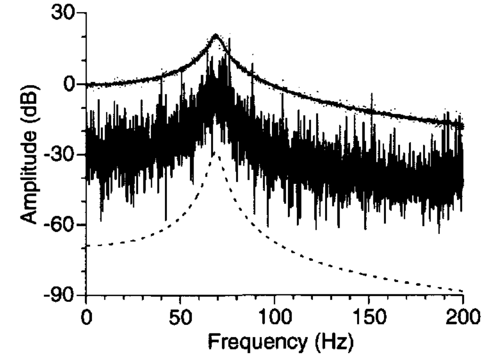


Figure 12-11. Validation of the identified plant model with the random excitation measurements. Dots: frequency response function (FRF) $\hat{G}_r(s_k)$ (12-4) obtained from the random excitation experiment, solid line: difference between FRF and plant model $\hat{G}_r(s_k) - G(s_k, \hat{\theta}_{\mathrm{BJ}}(Z))$, dashes: variance of the plant model $\mathrm{var}(G(s_k, \hat{\theta}_{\mathrm{BJ}}(Z)))$.

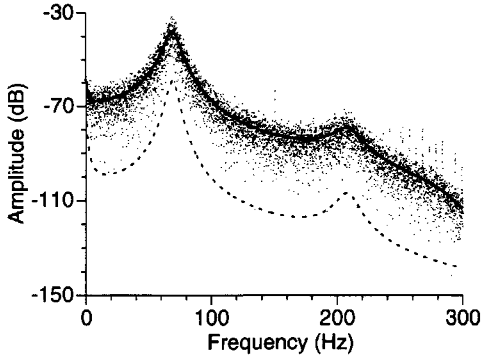


Figure 12-12. Validation of the identified noise model with the random excitation measurements. Dots: noise residuals $\hat{N}_Y(k)$ (12-6), solid line: noise model $\hat{\sigma}H(z_k^{-1}, \hat{\theta}_{\mathrm{BJ}}(Z))$, and dashes: variance of the noise model $\mathrm{var}(\hat{\sigma}H(z_k^{-1}, \hat{\theta}_{\mathrm{BJ}}(Z)))$.

$$\hat{\sigma}^2 = V_{\text{PE}}(\hat{\theta}_{\text{BJ}}(Z), Z)/(F - n_\theta/2) \quad (12\text{-}5)$$

with the noise residuals

$$\hat{N}_Y(k) = Y(k) - G(s_k, \hat{\theta}_{\text{BJ}}(Z))U(k) - T(s_k, \hat{\theta}_{\text{BJ}}(Z)) - T_H(z_k^{-1}, \hat{\theta}_{\text{BJ}}(Z)) \quad (12\text{-}6)$$

Note that the plant model is shown only within the bandwidth of the excitation signal, whereas the noise model is shown from DC ($f = 0$) to Nyquist ($f = f_s/2$).

12.5.3 Cross-Validation

The estimated Box-Jenkins model is cross-validated with the periodic measurements in Figures 12-13 and 12-14. For the periodic measurements, the identification starts from the measured frequency response function. The sample mean $\hat{G}(s_k)$ and sample variance $\hat{\sigma}_{\hat{G}}^2(k)$ required by the SML estimator (8-10) are calculated from $M = 10$ consecutive periods. Figure 12-13 compares the estimated plant model $G(s_k, \hat{\theta}_{\text{BJ}}(Z))$ with the frequency response function $\hat{G}(s_k)$ obtained from the periodic excitation experiment, and Figure 12-14 compares the estimated parametric noise model $\hat{\sigma}H(z_k^{-1}, \hat{\theta}_{\text{BJ}}(Z))$ with the nonparametric noise model $\hat{\sigma}_G(k)|U(k)|$ obtained from the periodic measurements. It follows that the estimated Box-Jenkins model explains the periodic excitation experiment very well, which is a nice illustration of Theorems 3.12 and 3.16. Note that the cross-validation of the noise model is possible only within the bandwidth of the excitation signal.

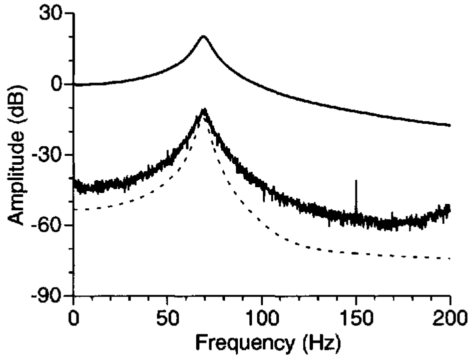


Figure 12-13. Cross-validation of the identified plant model with the periodic excitation measurements. Bold line: frequency response function (FRF) $\hat{G}(s_k)$ obtained from the periodic excitation experiment, solid line: variance measured FRF $\hat{\sigma}_{\hat{G}}^2(k)$, and dashes: difference between the plant model $G(s_k, \hat{\theta}_{\text{SML}}(Z))$ identified from the periodic excitation measurements and the plant model $G(s_k, \hat{\theta}_{\text{BJ}}(Z))$ identified from the random excitation measurements.

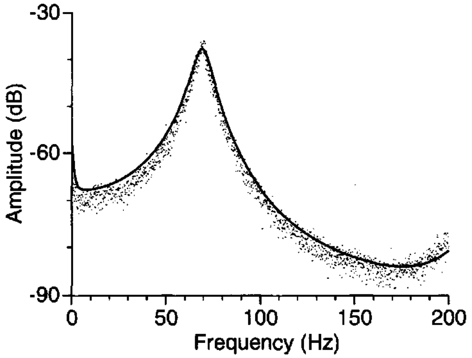


Figure 12-14. Cross-validation of the identified noise model with the periodic excitation measurements. Dots: nonparametric noise model $\hat{\sigma}_G(k)|\hat{U}(k)|$ obtained from the periodic excitation experiments, and solid line: parametric noise model $\hat{\sigma}H(z_k^{-1}, \hat{\theta}_{\text{BJ}}(Z))$ obtained from the random excitation experiments.

12.6 SYNCHRONOUS MACHINE

12.6.1 Measurement Setup

Device under Test. A 20 kVA salient pole synchronous machine: four poles, nominal voltage 220 V, nominal frequency 50 Hz, and field-to-armature turns ratio $N_{af} = 2.83$.

Generator. Two generators (HPE1445A VXI cards) in the ZOH mode without reconstruction filter. The sampling frequency was set to $f_s = 10\text{ MHz}/2^{14} \approx 610\text{ Hz}$.

Acquisition. Four HPE1430A VXI cards with antialias protection on and sampling frequency equal to f_s. The four acquisition cards are synchronized with the two generator cards (the clock frequencies stem from the same 10 MHz mother clock). The measured signals pass through a voltage buffer (high input and low output impedance) with a gain of about 0.5 before being fed to the acquisition units, which have an input impedance of 50 Ω.

Excitation Signals. Two random phase multisines $u_1(t)$ and $u_2(t)$ (see Definition 3.2) with the same bandwidth (about 230 Hz) and $F = 57$ frequencies each. The signals $u_1(t)$ and $u_2(t)$ have a flat amplitude spectrum ($|U_{1k}|$ and $|U_{2k}|$ in (3-7) are independent of $k = 1, 2, \ldots, F$) and harmonic related frequencies $k_1 f_0$ and $k_2 f_0$, respectively, where $f_0 = f_s/2^{16}$ and $k_1, k_2 \in \mathbb{N}$ with $k_1 \neq k_2$. The values of k_1 and k_2 are chosen such that the frequencies $k_1 f_0$ and $k_2 f_0$ are alternating and logarithmically spaced in the band $[0.1\text{ Hz}, 230\text{ Hz}]$.

Measurements. $M = 8$ consecutive periods of the steady-state response are measured. The number of samples per period and per signal equals $N = 2^{16} = 65{,}536$.

12.6.2 Measurement Results

Jef Verbeeck (Department ELEC of the Vrije Universiteit Brussel, Belgium) and Philippe Lataire (Department ETEC of the Vrije Universiteit Brussel, Belgium) have provided us with the experimental data. Figure 12-15 shows the measurement setup where the electrical machine is oriented in its d-axis. The rotor (field winding) and stator (armature winding) of the electrical machine are simultaneously excited using two thyristor rectifier bridges. The rms values of the field $i_f(t)$ and armature $i_a(t)$ currents are chosen such that the magnetization level is zero: $(i_f)_{\text{rms}}/(i_a)_{\text{rms}} = N_{af}$. The DFT spectra $I_f(k)$ and $I_a(k)$ of the field and armature currents are related to the DFT spectra $E_f(k)$ and $E_a(k)$ of the field $e_f(t)$ and armature $e_a(t)$ voltages by the impedance matrix $Z(j\omega_k)$

$$\begin{bmatrix} E_a(k) \\ E_f(k) \end{bmatrix} = Z(j\omega_k) \begin{bmatrix} I_a(k) \\ I_f(k) \end{bmatrix} \text{ with } Z(s) = \begin{bmatrix} Z_{11}(s) & Z_{12}(s) \\ Z_{12}(s) & Z_{22}(s) \end{bmatrix} \tag{12-7}$$

Two experiments are performed to estimate the frequency response matrix (FRM) $Z(j\omega_k)$ (see also Section 2.7): in the first experiment $i_a^{[1]}(t) = u_1(t)$ and $i_f^{[1]}(t) = u_2(t)$; in the second experiment $i_a^{[2]}(t) = u_2(t)$ and $i_f^{[2]}(t) = u_1(t)$. The current and voltage DFT spectra of the first experiment are shown in Figure 12-16. In the frequency region below 0.1 Hz, the frequency components are alternating dominant in the armature $i_a(t)$ and field $i_f(t)$ current. Above 0.1 Hz, the frequency components injected in the armature winding become visible at the field side and vice versa. This effect is caused by the nonideal output impedance of the thyristor rectifier bridges. The FRM estimate is given by

$$\hat{Z}(j\omega_k) = \hat{\mathrm{Y}}(k)\hat{\mathrm{U}}^{-1}(k) \tag{12-8}$$

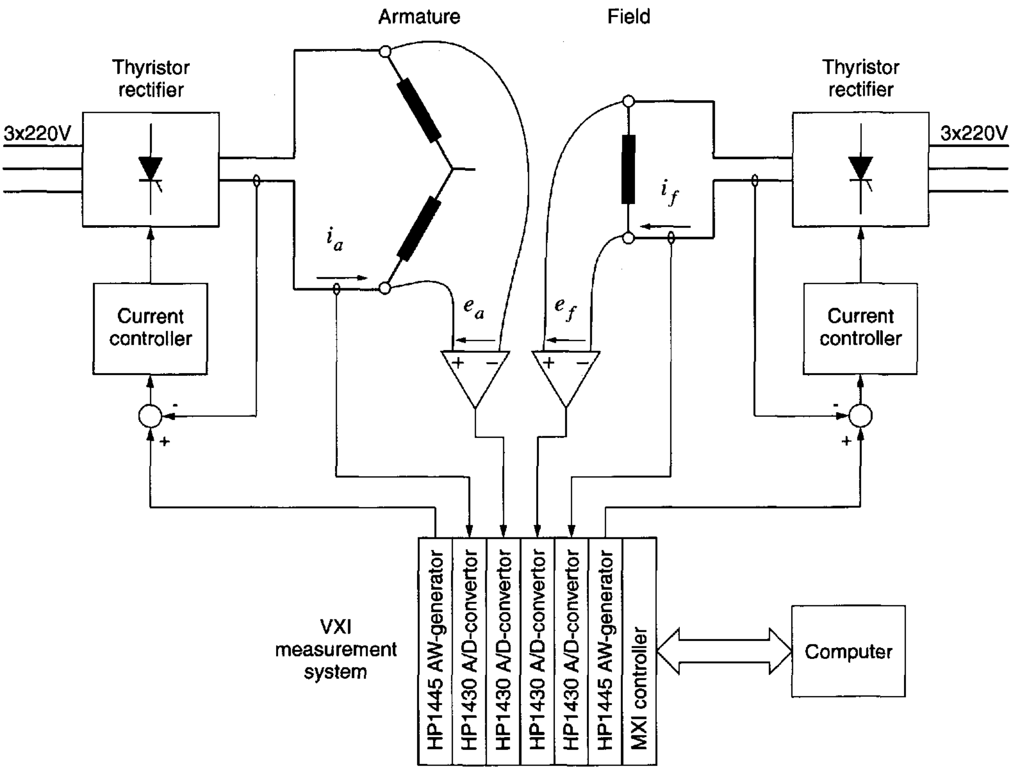


Figure 12-15. Double excited measurement setup for the d-axis.

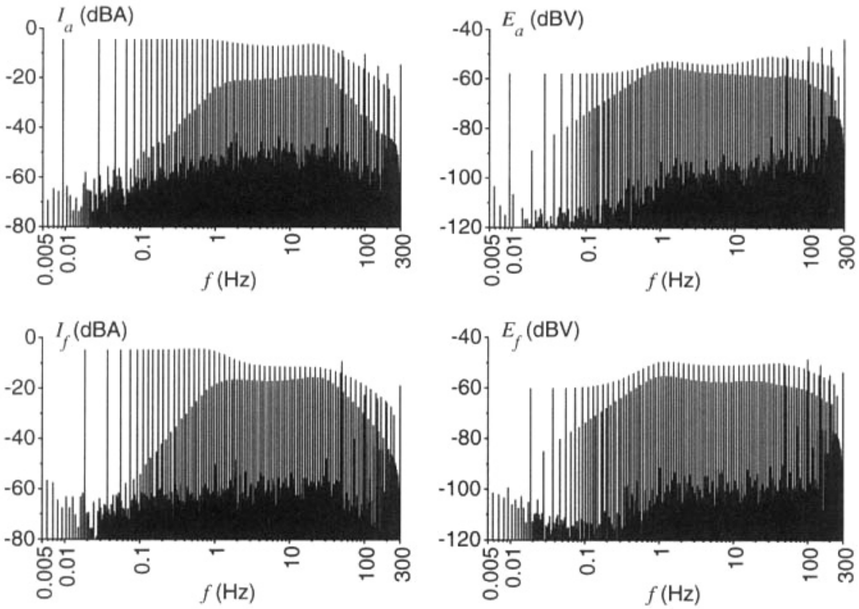


Figure 12-16. DFT spectra of the measured armature $E_a(k)$ and field $E_f(k)$ voltages and armature $I_a(k)$ and field $I_f(k)$ currents of the first MIMO experiment.

with

$$\hat{\mathrm{Y}}(k) = \begin{bmatrix} \hat{E}_a^{[1]}(k) & \hat{E}_a^{[2]}(k) \\ \hat{E}_f^{[1]}(k) & \hat{E}_f^{[2]}(k) \end{bmatrix} \quad \text{and} \quad \hat{\mathrm{U}}(k) = \begin{bmatrix} \hat{I}_a^{[1]}(k) & \hat{I}_a^{[2]}(k) \\ \hat{I}_f^{[1]}(k) & \hat{I}_f^{[2]}(k) \end{bmatrix}$$

the sample means over the $M = 8$ periods. Figure 12-17 shows the measured impedance matrix together with its 95% uncertainty bound. It follows that the signal-to-noise ratios are larger than 40 dB except at the low frequencies (<0.1 Hz) in $Z_{12}(j\omega_k)$. The FRM in Figure 12-17 is shown in per unit values; that is, it is normalized by the ratio of the nominal voltage to the nominal current.

12.6.3 Identification Results

The measurements are identified using the common denominator model (5-55). Taking into account the physical properties of the electrical machine, it has the form

$$Z(s, \theta) = \frac{\sum_{r=0}^{n_b} B_r s^r}{\sum_{r=0}^{n_a} a_r s^r} \tag{12-9}$$

with $n_a = n_b - 1$, $B_r \in \mathbb{R}^{2 \times 2}$, $B_r^T = B_r$, and B_0 a diagonal matrix (Verbeeck et al., 1999a). Using the sample mean and the sample (co)variances of the $M = 8$ measurements, the SML estimate (cost function (8-10) generalized for common denominator models (5-55)) of (12-9) is calculated for model orders $n_b = 2, 3, \ldots, 7$. Table 12-1 compares the SML cost function with its expected value in case no model errors (unmodeled dynamics, nonlinear distortions) are present. It can be seen that the cost function is much too large for all models, which indicates the presence of model errors. Using the parsimony principle, the model $n_b = 6$ is selected. Figures 12-17 and 12-18 compare the measured and estimated (model $n_b = 6$) impedance matrix. It follows that the model explains the measurements very well (no systematic errors can be detected). However, the residuals $\hat{Z}(s_k)$

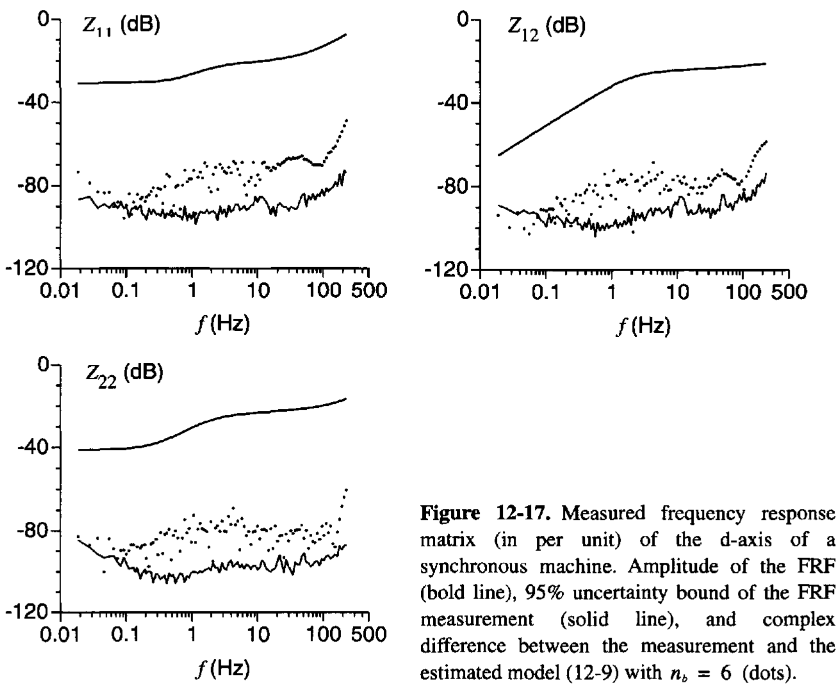


Figure 12-17. Measured frequency response matrix (in per unit) of the d-axis of a synchronous machine. Amplitude of the FRF (bold line), 95% uncertainty bound of the FRF measurement (solid line), and complex difference between the measurement and the estimated model (12-9) with $n_b = 6$ (dots).

TABLE 12-1 Sample Maximum Likelihood (SML) Cost Function as a Function of the Model Order n_b

Model (12-9) with n_b	SML Cost Function	Expected Value Cost Function (No Model Errors)
2	3.97e5	625.8
3	3.21e4	620.2
4	8.15e3	614.6
5	3.12e3	609
6	2.18e3	603.4
7	2.14e3	597.8

$-Z(s_k, \hat{\theta}_{\text{SML}}(Z))$, $k = 1, 2, \ldots, F$, are too large w.r.t. the uncertainty of the FRM measurement $\hat{Z}(s_k)$ (see Figure 12-17). This is most probably due to the stochastic nonlinear distortions created by the nonlinear magnetization behavior of the iron of the electrical machine. Indeed, the sample noise (co)variances in this experiment are calculated from eight consecutive signal periods and, hence, do not contain any information about the stochastic nonlinear contributions (see also Section 8.7).

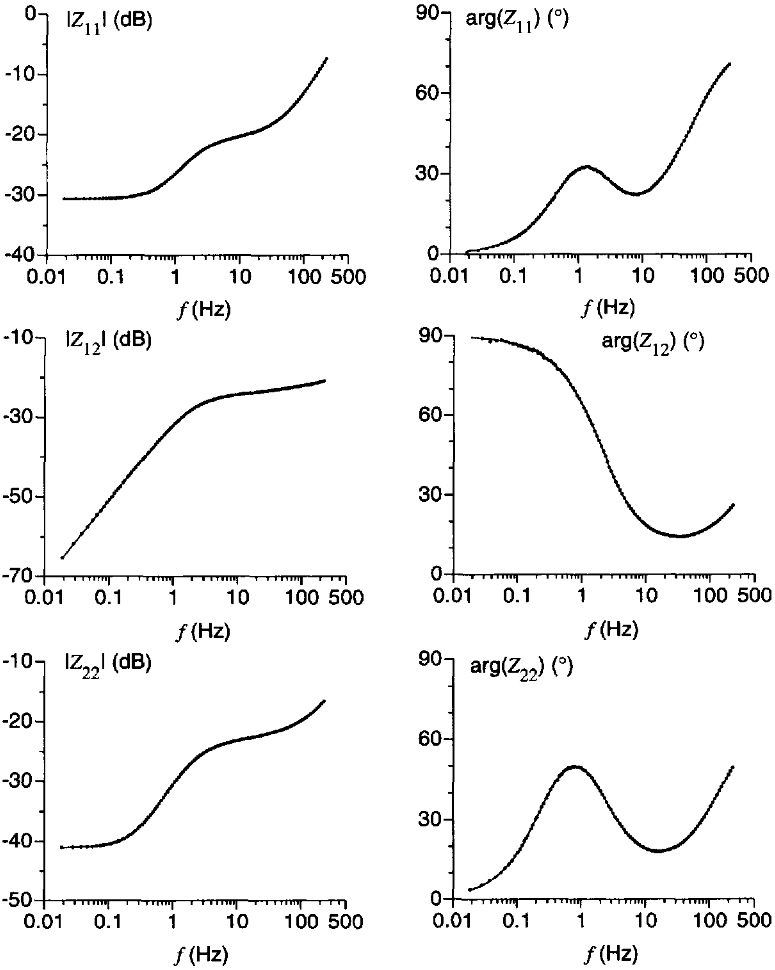


Figure 12-18. Comparison of the measured (dots) and estimated (solid line) d-axis impedance matrix (in per unit) for model (12-9) with $n_b = 6$. From left to right, amplitude and phase.

12.7 ELECTROCHEMICAL PROCESSES

Two examples of electrochemical processes are shown. Although these processes are strongly nonlinear, the linearized behavior at an operating point is typically studied.

In the first example, the AC impedance of a plumb-acid traction battery (24 V, 100 Ah) is measured using a multisine signal consisting of $F = 88$ components superimposed on a DC current of 4 A. Kamba Beya (Department ETEC of the Vrije Universiteit Brussel, Belgium) has provided us with the experimental data. The sample noise (co)variances are obtained from $M = 6$ consecutive periods. Figure 12-19 compares the measured impedance $Z(j\omega_k)$ with identified transfer functions $G(\Omega_k, \hat{\theta}_{\text{SML}}(Z))$ (rational forms of order 2/2) in s-and $\sqrt{s}$-domains, respectively. Notice the low input impedance of the traction battery. Clearly, the s-domain model is unable to explain well the amplitude and phase characteristics in the band [6 Hz, 20 Hz]. Because the same number of parameters are used in both rational forms, it is clear that the $\sqrt{s}$-domain model explains the measurements much better than the s-domain model. The poles and zeros of the identified $\sqrt{s}$-model equal $-0.2652 \pm 2.5125j$ and $-0.1113 \pm 0.6250j$, respectively, which corresponds to a stable minimum phase system (see Section 5.1).

The second example shows the AC impedance of the reduction of iron ($Fe^{3+} + e \rightarrow Fe^{2+}$) in the frequency band [100 mHz, 70 kHz]. Sven Dumortier (Department META of the Vrije Universiteit Brussel, Belgium) has provided us with the experimental data. The measurements are obtained using $F = 41$ single sine excitations. No noise variance information was available, so the NLS estimator (7-46) is used. Figure 12-20 compares the measured impedance $Z(j\omega_k)$ with the identified transfer functions $G(\Omega_k, \hat{\theta}_{\text{NLS}}(Z))$ (rational forms of order 2/2) in s- and $\sqrt{s}$-domains, respectively. Notice the high input impedance of the electrochemical process. Because the same number of parameters is used in both rational forms, it is clear that the $\sqrt{s}$-domain model outperforms the s-domain model. The poles and zeros of the identified $\sqrt{s}$-model equal $-4.0101 \pm 13.1867j$ and $-32.6908 \pm 94.4456j$, respectively, which correspond to a stable minimum phase system (see Section 5.1).

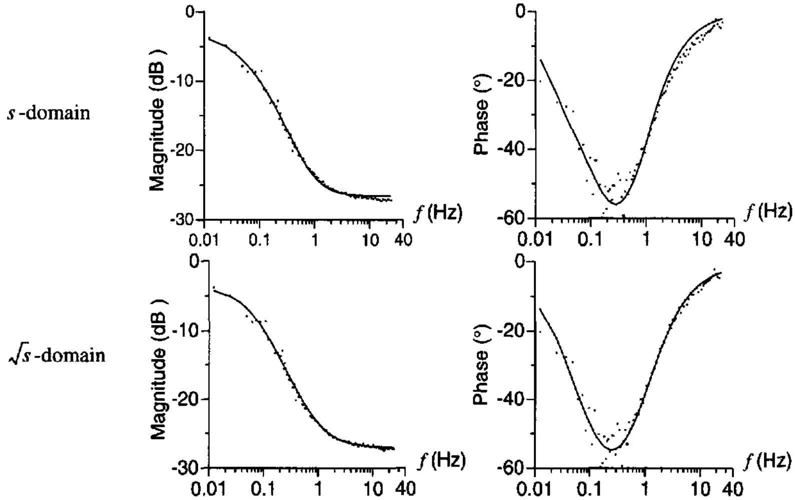


Figure 12-19. Measured (dots) and modeled (solid line) AC impedance of a traction battery. From left to right, amplitude and phase. From top to bottom, s- and $\sqrt{s}$-domain models of order 2/2.

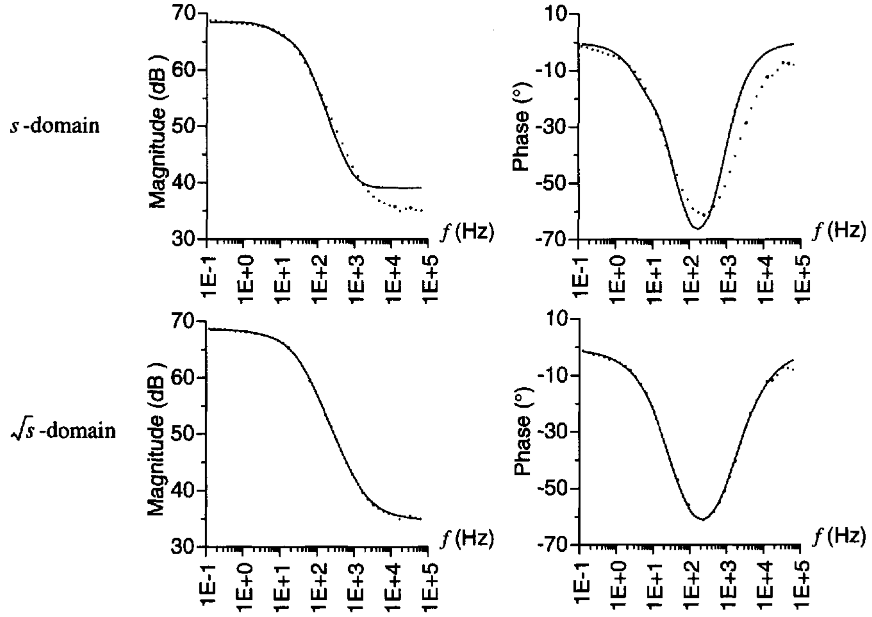


Figure 12-20. Measured (dots) and modeled (solid line) AC impedance of the reduction of iron. From left to right, amplitude and phase. From top to bottom, s- and $\sqrt{s}$-domain models of order 2/2.

12.8 MICROWAVE DEVICE

12.8.1 Measurement Setup

Device under Test. A 50 Ω semirigid coaxial cable of about 15 m (UT-141A of Suhner) terminated by a tubular bandpass filter (Mini-Circuits). The bandpass filter is the cascade of a low-pass filter with a cutoff frequency of 1.2 GHz followed by a high-pass filter with a cutoff frequency of 0.8 GHz.

Generator. A microwave sine wave generator (HP 8648B).

Acquisition. A frequency down converter (HP 85120A), followed by four HPE1437A VXI cards with antialias protection on and sampling frequency $f_s = 20$ MHz. The down conversion frequency depends on the excitation frequency and is about 19 MHz. The down conversion frequency and the sampling frequencies of the four acquisition units all stem from the same 10 MHz mother clock of the sine wave generator.

Excitation Signals. $F = 201$ single sine excitations in the band [0.4 GHz, 1.5 GHz]: $f = f_0 + k\Delta f$ with $f_0 = 0.4$ GHz and $\Delta f = 5.5$ MHz.

Measurements. $P = 5$ consecutive periods of the steady-state response of each single sine experiment are measured. The complete measurement procedure (reconnection and recalibration) is repeated $M = 5$ times.

12.8.2 Measurement Results

Wendy Van Moer and Yves Rolain (Department ELEC of the Vrije Universiteit Brussel, Belgium) have provided us with the experimental data. Figure 12-21 shows the measurement setup. At microwave frequencies, it is no longer possible to measure currents and voltages; instead, waves are measured that are linear combinations of the currents and voltages:

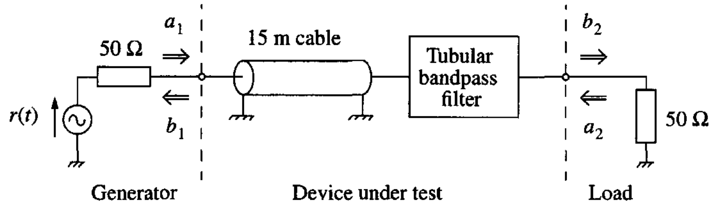


Figure 12-21. Measurement of a microwave device: a_1, a_2 are the incident waves and b_1, b_2 are the reflected waves.

$$\text{incident wave:} \quad a = (V + Z_0 I)/2$$

$$\text{reflected wave:} \quad b = (V - Z_0 I)/2$$

where Z_0 is an arbitrarily chosen impedance. Mostly, Z_0 is chosen to be equal to the characteristic impedance of the transmission lines (here 50 Ω). The incident and reflected waves are related by the scattering parameters (S-matrix)

$$\begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = S \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} \text{ with } S = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} \tag{12-10}$$

Two experiments are necessary to measure the S-matrix (see Section 2.7). In the first experiment, the setup of Figure 12-21 is used, giving the waves $a_1^{[1]}$, $a_2^{[1]}$, $b_1^{[1]}$, and $b_2^{[1]}$. In the second experiment the generator and the load in Figure 12-21 are interchanged, giving the waves $a_1^{[2]}$, $a_2^{[2]}$, $b_1^{[2]}$, and $b_2^{[2]}$. Finally, we get

$$S = \begin{bmatrix} b_1^{[1]} & b_1^{[2]} \\ b_2^{[1]} & b_2^{[2]} \end{bmatrix} \begin{bmatrix} a_1^{[1]} & a_1^{[2]} \\ a_2^{[1]} & a_2^{[2]} \end{bmatrix}^{-1} \tag{12-11}$$

Figure 12-22 compares the sample standard deviation of the S_{21}-measurement over the $P = 5$ consecutive periods (measurement noise only) with the sample standard deviation of the S_{21}-measurement over the $M = 5$ experiments (measurement, connection, and calibration noise). It follows that the connection and calibration noise are dominant over the measurement noise, which is mostly the case in microwave measurements (Verschueren et al., 1998). Because we are interested in the characteristics of the device under test and not in the (nonrepeatable) connection characteristics of the device under test and the calibration standards, the sample standard deviation used in the SML estimator (8-10) should include the connection and calibration noise. Therefore, similar to the measurement strategy of Section 8.7, the sample means of the S_{21}-measurements over the $P = 5$ signal periods are calculated before calculating the sample mean and sample variance over the $M = 5$ experiments.

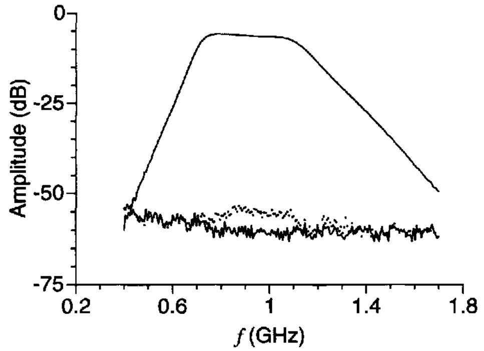


Figure 12-22. Measured S_{21} -parameter (bold line) of the microwave device together with its one sigma bound (referred to the measurement of one signal period). Uncertainty due to the measurement noise only (solid line) and uncertainty due to the measurement noise, the reconnection, and the recalibration of the measurement setup (dots).

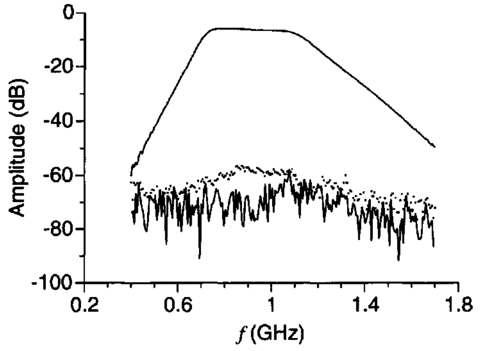


Figure 12-23. Validation of the estimated S_{21} -parameter: measured S_{21} -parameter (bold line) together with its 95% confidence bound (dots) and complex difference between the measurements and the 12th-order model (solid line).

A rational form in s with time delay (5-29) is used to model S_{21}. An initial guess of the delay is obtained via the mean slope of the unwrapped measured phase of S_{21} (see Eq. 7-142): $\tau^{(0)} = 75.86$ ns. Because this value is an overestimate (the sum of the true delay plus the delay of the rational part of the DUT), the rational part of the model is first identified for the following fixed values of the delay [73:0.1:76] ns (the numerator and denominator coefficients are free parameters, and the delay is fixed). The delay corresponding to the smallest value of the cost function is then used as improved starting value of the complete minimization problem (the numerator and denominator coefficients and the delay are free parameters). It turns out that a 12th-order model ($n_a = n_b = 12$) with a delay of $\tau = 74.23$ ns explains the measurements very well (see Figure 12-23): only 7% of the residuals (solid line in Figure 12-23) lie outside the 95% confidence bound of the measurements (dots in Figure 12-23), and the cost function $V_{\mathrm{ML}}(\hat{\theta}_{\mathrm{ML}}(Z), Z) = 235$ lies within the 95% confidence interval of the theoretical value without model errors 251 ± 53 (apply Eq. 9-15 with $M = 5$, $F = 201$, and $n_\theta = 26$).

13

Some Linear Algebra Fundamentals

Abstract: This chapter states and reviews linear algebra notations and basic concepts that are used throughout this book. In order to become familiar with these concepts, many exercises are provided at the end of the chapter. Elaborated discussions and proofs on the topic can be found in Gantmacher (1990), Golub and Van Loan (1996), Lancaster and Tismenetsky (1985), and Wilkinson (1988). Elementary matrix operations such as the sum, the inverse, the transpose, and the determinant are assumed to be known.

13.1 NOTATIONS AND DEFINITIONS

The entries of a matrix $A \in \mathbb{C}^{n \times m}$ are denoted by $A_{[i,j]}$

$$A = \begin{bmatrix} A_{[1,1]} & \cdots & A_{[1,m]} \\ \cdots & \cdots & \cdots \\ A_{[n,1]} & \cdots & A_{[n,m]} \end{bmatrix} \tag{13-1}$$

$A_{[:,k]}$ ($A_{[k,:]}$) stands for the kth column (row) of A. $A_{[i:j,k:l]}$, with $j \geq i$ and $l \geq k$, selects a $(j-i+1) \times (k-l+1)$ block of A containing rows i to j and columns k to l. Superscript T (H) is for the matrix transpose (complex conjugate transpose) and superscript −T (−H) denotes the transpose (complex conjugate transpose) of the inverse matrix. A matrix A is *Hermitian (skew Hermitian)* if $A^H = A$ ($A^H = -A$) and it is *symmetric (skew-symmetric)* if $A^T = A$ ($A^T = -A$). I_n (O_n) denotes the $n \times n$ identity (zero) matrix.

The *row (column) rank* of a matrix is the maximum number of linearly independent rows (columns). A matrix $A \in \mathbb{C}^{n \times m}$ has a *full row (column) rank* if its row (column) rank is n (m). For any matrix the column rank equals the row rank (Lancaster and Tismenetsky, 1985). This motivates the following definition of the *rank* of a matrix A:

$$\text{rank}(A) = \text{column rank of } A = \text{row rank of } A \tag{13-2}$$

System Identification: A Frequency Domain Approach. Rik Pintelon and Johan Schoukens
Copyright © 2001 Institute of Electrical and Electronics Engineers. ISBN: 0-780-36000-1

A square matrix is called *regular* if it is of full rank.

For $A \in \mathbb{C}^{n \times m}$, null($A$) is the linear subspace of $\mathbb{C}^m$ defined by $Ax = 0$

$$\text{null}(A) = \{x \in \mathbb{C}^m | Ax = 0\} \tag{13-3}$$

The *range (column space)* of a matrix $A \in \mathbb{C}^{n \times m}$ is the linear subspace of $\mathbb{C}^n$ that is obtained by making all possible linear combinations of the columns of A

$$\text{range}(A) = \{y \in \mathbb{C}^n | y = Ax, x \in \mathbb{C}^m\} \tag{13-4}$$

Note that $\text{range}(A) = \text{range}(AA^H) = (\text{null}(A^H))^{\perp}$ where superscript $\perp$ stands for the orthogonal complement of a subspace (proof: see Exercise 13.1).

The *span* of m vectors $a_1, a_2, \ldots, a_m \in \mathbb{C}^n$ is the linear subspace of $\mathbb{C}^n$ obtained by making all possible linear combinations of $a_1, a_2, \ldots, a_m$

$$\text{span}\{a_1, a_2, \ldots, a_m\} = \{x \in \mathbb{C}^n | x = \textstyle\sum_{i=1}^{m} \alpha_i a_i, \alpha_i \in \mathbb{C}\} \tag{13-5}$$

The *eigenvalues* $\lambda(A)$ of a matrix $A \in \mathbb{C}^{n \times n}$ are the roots of the *characteristic polynomial* $\det(A - \lambda I_n) = 0$, where det( ) denotes the determinant. The nonzero vectors $x \neq 0$ that satisfy $Ax = \lambda x$ are the corresponding *eigenvectors*. The eigenvalues are invariant with respect to a regular transformation $T \in \mathbb{C}^{n \times n}$ (Golub and Van Loan, 1996):

$$B = TAT^{-1} \text{ with } \det(T) \neq 0 \tag{13-6}$$

whence, after ordering of the eigenvalues, $\lambda_k(B) = \lambda_k(A)$, $(k = 1, 2, \ldots, n)$. We note that B and A are *similar*, and T is called a *similarity transformation*. Hermitian matrices have real eigenvalues (Wilkinson, 1988).

By definition, a real matrix $A \in \mathbb{R}^{n \times n}$ is *positive (semi-)definite* if for any $x \in \mathbb{R}_0^n$, the *quadratic form* $x^T A x$ is strictly positive (positive): $x^T A x > 0$ $(x^T A x \geq 0)$. Similarly, a matrix $A \in \mathbb{C}^{n \times n}$ is positive (semi-)definite if for any $x \in \mathbb{C}_0^n$, $x^H A x > 0$ $(x^H A x \geq 0)$. These conditions are satisfied if and only if all the eigenvalues of A are real and $\lambda_k(A) > 0$ $(\lambda_k(A) \geq 0)$, $k = 1, 2, \ldots, n$. Note that no symmetry is required in the real case while in the complex case the positive (semi-)definite condition implies that A is Hermitian. We shall write $A > 0$ for positive definite and $A \geq 0$ for positive semidefinite matrices.

The *right singular vectors* v (*left singular vectors* u) of a matrix $A \in \mathbb{C}^{n \times m}$ are the eigenvectors of the matrix $A^H A$ (AA^H). The *singular values* $\sigma_k(A)$, $k = 1, 2, \ldots, \min(n, m)$, are the positive square roots of the eigenvalues of $A^H A$ (AA^H) and are usually ordered from large to small values.

A matrix $U \in \mathbb{R}^{n \times n}$ $(U \in \mathbb{C}^{n \times n})$ is said to be *orthogonal (unitary)* if $U^T U = I_n$ $(U^H U = I_n)$. Orthogonal (unitary) matrices have the property $\det(U) = \pm 1$ $(|\det(U)| = 1)$ and $U^{-1} = U^T$ $(U^{-1} = U^H)$.

13.2 OPERATORS AND FUNCTIONS

Let $A_i \in \mathbb{C}^{n \times m}$, $i = 1, 2, \ldots, K$, then $\text{diag}(A_1, A_2, \ldots, A_K) \in \mathbb{C}^{nK \times mK}$ is a block diagonal matrix

$$\mathrm{diag}(A_1, A_2, \ldots, A_K) = \begin{bmatrix} A_1 & 0 & \ldots & 0 \\ 0 & A_2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & A_K \end{bmatrix} \tag{13-7}$$

The *Hermitian part* (*skew Hermitian part*) of $A \in \mathbb{C}^{n \times n}$ is $\mathrm{herm}(A) = (A + A^H)/2$ ($(A - A^H)/2$). Any matrix can be written as the sum of a Hermitian and a skew Hermitian matrix: $A = (A + A^H)/2 + (A - A^H)/2$.

Inverse of block matrices: if $\begin{bmatrix} A & D \\ C & B \end{bmatrix}^{-1}$ and A^{-1} exist, then (Kailath, 1980)

$$\begin{bmatrix} A & D \\ C & B \end{bmatrix}^{-1} = \begin{bmatrix} A^{-1} + E\Delta^{-1}F & -E\Delta^{-1} \\ -\Delta^{-1}F & \Delta^{-1} \end{bmatrix} \tag{13-8}$$

where $\Delta = B - CA^{-1}D$, $E = A^{-1}D$, and $F = CA^{-1}$.

The *trace* of $A \in \mathbb{C}^{n \times n}$ is defined as $\mathrm{tr}(A) = \sum_{k=1}^{n} A_{[k,k]}$. It is circular shift invariant: for any $A \in \mathbb{C}^{n \times m}$, $B \in \mathbb{C}^{m \times p}$, and $C \in \mathbb{C}^{p \times n}$, $\mathrm{tr}(ABC) = \mathrm{tr}(BCA)$.

For $A \in \mathbb{C}^{n \times m}$, $\mathrm{vec}(A) \in \mathbb{C}^{nm}$ is a column vector obtained by stacking the columns of A on top of each other

$$\mathrm{vec}(A) = \begin{bmatrix} A_{[:,1]} \\ A_{[:,2]} \\ \ldots \\ A_{[:,m]} \end{bmatrix} \tag{13-9}$$

13.3 NORMS

$\| \; \|$ is a matrix norm if the following properties are satisfied for all $A, B \in \mathbb{C}^{n \times m}$ and $\alpha \in \mathbb{C}$:

1. $\|A\| \geq 0$ and $\|A\| = 0 \Leftrightarrow A = 0$
2. $\|A + B\| \leq \|A\| + \|B\|$
3. $\|\alpha A\| = |\alpha| \|A\|$

The following matrix norms ($A \in \mathbb{C}^{n \times m}$) are used frequently: the Frobenius norm,

$$\|A\|_F = \sqrt{\mathrm{tr}(A^H A)} = \sqrt{\sum_{k=1}^{n} \sum_{l=1}^{m} |A_{[k,l]}|^2} \tag{13-10}$$

the 1-norm,

$$\|A\|_1 = \max_{1 \leq l \leq m} \sum_{k=1}^{n} |A_{[k,l]}| \tag{13-11}$$

the 2-norm,

$$\|A\|_2 = \max_{1 \le k \le m} \sigma_k(A) = \sigma_1(A) \tag{13-12}$$

and the ∞-norm,

$$\|A\|_\infty = \max_{1 \le k \le n} \sum_{l=1}^{m} |A_{[k,l]}| \tag{13-13}$$

The Frobenius, 1-, 2- and ∞-norms satisfy the submultiplicative property

$$\|AB\| \le \|A\|\|B\| \qquad \forall A \in \mathbb{C}^{n \times m}, \forall B \in \mathbb{C}^{m \times p} \tag{13-14}$$

Note that not all matrix norms satisfy (13-14). We also have

$$\|A\|_2 \le \|A\|_F \tag{13-15}$$

Perturbations and the inverse (Theorem 2.3.4 of Golub and Van Loan, 1996): take $A, E \in \mathbb{C}^{n \times n}$, if $\|A^{-1}E\| = r < 1$, then $\det(A + E) \ne 0$ and

$$\|(A+E)^{-1} - A^{-1}\| \le \frac{\|E\|\,\|A^{-1}\|^2}{1-r} \tag{13-16}$$

with $\|\ \|$ any matrix norm that satisfies the submultiplicative property (13-14).

13.4 DECOMPOSITIONS

13.4.1 Singular Value Decomposition

For any $A \in \mathbb{C}^{n \times m}$ with $n \ge m$ there exist $U \in \mathbb{C}^{n \times m}$ and $\Sigma, V \in \mathbb{C}^{m \times m}$ such that (Golub and Van Loan, 1996)

$$A = U \Sigma V^H \tag{13-17}$$

where $V^H V = V V^H = U^H U = I_m$ and $\Sigma = \operatorname{diag}(\sigma_1, \sigma_2, \cdots, \sigma_m)$ with $\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_m \ge 0$. The nonnegative real numbers σ_k are the *singular values* of A, and the columns $V_{[:,k]}$ and $U_{[:,k]}$ are the corresponding right and left singular vectors. (13-17) is called the *singular value decomposition* (SVD) of the matrix A. It can be expanded as

$$A = \sum_{k=1}^{m} \sigma_k U_{[:,k]} V_{[:,k]}^H \tag{13-18}$$

Taking the Hermitian transpose of (13-17) covers the case $n \le m$. A numerically stable calculation of the singular value decomposition is available in standard mathematical software packages.

The singular value decomposition (13-17) contains a lot of information about the structure of the matrix. Indeed, if $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma_m = 0$, then

$$\begin{aligned} \operatorname{rank}(A) &= r \\ \operatorname{null}(A) &= \operatorname{span}\{V_{[:,r+1]}, V_{[:,r+2]}, \ldots, V_{[:,m]}\} \\ \operatorname{range}(A) &= \operatorname{span}\{U_{[:,1]}, U_{[:,2]}, \ldots, U_{[:,r]}\} \end{aligned} \tag{13-19}$$

If $A \in \mathbb{C}^{m \times n}$ with $n \geq m$, then it can be decomposed into singular values as

$$A = V\Sigma U^H \tag{13-20}$$

with $V \in \mathbb{C}^{m \times m}$, $\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_m)$, and $U \in \mathbb{C}^{n \times m}$ (proof: apply (13-17) to A^H). If $\operatorname{rank}(A) = r$ then

$$\begin{aligned} \operatorname{null}(A) &= (\operatorname{span}\{U_{[:,1]}, U_{[:,2]}, \ldots, U_{[:,r]}\})^{\perp} \\ \operatorname{range}(A) &= \operatorname{span}\{V_{[:,1]}, V_{[:,2]}, \ldots, V_{[:,r]}\} \end{aligned} \tag{13-21}$$

If $\operatorname{rank}(A) = n$ then $\operatorname{null}(A) = \{x \in \mathbb{C}^n | x = U_{\perp}y, y \in \mathbb{C}^{n-m}\}$ with $U_{\perp} \in \mathbb{C}^{n \times (n-m)}$ the orthogonal complement of U: $U^H U_{\perp} = 0$ and $U_{\perp}^H U_{\perp} = I_{n-m}$.

The *condition number* of a matrix $A \in \mathbb{C}^{n \times m}$ is defined as the ratio of the largest singular value to the smallest singular value $\kappa(A) = \sigma_1/\sigma_m$. For regular square matrices $m = n$ it is a measure of the sensitivity of the solution of the linear system $Ax = b$, with $b \in \mathbb{C}^n$, to perturbations in A and b. It can be shown that (Golub and Van Loan, 1996)

$$\frac{\|\Delta x\|_2}{\|x\|_2} \leq \kappa(A)\left(\frac{\|\Delta A\|_2}{\|A\|_2} + \frac{\|\Delta b\|_2}{\|b\|_2}\right) \tag{13-22}$$

where Δ denotes the perturbation. For rectangular matrices $m > n$ of full rank, it is a measure of the sensitivity of the least squares solution $x_{\mathrm{LS}} = (A^H A)^{-1} A^H b$ of the overdetermined set of equations $Ax \approx b$, with $b \in \mathbb{C}^m$, to perturbations in A and b (see Section 13.13). For singular matrices $\kappa(A) = \infty$. If $\kappa(A)$ is large ($\log_{10}(\kappa(A))$ is of the order of the number of significant digits used in the calculations), then A is said to be *ill-conditioned*. Unitary (orthogonal) matrices are perfectly conditioned ($\kappa = 1$), while matrices with small condition number ($\kappa \sim 1$) are said to be *well-conditioned*.

13.4.2 Generalized Singular Value Decomposition

Let $A \in \mathbb{C}^{n \times m}$ with $n \geq m$, $B \in \mathbb{C}^{p \times m}$ with $p \geq m$ and $\operatorname{rank}([A^T B^T]^T) = m$ then there exist $U_A \in \mathbb{C}^{n \times m}$, $U_B \in \mathbb{C}^{p \times m}$, and a regular $X \in \mathbb{C}^{m \times m}$ such that (Golub and Van Loan, 1996; Paige, 1986)

$$A = U_A \Sigma_A X^{-1} \qquad B = U_B \Sigma_B X^{-1} \tag{13-23}$$

where $U_A^H U_A = U_B^H U_B = I_m$, $\Sigma_A = \operatorname{diag}(\alpha_1, \alpha_2, \ldots, \alpha_m)$, and $\Sigma_B = \operatorname{diag}(\beta_1, \beta_2, \ldots, \beta_m)$, with $\alpha_k \geq 0$, $\beta_k \geq 0$, and $\alpha_k^2 + \beta_k^2 = 1$. The ratios $\sigma_k(A, B) = \alpha_k/\beta_k$ are the *generalized singular values* of the matrix pair (A, B), and the columns $X_{[:,k]}$ are the corresponding *generalized right singular vectors*. If $B \in \mathbb{C}^{m \times m}$ is regular then the generalized singular values of (A, B) are equal to the singular values of AB^{-1}: $\sigma_k(A, B) = \sigma_k(AB^{-1})$.

The generalized singular value decomposition (13-23) can be used to solve the generalized eigenvalue problem

$$A^H Ax = \lambda B^H Bx \tag{13-24}$$

without forming $A^H A$ and $B^H B$. It is easy to verify that $x = X_{[:,k]}$ and $\lambda = \alpha_k^2/\beta_k^2$, $k = 1, 2, \ldots, m$, are the solutions of (13-24) (see Exercise 13.18). Fortran and C versions of the generalized singular value decomposition are available in public domain software (Anderson et al., 1992; Bai and Demmel, 1993). For $B^H B = I_m$ the generalized eigenvalue problem reduces to an ordinary eigenvalue problem that can be solved using the singular value decomposition (13-20) of A. $x = V_{[:,k]}$ and $\lambda = \sigma_k^2$, $k = 1, 2, \ldots, m$, are then the solutions of (13-24) (see Exercise 13.16).

13.4.3 The QR Factorization

The QR factorization of $A \in \mathbb{C}^{n\times m}$ with $n \geq m$ is given by

$$A = QR \tag{13-25}$$

where $Q \in \mathbb{C}^{n\times n}$ satisfies $Q^H Q = I_n$, and R is an upper triangular matrix (Golub and Van Loan, 1996). If A is of full rank then the QR factorization has the following properties: Q and R are unique, the diagonal elements of R are positive, and range(A) = range(Q).

13.4.4 Square Root of a Positive (Semi-)Definite Matrix

Any positive (semi-)definite matrix $A \in \mathbb{C}^{n\times n}$ can be decomposed as

$$A = \Lambda^H \Lambda \quad \text{or} \quad A = SS^H \tag{13-26}$$

where $\Lambda^H, S \in \mathbb{C}^{n\times m}$ and $m \geq \text{rank}(A)$. Λ, S are *square roots* of A that are not unique and often have no analytic solution. Numerical, $n \times n$, solutions can be calculated using the singular value decomposition. For example, if $A = V\Sigma V^H$ then any $\Lambda = T\Sigma^{1/2}V^H$ with $T \in \mathbb{C}^{n\times n}$ a unitary matrix satisfies (13-26). Choosing $T = V$ gives a Hermitian solution that motivates the following notation:

$$A = A^{1/2}A^{1/2} \tag{13-27}$$

with $A^{1/2} = V\Sigma^{1/2}V^H$. In the real case, similar results apply to symmetric positive (semi-) definite matrices.

13.5 MOORE-PENROSE PSEUDOINVERSE

For any matrix $A \in \mathbb{C}^{n\times m}$ there exists a unique generalized inverse $A^+ \in \mathbb{C}^{m\times n}$, also called a Moore-Penrose pseudoinverse, that satisfies the four Moore-Penrose conditions (Ben-Israel and Greville, 1974)

1. $AA^+A = A$
2. $A^+AA^+ = A^+$
3. $(AA^+)^H = AA^+$
4. $(A^+A)^H = A^+A$

For regular square matrices it is clear that $A^+ = A^{-1}$. The pseudoinverse can be constructed using, for example, the singular value decomposition (Golub and Van Loan, 1996). If $\text{rank}(A) = r$ then

$$A^+ = V\Sigma^+U^H \text{ with } \Sigma^+ = \text{diag}(\sigma_1^{-1}, \sigma_2^{-1}, \dots, \sigma_r^{-1}, 0, \dots, 0) \tag{13-28}$$

Using (13-28) it can easily be shown that for every matrix A, $(A^+)^+ = A$, $(A^+)^H = (A^H)^+$, and $A^+ = (A^HA)^+A^H = A^H(AA^H)^+$.

Although the properties of the pseudoinverse very much resemble those of the inverse, in general $(AB)^+ \neq B^+A^+$. If the matrices $A \in \mathbb{C}^{n\times r}$ and $B \in \mathbb{C}^{r\times m}$ with $r \le \min(n, m)$ are of full rank then $(AB)^+ = B^+A^+$ (Ben-Israel and Greville, 1974).

Theorem 13.1: For any $C \in \mathbb{C}^{n\times m}$ with $n \ge m$ and $\text{rank}(C) = m$ and any $B \in \mathbb{C}^{m\times m}$ with $\text{rank}(B) = m$ we have $C^H(CBC^H)^+C = B^{-1}$.

Proof. Apply condition 1 with $A = CBC^H$

$$CBC^H(CBC^H)^+CBC^H = CBC^H \tag{13-29}$$

Left multiplication with C^H and right multiplication with C of (13-29) results in

$$C^HCBC^H(CBC^H)^+CBC^HC = C^HCBC^HC \tag{13-30}$$

where C^HC and B are, by assumption, regular matrices. Left division by C^HCB and right division by BC^HC of (13-30) proves the theorem. □

13.6 IDEMPOTENT MATRICES

By definition, an *idempotent matrix* $P \in \mathbb{C}^{n\times n}$ satisfies $P^2 = P$. If P is an idempotent matrix then (Lancaster and Tismenetsky, 1985)

1. $\lambda_k(P) = 1$, $k = 1, 2, \dots, \text{rank}(P)$ and $\lambda_k(P) = 0$, $k = \text{rank}(P) + 1, \dots, n$.
2. $I_n - P$ is also a idempotent matrix
3. $\text{range}(I_n - P) = \text{null}(P)$ and $\text{null}(I_n - P) = \text{range}(P)$
4. $\text{null}(P) + \text{range}(P) = \mathbb{C}^n$ and $\text{null}(P) \cap \text{range}(P) = \{0\}$

where $\mathbb{A} + \mathbb{B} = \mathbb{D}$ means that for each element $d \in \mathbb{D}$ there exist an element $a \in \mathbb{A}$ and an element $b \in \mathbb{B}$ such that $d = a + b$.

An idempotent matrix P can be interpreted geometrically as a projection on range(P) along null(P). This projection is orthogonal for Hermitian idempotent matrices P: $\text{range}(P) = (\text{null}(P))^\perp$. Note that a Hermitian idempotent matrix is positive (semi-)definite.

Theorem 13.2: Let $P, Q \in \mathbb{C}^{n \times n}$ be Hermitian idempotent matrices with $\text{rank}(P) = r$ and $\text{rank}(Q) = n - r$, respectively. If $QP = 0$ then $Q + P = I_n$.

Proof. Because P and Q are Hermitian, it follows from $QP = 0$ that $PQ = 0$. Take any eigenvector v_k of P with $\lambda_k(P) = 1$, $k = 1, 2, \ldots, r$. Right multiplication of $QP = 0$ by v_k gives $Qv_k = 0$ and, hence, $\lambda_k(Q) = 0$, $k = 1, 2, \ldots, r$. Using $PQ = 0$ it follows, similarly, that any eigenvector v_k of Q with $\lambda_k(Q) = 1$ is an eigenvector of P with $\lambda_k(P) = 0$, $k = r + 1, \ldots, n$. Because the eigenvectors of a Hermitian positive (semi-)definite matrix form an orthonormal basis (Exercise 13.19), we have

$$P = V\begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix}V^H \text{ and } Q = V\begin{bmatrix} 0 & 0 \\ 0 & I_{n-r} \end{bmatrix}V^H \tag{13-31}$$

where $V = [v_1 v_2 \ldots v_n]$, so that $Q + P = I_n$. □

13.7 KRONECKER ALGEBRA

This section gives some basic properties of the Kronecker product of matrices. A complete overview and elaborated proofs can be found in Brewer (1978) and Lancaster and Tismenetsky (1985).

Consider the following matrices: $A, H \in \mathbb{C}^{p \times q}$, $B \in \mathbb{C}^{s \times t}$, $C \in \mathbb{C}^{r \times l}$, $D \in \mathbb{C}^{q \times s}$, $G \in \mathbb{C}^{t \times u}$, $N \in \mathbb{C}^{n \times n}$, and $M \in \mathbb{C}^{m \times m}$. The Kronecker product of two matrices is defined as

$$A \otimes B = \begin{bmatrix} A_{[1,1]}B & A_{[1,2]}B & \ldots & A_{[1,q]}B \\ A_{[2,1]}B & A_{[2,2]}B & \ldots & A_{[2,q]}B \\ \ldots & \ldots & \ldots & \ldots \\ A_{[p,1]}B & \ldots & \ldots & A_{[p,q]}B \end{bmatrix} \in \mathbb{C}^{ps \times qt} \tag{13-32}$$

It has the following properties:

$$(A \otimes B) \otimes C = A \otimes (B \otimes C) \tag{13-33}$$

$$(A + H) \otimes B = A \otimes B + H \otimes B \tag{13-34}$$

$$(A \otimes B)^T = A^T \otimes B^T \tag{13-35}$$

$$(A \otimes B)(D \otimes G) = (AD) \otimes (BG) \tag{13-36}$$

$$(N \otimes M)^{-1} = N^{-1} \otimes M^{-1} \tag{13-37}$$

$$\text{vec}(ADB) = (B^T \otimes A)\text{vec}(D) \tag{13-38}$$

$$\|A \otimes B\| = \|A\|\|B\| \tag{13-39}$$

where $\| \; \|$ denotes the 1-, 2-, ∞-, and Frobenius norm.

13.8 ISOMORPHISM BETWEEN COMPLEX AND REAL MATRICES

Any complex matrix $A \in \mathbb{C}^{n \times m}$ can be transformed into a real matrix $A_{\text{Re}} \in \mathbb{R}^{(2n) \times (2m)}$ through

$$A_{\text{Re}} = \begin{bmatrix} \text{Re}(A) & -\text{Im}(A) \\ \text{Im}(A) & \text{Re}(A) \end{bmatrix} \tag{13-40}$$

Lemma 13.3

$$A = B + C \Leftrightarrow A_{\text{Re}} = B_{\text{Re}} + C_{\text{Re}} \tag{13-41}$$

$$A = BC \Leftrightarrow A_{\text{Re}} = B_{\text{Re}} C_{\text{Re}} \tag{13-42}$$

$$A = B^{-1} \Leftrightarrow A_{\text{Re}} = B_{\text{Re}}^{-1} \tag{13-43}$$

$$A = B^{+} \Leftrightarrow A_{\text{Re}} = B_{\text{Re}}^{+} \tag{13-44}$$

$$A = B^{H} \Leftrightarrow A_{\text{Re}} = B_{\text{Re}}^{T} \tag{13-45}$$

$$\det(A_{\text{Re}}) = |\det(A)|^2 \tag{13-46}$$

$$\text{rank}(A_{\text{Re}}) = 2\text{rank}(A) \tag{13-47}$$

provided that the matrix dimensions are appropriate. Moreover, if $A \in \mathbb{C}^{n \times n}$ is unitary (positive definite) then $A_{\text{Re}} \in \mathbb{R}^{(2n) \times (2n)}$ is orthogonal (symmetric and positive definite) and vice versa.

Proof. Exercises 13.34 to 13.38. □

Lemma 13.3 defines an isomorphism between the complex $n \times m$ matrices and the real $(2n) \times (2m)$ matrices. Using (13-42), (13-43), and (13-45), the relationships between the eigenvectors, eigenvalues, singular values, and singular vectors of A and A_{Re} are readily obtained. For example, if the singular value decomposition of A is given by $U \Sigma V^H$, then that of A_{Re} equals $U_{\text{Re}} \Sigma_{\text{Re}} V_{\text{Re}}^T$. Similarly, if the generalized singular value decomposition of the matrix pair (A, B) is given by $A = U_A \Sigma_A X^{-1}$, $B = U_B \Sigma_B X^{-1}$, then that of $(A_{\text{Re}}, B_{\text{Re}})$ is given by $A_{\text{Re}} = (U_A)_{\text{Re}} (\Sigma_A)_{\text{Re}} X_{\text{Re}}^{-1}$, $B_{\text{Re}} = (U_B)_{\text{Re}} (\Sigma_B)_{\text{Re}} X_{\text{Re}}^{-1}$, and vice versa.

Another transformation between complex $(A \in \mathbb{C}^{n \times m})$ and real $(A_{\text{re}} \in \mathbb{R}^{(2n) \times m})$ matrices is given by

$$A_{\text{re}} = \begin{bmatrix} \text{Re}(A) \\ \text{Im}(A) \end{bmatrix} \tag{13-48}$$

It has the following properties.

Lemma 13.4: Take any $A \in \mathbb{C}^{n \times m}$, $B \in \mathbb{C}^{n \times p}$, $X \in \mathbb{C}^{p \times m}$, and $Y \in \mathbb{R}^{p \times m}$

$$A = BX \Leftrightarrow A_{\mathrm{re}} = B_{\mathrm{Re}} X_{\mathrm{re}} \tag{13-49}$$

$$A = BY \Leftrightarrow A_{\mathrm{re}} = B_{\mathrm{re}} Y \tag{13-50}$$

$$\mathrm{Re}(A^H B) = A_{\mathrm{re}}^T B_{\mathrm{re}} \tag{13-51}$$

Proof. Exercise 13.39. □

Lemmas 13.3 and 13.4 are very useful to generalize results obtained for the real-valued case to the complex-valued case. This is illustrated in the following example.

Example 13.5: Consider the following expression:

$$\frac{1}{2} x^T C_x^{-1} x \tag{13-52}$$

where $x \in \mathbb{R}^n$ is a real-valued random vector and $C_x = \mathrm{Cov}(x) \in \mathbb{R}^{n \times n}$ is the corresponding covariance matrix. To obtain the result for the complex-valued case ($x \in \mathbb{C}^n$), x and C_x are replaced in (13-52) by x_{re} and $C_{x_{\mathrm{re}}}$, respectively. Assuming that $x \in \mathbb{C}^n$ is a circular complex random vector (see Section 14.1), $\mathrm{Cov}(\mathrm{Re}(x)) = \mathrm{Cov}(\mathrm{Im}(x))$ and $\mathrm{Cov}(\mathrm{Re}(x), \mathrm{Im}(x)) = -\mathrm{Cov}(\mathrm{Im}(x), \mathrm{Re}(x))$, we have $C_{x_{\mathrm{re}}} = 0.5(C_x)_{\mathrm{Re}}$. Using Lemmas 13.3 and 13.4, we find

$$\begin{aligned} \frac{1}{2} x_{\mathrm{re}}^T C_{x_{\mathrm{re}}}^{-1} x_{\mathrm{re}} &= x_{\mathrm{re}}^T (C_x^{-1})_{\mathrm{Re}} x_{\mathrm{re}} && \text{(property (13-43))} \\ &= x_{\mathrm{re}}^T (C_x^{-1} x)_{\mathrm{re}} && \text{(property (13-49))} \\ &= \mathrm{Re}(x^H C_x^{-1} x) && \text{(property (13-51))} \\ &= x^H C_x^{-1} x && (C_x \text{ is positive definite}) \end{aligned}$$

We conclude that $\frac{1}{2} x_{\mathrm{re}}^T C_{x_{\mathrm{re}}}^{-1} x_{\mathrm{re}} = x^H C_x^{-1} x$. □

13.9 DERIVATIVES

13.9.1 Derivatives of Functions and Vectors w.r.t. a Vector

The first- and second-order derivatives of an analytic function $f(x) \in \mathbb{C}$ (vector function $F(x) \in \mathbb{C}^n$) with respect to a vector $x \in \mathbb{C}^m$ are defined as

$$\frac{\partial f(x)}{\partial x} \in \mathbb{C}^{1 \times m} \text{ with } \left(\frac{\partial f(x)}{\partial x}\right)_{[1,k]} = \frac{\partial f(x)}{\partial x_{[k]}} \tag{13-53}$$

$$\frac{\partial^2 f(x)}{\partial x^2} \in \mathbb{C}^{m\times m} \text{ with } \left(\frac{\partial^2 f(x)}{\partial x^2}\right)_{[k,l]} = \frac{\partial^2 f(x)}{\partial x_{[k]}\partial x_{[l]}} \tag{13-54}$$

$$\frac{\partial F(x)}{\partial x} \in \mathbb{C}^{n\times m} \text{ with } \left(\frac{\partial F(x)}{\partial x}\right)_{[k,l]} = \frac{\partial F_{[k]}(x)}{\partial x_{[l]}} \tag{13-55}$$

Let $g(x) \in \mathbb{C}^q$ be an analytic vector function of $x \in \mathbb{C}^m$ and $A \in \mathbb{C}^{n\times m}$, $B \in \mathbb{C}^{m\times m}$. Using definitions (13-53), (13-54), and (13-55) it can be verified that

$$\frac{\partial Ax}{\partial x} = A \qquad \frac{\partial}{\partial x}\left(\frac{x^TBx}{2}\right) = x^T\left(\frac{B+B^T}{2}\right) \qquad \frac{\partial^2}{\partial x^2}\left(\frac{x^TBx}{2}\right) = \frac{B+B^T}{2} \tag{13-56}$$

$$\frac{\partial}{\partial x}\left(\frac{g^T(x)g(x)}{2}\right) = g^T(x)\frac{\partial g(x)}{\partial x} \tag{13-57}$$

$$\frac{\partial^2}{\partial x^2}\left(\frac{g^T(x)g(x)}{2}\right) = \left(\frac{\partial g(x)}{\partial x}\right)^T\left(\frac{\partial g(x)}{\partial x}\right) + \sum\nolimits_{k=1}^{q} g_{[k]}(x)\frac{\partial^2 g_{[k]}}{\partial x^2} \tag{13-58}$$

The derivative of a real function $f(x, \bar{x}) \in \mathbb{R}$ with respect to the real and imaginary parts of the vector $x \in \mathbb{C}^m$ can be found using the chain rule and symbol derivation w.r.t. x and $\bar{x}$ ($f(x, \bar{x})$ is NOT an analytic function of x)

$$\begin{aligned}\frac{\partial f(x,\bar{x})}{\partial \mathrm{Re}(x)} &= \frac{\partial f(x,\bar{x})}{\partial x}\frac{\partial x}{\partial \mathrm{Re}(x)} + \frac{\partial f(x,\bar{x})}{\partial \bar{x}}\frac{\partial \bar{x}}{\partial \mathrm{Re}(x)} = \frac{\partial f(x,\bar{x})}{\partial x} + \frac{\partial f(x,\bar{x})}{\partial \bar{x}}\\ \frac{\partial f(x,\bar{x})}{\partial \mathrm{Im}(x)} &= \frac{\partial f(x,\bar{x})}{\partial x}\frac{\partial x}{\partial \mathrm{Im}(x)} + \frac{\partial f(x,\bar{x})}{\partial \bar{x}}\frac{\partial \bar{x}}{\partial \mathrm{Im}(x)} = j\left(\frac{\partial f(x,\bar{x})}{\partial x} - \frac{\partial f(x,\bar{x})}{\partial \bar{x}}\right)\end{aligned} \tag{13-59}$$

Because $f(x, \bar{x})$ is real, $\partial f(x, \bar{x})/\partial \bar{x} = \overline{\partial f(x, \bar{x})/\partial x}$, so that (13-59) can be written as

$$\left(\frac{\partial f(x,\bar{x})}{\partial x_{\mathrm{re}}}\right)^T = 2\left[\left(\frac{\partial f(x,\bar{x})}{\partial x}\right)^H\right]_{\mathrm{re}} \tag{13-60}$$

13.9.2 Derivative of a Function w.r.t. a Matrix

The derivative of an analytic function $f(A) \in \mathbb{C}$ with respect to a matrix $A \in \mathbb{C}^{n\times m}$ is defined as

$$\frac{\partial f(A)}{\partial A} \in \mathbb{C}^{n\times m} \text{ with } \left(\frac{\partial f(A)}{\partial A}\right)_{[k,l]} = \frac{\partial f(A)}{\partial A_{[k,l]}} \tag{13-61}$$

Using definition (13-61) it can be verified that

$$\frac{\partial \mathrm{tr}(BA)}{\partial A} = B^T \tag{13-62}$$

$$\frac{\partial \mathrm{tr}(CABA^TC^T)}{\partial A} = C^TCA(B+B^T) \tag{13-63}$$

$$\frac{\partial \ln(\det(A))}{\partial A} = A^{-T} \tag{13-64}$$

$$\frac{\partial \mathrm{tr}(BA^{-1})}{\partial A} = -(A^{-1}BA^{-1})^T \tag{13-65}$$

provided that the matrix dimensions are appropriate.

Following the same procedure as in Section 13.9.1, the derivative of a real function $f(A, \bar{A}) \in \mathbb{R}$ with respect to the real and imaginary parts of the matrix $A \in \mathbb{C}^{n \times m}$ can be found through symbolic derivation ($f(A, \bar{A})$ is *not* an analytic function of A)

$$\frac{\partial f(A, \bar{A})}{\partial A_{\mathrm{re}}} = 2\left(\frac{\partial f(A, \bar{A})}{\partial \bar{A}}\right)_{\mathrm{re}} \tag{13-66}$$

Take, for example, $f(A, \bar{A}) = \mathrm{tr}(CABA^HC^H)$ with $B^H = B$, then

$$\frac{\partial \mathrm{tr}(CABA^HC^H)}{\partial A_{\mathrm{re}}} = 2(C^HCAB)_{\mathrm{re}} \tag{13-67}$$

Note that the derivative of a function w.r.t. a vector in Section 13.9.1 corresponds to the derivative of a function w.r.t. a row in Section 13.9.2.

13.10 INNER PRODUCT

Consider a finite-dimensional linear space $\mathbb{L}$ over the field $\mathbb{F}$ of real or complex numbers ($\mathbb{F} = \mathbb{R}$ or $\mathbb{F} = \mathbb{C}$), and let $x, y \in \mathbb{L}$. The function $\langle x, y \rangle$ from $\mathbb{L} \times \mathbb{L}$ to $\mathbb{F}$ is an *inner product* on the linear space $\mathbb{L}$ if the following properties are satisfied for all $x, y \in \mathbb{L}$ and $\alpha, \beta \in \mathbb{F}$:

1. $\langle x, x \rangle \geq 0$ and $\langle x, x \rangle = 0 \Leftrightarrow x = 0$
2. $\langle \alpha x + \beta y, z \rangle = \alpha \langle x, z \rangle + \beta \langle y, z \rangle$
3. $\langle x, y \rangle = \overline{\langle y, x \rangle}$

These properties are known as the positivity, linearity in the first argument, and Hermitian symmetry ($\mathbb{F} = \mathbb{C}$) or symmetry ($\mathbb{F} = \mathbb{R}$), respectively. The inner product also defines a *norm* on the space $\mathbb{L}$: $\sqrt{\langle x, x \rangle} = \|x\|$ (Lancaster and Tismenetsky, 1985). Two nonzero elements $x, y \in \mathbb{L}$ are *orthogonal* if $\langle x, y \rangle = 0$.

Lemma 13.6: Let $\mathbb{P}_m(\mathbb{R})$ be the linear space of real polynomials (= polynomials with real coefficients) of order smaller than or equal to m. Take any $p(x), q(x) \in \mathbb{P}_m(\mathbb{R})$ and define

$$\langle p(x), q(x) \rangle = \mathrm{Re}\left(\textstyle\sum_{k=1}^{n} [w_{1k}p(x_k) + w_{2k}\bar{p}(x_k)][\bar{w}_{1k}\bar{q}(x_k) + \bar{w}_{2k}q(x_k)]\right) \tag{13-68}$$

where $w_{1k}, w_{2k} \in \mathbb{C}$ are the weights and $x_k \in \mathbb{C}$ the grid points. (13-68) defines an inner product if and only if the matrix $J \in \mathbb{C}^{n \times (m+1)}$ with $J_{[k,r]} = w_{1k}x_k^{r-1} + w_{2k}\bar{x}_k^{r-1}$, $k = 1, 2, \ldots, n$ and $r = 1, 2, \ldots, m+1$ satisfies the rank condition

$$\text{rank}(J_{\text{re}}) = m + 1 \tag{13-69}$$

Proof. The linearity (only real linear combinations are considered for real polynomials) and symmetry of (13-68) follow directly. To show that the positivity condition is satisfied, we rewrite (13-68) using $p(x) = \sum_{r=0}^{m} p_r x^r$ and $p = [p_0 p_1 \cdots p_m]^T \in \mathbb{R}^{m+1}$ as

$$\langle p(x), p(x) \rangle = \text{Re}(p^T J^H J p) = p^T \text{Re}(J^H J) p = p^T J_{\text{re}}^T J_{\text{re}} p \tag{13-70}$$

The last equivalence is due to property (13-51). Under rank condition (13-69), the matrix $J_{\text{re}}^T J_{\text{re}}$ is positive definite and, hence, $\langle p(x), p(x) \rangle = 0$ if and only if $p = 0$. □

Lemma 13.7: Let $\mathbb{P}_m^2(\mathbb{R})$ be the linear space of real 2 by 1 vector polynomials of order smaller than or equal to m. If $p(x) \in \mathbb{P}_m^2(\mathbb{R})$, then $p_{[i]}(x) \in \mathbb{P}_m(\mathbb{R})$, $i = 1, 2$. Take any $p(x), q(x) \in \mathbb{P}_m^2(\mathbb{R})$ and define

$$\langle p(x), q(x) \rangle = \text{Re}(\textstyle\sum_{k=1}^{n} q^H(x_k) W_k^H W_k p(x_k)) \tag{13-71}$$

with $W_k \in \mathbb{C}^2$ the weighting matrices and $x_k \in \mathbb{C}$ the grid points. Define the 2 by 1 vector polynomials $E_{2r}^T(x) = [x^r \; 0]$ and $E_{2r+1}^T(x) = [0 \; x^r]$, $r = 0, 1, \ldots, m$. (13-71) defines an inner product if and only if the matrix $J \in \mathbb{C}^{n \times (2m+2)}$ with $J_{[2k-1:2k, r]} = W_k E_{r-1}(x_k)$, $k = 1, 2, \ldots, n$ and $r = 1, 2, \ldots, 2m+2$ satisfies the rank condition

$$\text{rank}(J_{\text{re}}) = 2m + 2 \tag{13-72}$$

Proof. The linearity (only real linear combinations are considered for real polynomials) and symmetry of (13-71) follow directly. The proof of the positivity condition is along the lines of Lemma 13.6. Using $p(x) = \sum_{r=0}^{2m+1} p_r E_r(x)$ and $p = [p_0 p_1 \cdots p_{2m+1}]^T \in \mathbb{R}^{2m+2}$, (13-71) becomes $\langle p(x), p(x) \rangle = p^T J_{\text{re}}^T J_{\text{re}} p$. Under the rank condition (13-72) $J_{\text{re}}^T J_{\text{re}}$ is positive definite so that $\langle p(x), p(x) \rangle = 0$ if and only if $p = 0$. □

Lemma 13.8: Let $\mathbb{P}_m(\mathbb{C})$ be the linear space of complex polynomials (= polynomials with complex coefficients) of order smaller than or equal to m. Take any $p(x), q(x) \in \mathbb{P}_m(\mathbb{C})$ and define

$$\langle p(x), q(x) \rangle = \textstyle\sum_{k=1}^{n} |w_k|^2 p(x_k) \bar{q}(x_k) \tag{13-73}$$

with $w_k \in \mathbb{C}$ the weights and $x_k \in \mathbb{C}$ the grid points. (13-73) is an inner product if and only if the matrix $J \in \mathbb{C}^{n \times (m+1)}$ with $J_{[k,r]} = w_k x_k^{r-1}$, $k = 1, 2, \ldots, n$ and $r = 1, 2, \ldots, m+1$ has rank $m + 1$.

Proof. Similar to Lemma 13.6. □

Lemma 13.9: Let $\mathbb{P}_m^2(\mathbb{C})$ be the linear space of complex 2 by 1 vector polynomials (= polynomials with complex coefficients) of order smaller than or equal to m. Take any $p(x), q(x) \in \mathbb{P}_m^2(\mathbb{C})$ and define

$$\langle p(x), q(x)\rangle = \sum_{k=1}^{n} q^H(x_k) W_k^H W_k p(x_k) \tag{13-74}$$

with $W_k \in \mathbb{C}^2$ the weighting matrices and $x_k \in \mathbb{C}$ the grid points. Define the 2 by 1 vector polynomials $E_{2r}^T(x) = [x^r \; 0]$ and $E_{2r+1}^T(x) = [0 \; x^r]$, $r = 0, 1, \dots, m$. (13-73) is an inner product if and only if the matrix $J \in \mathbb{C}^{n \times (2m+2)}$ with $J_{[2k-1\,:\,2k,\, r]} = W_k E_{r-1}(x_k)$, $k = 1, 2, \dots, n$ and $r = 1, 2, \dots, 2m+2$ has rank $2m+2$.

Proof. Similar to Lemma 13.7. □

Note that Lemmas 13.7 and 13.9 can easily be generalized to vector polynomials with more than two entries.

13.11 GRAM-SCHMIDT ORTHOGONALIZATION

The *Gram-Schmidt orthogonalization* calculates an orthonormal set $\{y_1, y_2, \dots, y_n\}$ from a given linear independent set $\{x_1, x_2, \dots, x_n\}$ with the property

$$\text{span}\{y_1, y_2, \dots, y_s\} = \text{span}\{x_1, x_2, \dots, x_s\} \text{ for } s = 1, 2, \dots, n \tag{13-75}$$

It works as follows. In the first step we assign $z_1 = x_1$ and calculate $y_1 = z_1/\|z_1\|$. In the second step we choose an element $z_2 \in \text{span}\{x_1, x_2\}$ that is orthogonal to y_1: $z_2 = x_2 + \alpha_{21} y_1$ and $\langle z_2, y_1\rangle = 0$. We find $\alpha_{21} = -\langle x_2, y_1\rangle$ and calculate $y_2 = z_2/\|z_2\|$. In the sth step we take an element $z_s \in \text{span}\{x_1, x_2, \dots, x_s\}$ that is orthogonal to $y_1, y_2, \dots, y_{s-1}$: $z_s = x_s + \sum_{r=1}^{s-1} \alpha_{sr} y_r$ and $\langle z_s, y_r\rangle = 0$, $r = 1, 2, \dots, k-1$. We find $\alpha_{sr} = -\langle x_s, y_r\rangle$, $r = 1, 2, \dots, s-1$, and calculate $y_s = z_s/\|z_s\|$.

It is well known that the Gram-Schmidt orthogonalization has poor numerical properties (Golub and Van Loan, 1996). There is, typically, a (severe) loss of orthogonality among the computed basis vectors. The method is, however, still very useful in applications where the orthogonality of the basis vectors is not explicitly taken into account during the calculations.

Example 13.10: Consider the space of real polynomials (Lemma 13.6) with inner product (13-68). Starting from the linear independent set $\{1, x, x^2, \dots, x^m\}$ the $(s+1)$th step of the Gram-Schmidt method becomes

$$\begin{aligned} q_s(x) &= x p_{s-1}(x) - \sum_{l=0}^{s-1} \langle x p_{s-1}(x), p_l(x)\rangle p_l(x) \\ p_s(x) &= q_s(x)/\|q_s(x)\| \end{aligned} \tag{13-76}$$

If we take imaginary grid points, $\bar{x}_k = -x_k$, and $w_{2k} = 0$ for any k in the inner product (13-68), then (13-76) reduces to a three-term recursion formula

$$\begin{aligned} q_s(x) &= x p_{s-1}(x) + \|q_{s-1}(x)\| p_{s-2}(x) \\ p_s(x) &= q_s(x)/\|q_s(x)\| \end{aligned} \tag{13-77}$$

(proof: see Exercise 13.44 and Forsythe, 1957). Note that (13-77) generates only even $p_{2s}(x)$ and odd $p_{2s+1}(x)$ polynomials. □

Example 13.11: Consider the space of real 2 by 1 vector polynomials (Lemma 13.7) with inner product (13-71). Applying the Gram-Schmidt method on the linear independent set $\{E_0(x), E_1(x), \ldots, E_{2m+1}(x)\}$, with $E_{2r}^T(x) = [x^r\ 0]$ and $E_{2r+1}^T(x) = [0\ x^r]$, $r = 0, 1, \ldots, m$, gives the following recursion formula:

$$\begin{aligned} q_s(x) &= E_s(x) - \textstyle\sum_{l=0}^{s-1} \langle E_s(x), p_l(x)\rangle p_l(x) \\ p_s(x) &= q_s(x)/\|q_s(x)\| \end{aligned} \tag{13-78}$$

Using $E_s(x) = xE_{s-2}(x)$ and $E_{s-2}(x) \in \operatorname{span}\{p_0(x), p_1(x), \ldots, p_{s-2}(x)\}$, the recursion can be written as

$$\begin{aligned} q_s(x) &= xp_{s-2}(x) - \textstyle\sum_{l=0}^{s-1} \langle xp_{s-2}(x), p_l(x)\rangle p_l(x) \\ p_s(x) &= q_s(x)/\|q_s(x)\| \end{aligned} \tag{13-79}$$

If we take imaginary grid points, $\bar{x}_k = -x_k$ for any k, then (13-79) reduces to a five-term recursion formula

$$\begin{aligned} q_s(x) &= xp_{s-2}(x) - \beta(p_{s-1}(x) - p_{s-3}(x)) + p_{s-4}(x)\|q_{s-2}(x)\| \\ p_s(x) &= q_s(x)/\|q_s(x)\| \end{aligned} \tag{13-80}$$

with $\beta = \langle xp_{s-2}(x), p_{s-1}(x)\rangle$ (proof: see Exercise 13.46). A numerically stable and time-efficient implementation of the orthogonalization can be found in Van Barel and Bultheel (1992) for real polynomials and real grid point $x_k \in \mathbb{R}$, and Van Barel and Bultheel (1994) for real polynomials and grid points on the unit circle, $|x_k| = 1$ for any k. □

Example 13.12: Consider the space of complex polynomials (Lemma 13.8) with inner product (13-73). Applying the Gram-Schmidt method on the linear independent set $\{1, x, x^2, \ldots, x^m\}$ gives the same full recursion formula (13-76). If we take imaginary grid points, $\bar{x}_k = -x_k$ for any k, then the orthogonalization reduces to

$$\begin{aligned} q_s(x) &= (x - \alpha)p_{s-1}(x) - \beta p_{s-2}(x) \\ p_s(x) &= q_s(x)/\|q_s(x)\| \end{aligned} \tag{13-81}$$

with $\alpha = \langle xp_{s-1}(x), p_{s-1}(x)\rangle$ and $\beta = \langle xp_{s-1}(x), p_{s-2}(x)\rangle$ (proof: similar to Example 13.10). □

Example 13.13: Consider the space of complex 2 by 1 vector polynomials (Lemma 13.9) with inner product (13-74). Applying the Gram-Schmidt method on the linear independent set $\{E_0(x), E_1(x), \ldots, E_{2m+1}(x)\}$, with $E_{2r}^T(x) = [x^r\ 0]$ and $E_{2r+1}^T(x) = [0\ x^r]$, $r = 0, 1, \ldots, m$, gives the same full recursion formula (13-79). If we take imaginary grid points, $\bar{x}_k = -x_k$ for any k, then (13-79) reduces to a five-term recursion formula

$$\begin{aligned} q_s(x) &= (x - \alpha)p_{s-2}(x) - \beta(p_{s-1}(x) + \bar{\beta}p_{s-3}(x)) - \gamma p_{s-4}(x) \\ p_s(x) &= q_s(x)/\|q_s(x)\| \end{aligned} \tag{13-82}$$

with $\alpha = \langle xp_{s-2}(x), p_{s-2}(x)\rangle$, $\beta = \langle xp_{s-2}(x), p_{s-1}(x)\rangle$, $\gamma = \langle xp_{s-2}(x), p_{s-4}(x)\rangle$ (proof: similar to Example 13.11). A numerically stable and time-efficient implementation of the orthogonalization can be found in Van Barel and Bultheel (1994) for complex polynomials and grid points on the unit circle, $|x_k| = 1$ for any k. □

Note that a particular value of the orthogonal polynomial $p_s(x_k)$ is calculated via the recursion formula used for the orthogonalization AND NOT via the expansion of the orthogonal polynomials in powers of x. The last approach is numerically ill-conditioned for high-order polynomials. Similarly, the poles and zeros of orthogonal polynomials are calculated via a companion matrix based on the recursion formula AND NOT via the expansion of the orthogonal polynomials in powers of x (see Section 13.12).

13.12 CALCULATING THE ROOTS OF POLYNOMIALS

13.12.1 Scalar Orthogonal Polynomials

In this section we study the problem of calculating the roots of a polynomial $A(x)$ that is written as a linear combination of scalar orthogonal polynomials $p_r(x)$

$$A(x) = \sum_{r=0}^{n_a} a_r p_r(x) \tag{13-83}$$

The coefficients a_r are known and the orthogonal basis $p_r(x)$, $r = 0, 1, \ldots, n_a$, is defined by the following recursion formula:

$$\begin{aligned} q_r(x) &= xp_{r-1} + \sum_{s=0}^{r-1} \alpha_{rs} p_s(x) \\ p_r(x) &= q_r(x)/\|q_r(x)\| \end{aligned} \tag{13-84}$$

with $q_0(x) = 1$, $q_1(x) = x$ and $\alpha_{rs} = -\langle xp_{r-1}(x), p_s(x)\rangle$ (see Example 13.10).

To maintain good numerical conditioning, the calculation of the roots must use only the orthogonal decomposition of the polynomials and not their explicit form as a polynomial in powers of x. The eigenvalues of the modified companion matrix A,

$$A = A_1^{-1}(A_2 - A_3) \tag{13-85}$$

with A_1, A_2 and A_3, n_a by n_a matrices,

$$A_1 = \operatorname{diag}(1/\|q_{n_a}(x)\|, 1/\|q_{n_a-1}(x)\|, \ldots, 1/\|q_1(x)\|)$$

$$A_2 = \begin{bmatrix} -\dfrac{a_{n_a-1}}{a_{n_a}} & -\dfrac{a_{n_a-2}}{a_{n_a}} & \ldots & -\dfrac{a_1}{a_{n_a}} & -\dfrac{a_0}{a_{n_a}} \\ 1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & \ldots & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & \ldots & 1 & 0 \end{bmatrix}, A_3 = \begin{bmatrix} \dfrac{\alpha_{n_a(n_a-1)}}{\|q_{n_a}(x)\|} & \dfrac{\alpha_{n_a(n_a-2)}}{\|q_{n_a}(x)\|} & \ldots & \dfrac{\alpha_{n_a0}}{\|q_{n_a}(x)\|} \\ 0 & \dfrac{\alpha_{(n_a-1)(n_a-2)}}{\|q_{n_a-1}(x)\|} & \ldots & \dfrac{\alpha_{(n_a-1)0}}{\|q_{n_a-1}(x)\|} \\ \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & 0 & \dfrac{\alpha_{10}}{\|q_1(x)\|} \end{bmatrix} \tag{13-86}$$

are the required roots of $A(x)$ (see Appendix 13.A). For real polynomials with grid points on the imaginary axis (see Example 13.10), (13-84) reduces to the three-term recursion (13-77), and A_3 in (13-86) contains only one nonzero diagonal.

13.12.2 Vector Orthogonal Polynomials

In this section we study the problem of finding the roots of the entries of a 2 by 1 polynomial vector $[A(x)\ B(x)]^T$ that is written as a linear combination of vector orthogonal polynomials

$$\begin{bmatrix} A(x) \\ B(x) \end{bmatrix} = \sum\nolimits_{r=0}^{n_a+n_b+1} a_r \begin{bmatrix} p_r(x) \\ q_r(x) \end{bmatrix} \tag{13-87}$$

with n_a, n_b the orders of respectively $A(x)$, $B(x)$, and where the coefficients a_r are known. The vector orthogonal basis $[p_r(x)\ q_r(x)]^T$, $r = 0, 1, \dots, n_a + n_b + 1$, stems from a rational approximation $B(x)/A(x)$ of a frequency response function $G(x)$ at the grid points x_k, $k = 1, 2, \dots, F$ (see Section 7.16.2) and is calculated through a recursion formula (see, for example, the five-term recursion (13-80) in Example 13.11).

The calculation of the roots of $A(x) = 0$ and $B(x) = 0$ can be reduced to finding the roots of scalar orthogonal polynomials. This is done as follows.

1. Fit a scalar orthogonal polynomial $A_1(x) = \sum_{r=0}^{n_a} \alpha_r p_{1r}(x)$ of order n_a to the denominator $A(x)$ polynomial in (13-87) by minimizing

$$\sum\nolimits_{k=1}^{F} \left| \frac{A_1(x_k) - A(x_k)}{B(x_k)} \right|^2 = \sum\nolimits_{k=1}^{F} \left| \frac{\sum_{r=0}^{n_a} \alpha_r p_{1r}(x_k) - \sum_{r=0}^{n_a+n_b+1} a_r p_r(x_k)}{\sum_{r=0}^{n_a+n_b+1} a_r q_r(x_k)} \right|^2$$

w.r.t. α_0, α_1, ..., α_{n_a} (see Section 7.16.1).

2. Fit a scalar orthogonal polynomial $B_1(x) = \sum_{r=0}^{n_b} \beta_r q_{1r}(x)$ of order n_b to the numerator $B(x)$ polynomial in (13-87) by minimizing

$$\sum\nolimits_{k=1}^{F} \left| \frac{B_1(x_k) - B(x_k)}{A(x_k)} \right|^2 = \sum\nolimits_{k=1}^{F} \left| \frac{\sum_{r=0}^{n_b} \beta_r q_{1r}(x) - \sum_{r=0}^{n_a+n_b+1} a_r q_r(x_k)}{\sum_{r=0}^{n_a+n_b+1} a_r p_r(x_k)} \right|^2$$

w.r.t. β_0, β_1, ..., β_{n_b} (see Section 7.16.1).

3. Calculate the roots of $A_1(x) = \sum_{r=0}^{n_a} \alpha_r p_{1r}(x)$ and $B_1(x) = \sum_{r=0}^{n_b} \beta_r q_{1r}(x)$ using the modified companion matrix approach of Section 13.12.1.

Note that the first and the second step of this procedure introduce no approximation errors because (i) there are no model errors (the true orders of $A(x)$ and $B(x)$ are respectively n_a and n_b), and (ii) there is no disturbing noise ($A(x)$, $B(x)$ are known exactly, within the numerical precision, at the grid points x_k, $k = 1, 2, \dots, F$). Note also that the grid points x_k, $k = 1, 2, \dots, F$, in the first and second steps are the same as those used to calculate the vector orthogonal basis $[p_r(x)\ q_r(x)]^T$, $r = 0, 1, \dots, n_a + n_b + 1$.

13.13 SENSITIVITY OF THE LEAST SQUARES SOLUTION

Consider the overdetermined set of equations

$$Ax \approx b \tag{13-88}$$

with $A \in \mathbb{C}^{n \times m}$, $n > m$, regular and $b \in \mathbb{C}^{m}$. The least squares solution of (13-88) is

$$x_{\mathrm{LS}} = (A^H A)^{-1} A^H b \tag{13-89}$$

The sensitivity of the least squares solution (13-89) to perturbations in A and b equals

$$\|\Delta x_{\mathrm{LS}}\|_2 / \|x_{\mathrm{LS}}\|_2 \le \varepsilon\left(\frac{2\kappa(A)}{\cos(\alpha)} + \mathrm{tg}(\alpha)\kappa^2(A)\right) \tag{13-90}$$

with $\sin(\alpha) = \|r_{\mathrm{LS}}\|_2/\|b\|_2$, $r_{\mathrm{LS}} = Ax_{\mathrm{LS}} - b$, $\varepsilon = \max(\|\Delta A\|_2/\|A\|_2, \|\Delta b\|_2/\|b\|_2)$, and Δ the perturbation; that of the least square residual r_{LS} is given by

$$\|\Delta r_{\mathrm{LS}}\|_2 / \|b\|_2 \le 2\varepsilon\kappa(A) \tag{13-91}$$

(Golub and Van Loan, 1996). It shows that for nonzero residual problems ($r_{\mathrm{LS}} \neq 0$) the sensitivity of x_{LS} depends on the square of $\kappa(A)$, while the sensitivity of r_{LS} just depends linearly on $\kappa(A)$.

The loss in numerical precision (high sensitivity) of the least squares solution (13-89) is basically due to the calculation of $A^H A$ ($\kappa(A^H A) = \kappa^2(A)$). There exist algorithms that calculate x_{LS} without forming the product $A^H A$ explicitly and, hence, have better sensitivity. For example, using the singular value decomposition $A = U\Sigma V^H$, (13-89) can be calculated as

$$x_{\mathrm{LS}} = V\Sigma^{-1}U^H b \text{ or } x_{\mathrm{LS}} = A^+ b, \tag{13-92}$$

while using the QR-factorization $A = QR$, (13-89) is calculated via back-substitution

$$Rx_{\mathrm{LS}} = Q^H b \tag{13-93}$$

The sensitivity of the SVD (13-92) and QR (13-93) solutions is approximately given by

$$\|\Delta x_{\mathrm{LS}}\|_2 / \|x_{\mathrm{LS}}\|_2 \le \varepsilon\left(\frac{2\kappa(A)}{\cos(\alpha)} + \mathrm{tg}(\alpha)\kappa^2(A)\right)10^{-d} \tag{13-94}$$

where d is the number of significant digits used in the calculations (Golub and Van Loan, 1996).

13.14 EXERCISES

13.1. Prove that $\text{range}(A) = (\text{null}(A^H))^{\perp}$ (hint: take any $z_1 \in \text{null}(A^H)$, $z_2 \in \text{range}(A)$ and show that $z_1^H z_2 = 0$).

13.2. Show that the eigenvalues of a Hermitian matrix are real (hint: use the equality $x^H A x = \lambda x^H x$ valid for the eigenvalue, eigenvector pair λ, x and take the complex conjugate).

13.3. Show that the eigenvalues of A and A^T are the same (hint: use $\det(A) = \det(A^T)$).

13.4. Show that the eigenvalues of $A \in \mathbb{C}^{n \times n}$ are invariant w.r.t. a similarity transformation T (hint: use $I_n = TT^{-1}$).

13.5. Show that the eigenvalues of an upper or lower triangular matrix are the diagonal elements.

13.6. Consider the following matrix with a Vandermonde structure:

$$V(x_1, x_2, x_3) = \begin{bmatrix} 1 & x_1 & x_1^2 & \dots & x_1^n \\ 1 & x_2 & x_2^2 & \dots & x_2^n \\ 1 & x_3 & x_3^2 & \dots & x_3^n \end{bmatrix}$$

Show, via linear combinations, that it can be reduced to

$$\alpha \begin{bmatrix} 1 & x_1 & x_1^2 & \dots & x_1^n \\ 0 & 1 & x_1 + x_2 & \dots & \sum_{r_1=0}^{n-1} x_2^{n-1-r_1} x_1^{r_1} \\ 0 & 0 & 1 & \dots & \sum_{r_1=0}^{n-2} \sum_{r_2=0}^{n-2-r_1} x_3^{n-2-r_1-r_2} x_2^{r_2} x_1^{r_1} \end{bmatrix}$$

with $\alpha = (x_3 - x_1)(x_2 - x_1)(x_3 - x_2)$. Note that $V(x_1, x_2, x_3)$ has a full rank if and only if all x_i are different (hint: use $(y^n - x^n)/(y - x) = \sum_{r=0}^{n-1} y^{n-1-r} x^r$).

13.7. Prove that only the symmetric part of a real matrix $A \in \mathbb{R}^{n \times n}$ contributes to $x^T A x$ (hint: use $A = (A + A^T)/2 + (A - A^T)/2$).

13.8. Show that for any $A \in \mathbb{C}^{n \times n}$ and $x \in \mathbb{C}^n$: $\text{Re}(x^H A x) = x^H \text{herm}(A) x$. Conclude that only the Hermitian part of a matrix contributes to the real value of $x^H A x$.

13.9. If $\begin{bmatrix} A & D \\ C & B \end{bmatrix}^{-1}$ and B^{-1} exist, show that

$$\det\left(\begin{bmatrix} A & D \\ C & B \end{bmatrix}\right) = \det(B)\det([A - DB^{-1}C]) \tag{13-95}$$

(hint: $\begin{bmatrix} A & D \\ C & B \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & B \end{bmatrix} \begin{bmatrix} A & D \\ B^{-1}C & I \end{bmatrix}$ and $\det(AB) = \det(A)\det(B)$).

13.10. If B^{-1} exists, show that

$$\begin{bmatrix} A & D \\ C & B \end{bmatrix}^{-1} = \begin{bmatrix} \Delta^{-1} & -\Delta^{-1}F \\ -E\Delta^{-1} & B^{-1} + E\Delta^{-1}F \end{bmatrix} \tag{13-96}$$

where $\Delta = A - DB^{-1}C$, $E = B^{-1}C$, and $F = DB^{-1}$. Show, using Exercise 13.9, that Δ^{-1} exists.

13.11. Show that for any $A \in \mathbb{C}^{n\times m}$ and $B \in \mathbb{C}^{m\times n}$, $\mathrm{tr}(AB) = \mathrm{tr}(BA)$.

13.12. Show that for any $A \in \mathbb{C}^{n\times n}$, $\mathrm{tr}(A) = \sum_{k=1}^{n} \lambda_k(A)$.

13.13. Show that for any $A, B \in \mathbb{C}^{n\times m}$, $\mathrm{tr}(A^T B) = (\mathrm{vec}(A))^T \mathrm{vec}(B)$.

13.14. Prove the submultiplicative property (13-14) for the Frobenius norm.

13.15. Show that $\|A\|_2 \le \|A\|_F$.

13.16. Using the definition of the singular value decomposition of a matrix $A \in \mathbb{C}^{n\times m}$ $(n \ge m)$ show that the right singular vectors v_k and the squared singular values σ_k^2, $k = 1, 2, \ldots, m$, are the eigenvectors and eigenvalues of $A^H A$.

13.17. Using the definition of the singular value decomposition of a matrix $A \in \mathbb{C}^{n\times m}$ $(n \ge m)$, show that the left singular vectors u_k and the squared singular values σ_k^2, $k = 1, 2, \ldots, n$, are the eigenvectors and eigenvalues of AA^H.

13.18. Let $A = U_A \Sigma_A X^{-1}$, $B = U_B \Sigma_B X^{-1}$ be the generalized singular value decomposition of the matrix pair (A, B). Show that $x = X_{[:,k]}$ and $\lambda = \alpha_k^2/\beta_k^2$, $k = 1, 2, \ldots, m$, are the solutions of the generalized eigenvalue problem $A^H Ax = \lambda B^H Bx$.

13.19. Show that a Hermitian positive definite matrix $A \in \mathbb{C}^{n\times m}$ can be written as $A = V\Sigma V^H$ where $V^H V = VV^H = I_n$ and $\Sigma = \mathrm{diag}(\lambda_1(A), \lambda_2(A), \ldots, \lambda_n(A))$ (hint: show that $\lambda_k(A) = \sigma_k(A)$ and that $U_{[:,k]} = V_{[:,k]}$).

13.20. Prove that a Hermitian matrix $A \in \mathbb{C}^{n\times m}$ is positive (semi-)definite if and only if $\lambda_k(A) > 0$ $(\lambda_k(A) \ge 0)$ (hint: apply the results of Exercise 13.19, $A = V\Sigma V^H$ with $V^H = V^{-1}$, to the quadratic form $x^H Ax$).

13.21. Let $Q \in \mathbb{C}^{n\times n}$ be a unitary matrix $(Q^H Q = I_n)$. Show that $\kappa(Q) = 1$.

13.22. Prove inequality (13-16) (hint: first show that $(A+E)^{-1} - A^{-1} = -A^{-1}E(A+E)^{-1}$, next use $(I_n + A^{-1}E)^{-1} = \sum_{k=0}^{\infty}(-A^{-1}E)^k$).

13.23. Let $A \in \mathbb{C}^{n\times m}$ with $n \ge m$ and $B \in \mathbb{C}^{m\times m}$ nonsingular. Show that $\sigma(A, B) = \sigma(AB^{-1})$.

13.24. Let $x \in \mathbb{C}^n$. Show that $x^+ = x^H/(x^H x)$.

13.25. Show that expression (13-28) satisfies the four Moore-Penrose conditions.

13.26. Let $A \in \mathbb{C}^{n\times r}$ and $B \in \mathbb{C}^{r\times m}$ with $r \le \min(n, m)$ and $\mathrm{rank}(A) = \mathrm{rank}(B) = r$. Show that $(AB)^+ = B^+A^+$ (hint: verify that the four Moore-Penrose conditions are satisfied with $A^+ = (A^H A)^{-1}A^H$ and $B^+ = B^H(BB^H)^{-1}$).

13.27. Let $P \in \mathbb{C}^{n\times n}$ be a Hermitian idempotent matrix. Show that $P^+ = P$.

13.28. Show that the eigenvalues of an idempotent matrix P are one or zero (hint: left multiply $Px = \lambda x$ by P and work out).

13.29. Prove properties 2, 3, and 4 of the idempotent matrices (see Section 13.6).

13.30. Show that $(A \otimes B)(D \otimes G) = (AD) \otimes (BG)$.

13.31. Show that $\mathrm{vec}(ADB) = (B^T \otimes A)\mathrm{vec}(D)$ (hint: calculate the kth column of ADB and use $(XY)_{[:,k]} = XY_{[:,k]}$).

13.32. Show that $(N \otimes M)^{-1} = N^{-1} \otimes M^{-1}$ (hint: use (13-36)).

13.33. Take any $A \in \mathbb{C}^{n \times m}$ and $B \in \mathbb{C}^{p \times q}$. Show that $\|A \otimes B\| = \|A\|\|B\|$ where $\| \ \|$ denotes the 1-, 2-, ∞-, and Frobenius norm (hint: for the 2- and Frobenius norm first show using (13-36) and (13-35) that $A \otimes B = (U_A \otimes U_B)(\Sigma_A \otimes \Sigma_B)(V_A \otimes V_B)^H$ with $A = U_A \Sigma_A V_A^H$ and $B = U_B \Sigma_B V_B^H$ the corresponding singular valued decompositions).

13.34. Prove properties (13-41), (13-42), (13-43), and (13-45) (hint: use $A = B^{-1} \Leftrightarrow AB = I_n$ for (13-43) and $A = B^{+} \Leftrightarrow BAB = B$ for (13-44)).

13.35. Take a positive definite matrix $A \in \mathbb{C}^{n \times n}$. Show that the real matrix A_{Re} is symmetric and positive definite.

13.36. Take a unitary matrix A. Show that A_{Re} is orthogonal.

13.37. Prove property (13-46) (hint: using linear combinations of block rows and block columns show that

$$\det\left(\begin{bmatrix} \mathrm{Re}(A) & -\mathrm{Im}(A) \\ \mathrm{Im}(A) & \mathrm{Re}(A) \end{bmatrix}\right) = \det\left(\begin{bmatrix} A & 0 \\ \mathrm{Im}(A) & \bar{A} \end{bmatrix}\right)$$

13.38. Prove property (13-47) (hint: first show that the singular value decompositions of A and A_{Re} are related to each other by $A = U \Sigma V^H$ and $A_{\mathrm{Re}} = U_{\mathrm{Re}} \Sigma_{\mathrm{Re}} V_{\mathrm{Re}}^T$).

13.39. Prove properties (13-49), (13-50), and (13-51).

13.40. Verify results (13-56), (13-57), (13-58), and (13-60).

13.41. Show that $\partial f(x, \bar{x})/\partial \bar{x} = \overline{\partial f(x, \bar{x})/\partial x}$ if $f(x, \bar{x}) \in \mathbb{R}$ (hint: use $\bar{f}(x, \bar{x}) = f(\bar{x}, x)$).

13.42. Verify results (13-62), (13-63), (13-64), (13-65), (13-66), and (13-67) (hint: use $A^{-1} = \mathrm{adj}(A)/\det(A)$ for (13-64), where $\mathrm{adj}(A)$ is the transposed matrix of cofactors of A; show first that $\partial A^{-1}/\partial A_{[k,l]} = -A^{-1}(\partial A/\partial A_{[k,l]})A^{-1}$ for (13-65)).

13.43. Prove Lemma 13.8.

13.44. Show that (13-76) reduces to a three-term recursion if for any k, $w_{2k} = 0$ and $\bar{x}_k = -x_k$ in (13-68) (hint: first use $\langle xp_{s-1}(x), p_l(x)\rangle = -\langle p_{s-1}(x), xp_l(x)\rangle$ and $xp_l(x) \in \mathrm{span}\{p_0(x), p_1(x), \ldots, p_{l+1}(x)\}$ to prove the three-term recursion; next show that $\langle p_{s-1}(x), xp_{s-1}(x)\rangle = 0$ and $\langle p_{s-1}(x), xp_{s-2}(x)\rangle = \|q_{s-1}(x)\|$).

13.45. Consider Example 13.10 with real grid points ($x_k \in \mathbb{R}$) and $w_{2k} = 0$ for any k in (13-68). Show that (13-76) reduces to a three-term recursion formula (hint: use $\langle xp_{s-1}(x), p_l(x)\rangle = \langle p_{s-1}(x), xp_l(x)\rangle$, $xp_l(x) \in \mathrm{span}\{p_0(x), p_1(x), \ldots, p_{l+1}(x)\}$).

13.46. Show that (13-79) reduces to a five-term recursion if $\bar{x}_k = -x_k$ for any k in (13-71) (hint: first use $\langle xp_{s-2}(x), p_l(x)\rangle = -\langle p_{s-2}(x), xp_l(x)\rangle$ and $xp_l(x) \in \mathrm{span}\{p_0(x), p_1(x), \ldots, p_{l+2}(x)\}$ to prove the five-term recursion; next show that $\langle p_{s-2}(x), xp_{s-2}(x)\rangle = 0$, $\langle p_{s-2}(x), xp_{s-1}(x)\rangle = -\langle p_{s-2}(x), xp_{s-3}(x)\rangle$ and $\langle p_{s-2}(x), xp_{s-4}(x)\rangle = \|q_{s-2}(x)\|$).

13.15 APPENDIX

Appendix 13.A: Calculation of the Roots of a Polynomial

The roots of the polynomial $A(x)$ are those values of x such that $A(x) = 0$ or

$$p_{n_a}(x) = -\frac{1}{a_{n_a}} \sum_{r=0}^{n_a - 1} a_r p_r(x) \tag{13-97}$$

(use (13-83)). Adding the equations $p_r(x) = p_r(x)$, $r = n_a - 1, n_a - 2, \ldots, 1$, to (13-97) gives the following set of n_a equations:

$$A_2 Z = Z_1 \tag{13-98}$$

with $Z_1^T = [p_{n_a}(x)\ p_{n_a-1}(x)\ \ldots p_1(x)\ ]$, $Z^T = [\ p_{n_a-1}(x)\ p_{n_a-2}(x)\ \ldots\ p_0(x)]$ and where A_2 is defined in (13-86). Recursion (13-84) can be written in matrix form as

$$Z_1 = xA_1 Z + A_3 Z \tag{13-99}$$

where A_1 and A_3 are n_a by n_a matrices defined in (13-86). Combining (13-98) and (13-99) shows that the roots x of $A(x) = 0$ are the solutions of the eigenvalue problem $A_1^{-1}(A_2 - A_3)Z = xZ$. □

14

Some Probability and Stochastic Convergence Fundamentals

Abstract: The goal of this chapter is to give insight into the way to analyze the stochastic properties of an estimator. Therefore, a great deal of attention is paid to the different concepts of stochastic convergence. The main ideas behind the stochastic convergence proofs used throughout this book are explained and some basic analysis tools are provided. More information on the topic can be found in the following textbooks: Billingsley (1995), Chow and Teicher (1988), Brillinger (1981), Lukacs (1975), Stout (1974), and Jazwinski (1970). The calculation of probabilities, expected values, and higher order (central) moments and the properties of standard distributions are assumed to be known. More information can be found in Anderson (1958), Stuart and Ord (1987), and Mathai and Provost (1992). This chapter also includes a study of the properties of the noise after a discrete Fourier transform, which is essential in frequency domain identification.

14.1 NOTATIONS AND DEFINITIONS

$\mathcal{E}\{\ \}$ and Prob( ) denote the expected value and the probability function, respectively. If $f_x(x)$ and $F(x)$ are the respective probability density function and distribution function of the random variable x, then the expected value of $g(x)$ is given by

$$\mathcal{E}\{g(x)\} = \int_X g(x)dF(x) = \int_X g(x)f_x(x)dx \tag{14-1}$$

where X is the domain of $F(x)$.

Let $x, y \in \mathbb{C}$ be complex random variables: then the *mean* μ_x and *variance* σ_x^2 of x and the *covariance* σ_{xy}^2 between x and y are defined as

$$\mu_x = \mathcal{E}\{x\} \qquad \sigma_x^2 = \text{var}(x) = \mathcal{E}\{|x - \mathcal{E}\{x\}|^2\} \tag{14-2}$$

$$\sigma_{xy}^2 = \text{covar}(x, y) = \mathcal{E}\{(x - \mathcal{E}\{x\})\overline{(y - \mathcal{E}\{y\})}\} \tag{14-3}$$

System Identification: A Frequency Domain Approach. Rik Pintelon and Johan Schoukens
Copyright © 2001 Institute of Electrical and Electronics Engineers. ISBN: 0-780-36000-1

Let $x, y \in \mathbb{C}^n$ be complex random vectors: then the *covariance matrix* C_x of x and the *cross-covariance matrix* C_{xy} between x and y are given by

$$C_x = \mathrm{Cov}(x) = \mathcal{E}\{(x-\mathcal{E}\{x\})(x-\mathcal{E}\{x\})^H\} \tag{14-4}$$

$$C_{xy} = \mathrm{Cov}(x, y) = \mathcal{E}\{(x-\mathcal{E}\{x\})(y-\mathcal{E}\{y\})^H\} \tag{14-5}$$

Let $\hat{x} \in \mathbb{C}^n$ be an estimate of the true value x_0. The *bias* b_x and the *mean square error* $\mathrm{MSE}(\hat{x})$ of the estimate are given, respectively, by

$$b_x = \mathcal{E}\{\hat{x}\} - x_0 \tag{14-6}$$

$$\mathrm{MSE}(\hat{x}) = \mathcal{E}\{(\hat{x}-x_0)(\hat{x}-x_0)^H\} = \mathrm{Cov}(\hat{x}) + b_x b_x^H \tag{14-7}$$

A stochastic process $x(t) \in \mathbb{C}^n$, $t \in \mathbb{Z}$, is *strictly stationary* if the joint distribution of $x(t_1+t)$, $x(t_2+t)$, ... $x(t_k+t)$ does not depend on t for every $t, t_1, \ldots, t_k \in \mathbb{Z}$ and $k = 1, 2, 3, \ldots$ For example, a series of independent, identically distributed random vectors is strictly stationary. A stochastic process $x(t) \in \mathbb{C}^n$, $t \in \mathbb{Z}$, is *second-order stationary* or *wide-sense stationary* if the first- and second-order moments are invariant under a common shift of the argument t

$$\mu_x(t_1+t) = \mu_x(t_1) \tag{14-8}$$

$$\mathrm{Cov}(x(t_1+t), x(t_2+t)) = \mathrm{Cov}(x(t_1), x(t_2)) \tag{14-9}$$

for every $t, t_1, t_2 \in \mathbb{Z}$. A strictly stationary process with finite second-order moments is second-order stationary.

The stochastic process $x(t) \in \mathbb{C}^n$, $t \in \mathbb{Z}$, is *independent* if $x(t)$ and $x(s)$ are independent whenever $t - s \neq 0$. It is *m-dependent* if $x(t), x(t+1), \ldots, x(t+r)$ is independent of $x(t+r+n), x(t+r+n+1), \ldots, x(t+r+s)$ for any $n > m \geq 0$, with $r, s > 0$. It is uniformly bounded if $|x(t)| \leq C < \infty$ for any realization and for any t.

Let $x(t) \in \mathbb{C}^n$ and $y(t) \in \mathbb{C}^m$ be second-order stationary stochastic processes; then the *autocorrelation matrix* $R_{xx}(\tau)$ of $x(t)$ and the *cross-correlation matrix* $R_{xy}(\tau)$ between x and y are defined as

$$R_{xx}(\tau) = \mathcal{E}\{x(t)x^H(t-\tau)\} \tag{14-10}$$

$$R_{xy}(\tau) = \mathcal{E}\{x(t)y^H(t-\tau)\} \tag{14-11}$$

The *auto-* and *cross-power spectra* are the Fourier transforms of the auto- and cross-correlation matrices.

A *real normally distributed* random vector $x \in \mathbb{R}^n$ with mean μ_x and covariance matrix C_x, will be denoted as $x \in N_n(\mu_x, C_x)$. If subscript n is omitted then this is equivalent to $n = 1$. Similarly, $x \in E(\mu_x, \sigma_x^2)$, $x \in L(\mu_x, \sigma_x^2)$, and $x \in U(\mu_x, \sigma_x^2)$ denote *real exponential, Laplace,* and *uniform* random variables, respectively. The sum of the squares $y = \sum_{k=1}^{n} x_k^2$ of n independent and identically distributed normal random variables $x_k \in N(0, 1)$ is *chi-squared distributed* with n degrees of freedom $y \in \chi^2(n)$. The ratio $z = (n_2 y_1)/(n_1 y_2)$ of two independent chi-squared distributed random variables $y_1 \in \chi^2(n_1)$ and $y_2 \in \chi^2(n_2)$ is *F-distributed* with n_1 and n_2 degrees of freedom $z \in F(n_1, n_2)$. The matrix-valued random variable $(y = \sum_{k=1}^{n} x_k x_k^T) \in \mathbb{R}^{p \times p}$, formed by the sum of the product of n independent and identically distributed normal vectors $x_k \in N_p(0, C_x)$, is *Wishart distributed*, $y \in W_p(n, C_x)$, with n degrees of freedom and associated parameter matrix C_x (Anderson, 1958; Mathai and Provost, 1992). As a special case, we have $W_1(n, 1) = \chi^2(n)$.

A complex random vector $x \in \mathbb{C}^n$ is said to be *circular* if

$$\mathcal{E}\{(x-\mathcal{E}\{x\})(x-\mathcal{E}\{x\})^T\} = 0 \tag{14-12}$$

(Picinbono, 1993). If $x \in \mathbb{C}^n$ is a circular complex random vector and $A \in \mathbb{C}^{m \times n}$, then $y = Ax$ is also circular (Exercise 14.3). Condition (14-12) is equivalent to

$$\begin{aligned} \text{Cov}(\text{Re}(x)) &= \text{Cov}(\text{Im}(x)) \\ \text{Cov}(\text{Re}(x), \text{Im}(x)) &= -(\text{Cov}(\text{Im}(x), \text{Re}(x)))^T \end{aligned} \tag{14-13}$$

If (14-13) is valid, then it can be verified that $\text{Cov}(x_{\text{re}}) = 0.5(\text{Cov}(x))_{\text{Re}}$ (Exercise 14.4). The probability density function $f_x(x)$ of complex random variables $x \in \mathbb{C}^n$ is given by the joint probability density function $f_{x_{\text{re}}}(x_{\text{re}})$ of the real and imaginary parts of x. For *circular complex normally* distributed random vectors $x \in \mathbb{C}^n$, the probability density function is uniquely determined by the mean value μ_x and the covariance matrix C_x (Picinbono, 1993)

$$f_x(x) = f_{x_{\text{re}}}(x_{\text{re}}) = \frac{1}{\pi^n \det(C_x)} \exp(-(x-\mu_x)^H C_x^{-1}(x-\mu_x)) \tag{14-14}$$

If the complex normally distributed noise is not circular, then the other second-order moment $\mathcal{E}\{(x-\mathcal{E}\{x\})(x-\mathcal{E}\{x\})^T\}$ is also required to construct $f_{x_{\text{re}}}(x_{\text{re}})$. In the univariate case, (14-13) implies that the real and imaginary parts of $x \in \mathbb{C}$ have equal variance and have zero covariance. If $x \in \mathbb{C}$ is circular complex normally distributed, then its real and imaginary parts are independent and $\mathcal{E}\{(x-\mathcal{E}\{x\})^n\} = 0$ (see Exercise 14.8 and Schoukens and Pintelon, 1990).

A *circular complex normally distributed* random vector $x \in \mathbb{C}^n$ with mean μ_x and covariance matrix C_x, will be denoted as $x \in N_n^c(\mu_x, C_x)$. Similarly, $x \in E^c(\mu_x, \sigma_x^2)$, $x \in L^c(\mu_x, \sigma_x^2)$, and $x \in U^c(\mu_x, \sigma_x^2)$ denote *circular complex exponential, Laplace,* and *uniform* random variables with independent real and imaginary parts. The matrix-valued random variable $(y = \sum_{k=1}^n x_k x_k^H) \in \mathbb{C}^{p \times p}$ with $x_k \in N_p^c(0, C_x)$ is *complex Wishart distributed*, $y \in W_p^c(n, C_x)$, with n degrees of freedom and associated parameter matrix C_x (Goodman, 1963; Brillinger, 1981). Its probability density function is given by

$$f_y(y) = \frac{(\det(y))^{n-p}\exp(-\text{tr}(C_x^{-1}y))}{\pi^{p(p-1)/2}(\det(C_x))^n \prod_{k=1}^n (n-k)!} \tag{14-15}$$

for $n \geq p$ and $y \geq 0$ (Goodman, 1963; Brillinger, 1981). Note that the real part of a complex Wishart distributed random variable is, in general, not Wishart distributed. This is due to the fact that the real and imaginary parts of the $x_k \in N_p^c(0, C_x)$ are not necessarily independent.

Take n complex random variables $x_k \in \mathbb{C}$ with $\mathcal{E}\{|x_k|^n\} < \infty$, $k = 1, 2, \ldots, n$. The *joint cumulant* $\text{cum}(x_1, x_2, \ldots, x_n)$ of order n is given by

$$\text{cum}(x_1, x_2, \ldots, x_n) = \sum (-1)^{p-1}(p-1)! \prod_{m=1}^p \mathcal{E}\{\prod_{k_m \in \mathbb{V}_m} x_{k_m}\} \tag{14-16}$$

where the summation extends over all partitions $\{\mathbb{V}_1, \mathbb{V}_2, \ldots, \mathbb{V}_p\}$ of $\mathbb{I} = \{1, 2, \ldots, n\}$. If for any k, $x_k = x$, then the definition gives the *cumulant* of order n of x.

Example 14.1: Calculation of the third-order joint cumulant. All the partitions $\{\mathbb{V}_1, \mathbb{V}_2, \ldots, \mathbb{V}_p\}$ of the set $\{1, 2, 3\}$ are $\{\{1\}, \{2\}, \{3\}\}$, $\{\{1\}, \{2, 3\}\}$, $\{\{2\}, \{1, 3\}\}$, $\{\{3\}, \{1, 2\}\}$, and $\{\{1, 2, 3\}\}$. Hence, formula (14-16) becomes

$$\begin{aligned}\mathrm{cum}(x_1, x_2, x_3) &= (-1)^2 2!\,\mathcal{E}\{x_1\}\mathcal{E}\{x_2\}\mathcal{E}\{x_3\} + (-1)^1 1!\,\mathcal{E}\{x_1\}\mathcal{E}\{x_2 x_3\} + \\ &\quad (-1)^1 1!\,\mathcal{E}\{x_2\}\mathcal{E}\{x_1 x_3\} + (-1)^1 1!\,\mathcal{E}\{x_3\}\mathcal{E}\{x_1 x_2\} + (-1)^0 0!\,\mathcal{E}\{x_1 x_2 x_3\}\end{aligned} \tag{14-17}$$

where $0! = 1$. □

The cumulants have the following properties for any random variables $x_k, y_l \in \mathbb{C}$ and constants $a_k, b_l \in \mathbb{C}$, $k, l = 1, 2, \ldots$ (Brillinger, 1981):

1. Symmetric in their arguments, for example, $\mathrm{cum}(x_1, x_2, x_3) = \mathrm{cum}(x_3, x_1, x_2)$.
2. Multilinear functions of their arguments, for example, $\mathrm{cum}(\sum_{k=1}^{n} a_k x_k, \sum_{l=1}^{m} b_l y_l) = \sum_{k=1}^{n}\sum_{l=1}^{m} a_k b_l \mathrm{cum}(x_k, y_l)$.
3. If any group of the x_k's, $k = 1, 2, \ldots, n$ are independent of the remaining x_k's, then $\mathrm{cum}(x_1, x_2, \ldots, x_n) = 0$.
4. If the random variables $x_1, x_2, \ldots, x_n$ are independent of $y_1, y_2, \ldots, y_n$, then $\mathrm{cum}(x_1 + y_1, x_2 + y_2, \ldots, x_n + y_n) = \mathrm{cum}(x_1, x_2, \ldots, x_n) + \mathrm{cum}(y_1, y_2, \ldots, y_n)$.
5. $\mathrm{cum}(x_1, x_2, \ldots, x_{k-1}, a_1, x_{k+1}, \ldots, x_r) = 0$ for $r = 2, 3, \ldots, n$.
6. $\mathrm{cum}(x_k) = \mathcal{E}\{x_k\}$, $\mathrm{cum}(x_k, \bar{x}_k) = \mathrm{var}(x_k)$ and $\mathrm{cum}(x_k, \bar{y}_l) = \mathrm{covar}(x_k, y_l)$.
7. For stationary random variables, $x(k) \in \mathbb{C}$, $k \in \mathbb{Z}$, we have $\mathrm{cum}(x(k_1), x(k_2), \ldots, \ldots, x(k_n)) = \mathrm{cum}(x(k_1 - k_n), x(k_2 - k_n), \ldots, x(0))$ for every $k_i \in \mathbb{Z}$, $i = 1, 2, \ldots, n$.

It can be shown that the joint cumulant (14-16) equals the coefficient of $j^r t_1 t_2 \ldots t_r$ in the Taylor series expansion of $\ln(\mathcal{E}\{\exp(j\sum_{k=1}^{r} x_k t_k)\})$ about the origin $t_1 = t_2 = \ldots = 0$ (Brillinger, 1981).

Example 14.2: Let $x \in \mathbb{R}^n$ be a multivariate normally distributed random variable with covariance matrix C_x and mean value μ_x. Its *characteristic function* $\phi(t)$ is given by (Stuart and Ord, 1987)

$$\phi(t) = \mathcal{E}\{\exp(j\sum_{k=1}^{n} x_{[k]} t_{[k]})\} = \exp(-\frac{1}{2} t^T C_x t + j t^T \mu_x) \tag{14-18}$$

with $t \in \mathbb{R}^n$. Note that this result remains valid if the covariance matrix C_x is not of full rank (Mathai and Provost, 1992). Taking the natural logarithm of (14-18) shows that all joint cumulants of order greater than two are zero

$$\mathrm{cum}(x_{[k_1]}, x_{[k_2]}, \ldots, x_{[k_r]}) = 0 \qquad k_i \in \{1, 2, \ldots, r\}, r > 2 \tag{14-19}$$

If $x \in \mathbb{C}^n$ is a multivariate complex normal random variable, then $x_{\mathrm{re}} \in \mathbb{C}^{2n}$ is a real-valued normal random variable for which result (14-19) applies. Because the cumulants of x can be written as a linear combination of the cumulants of x_{re} (multilinearity property 2), it follows that (14-19) is also valid for multivariate complex normal random variables. □

Chebyshev's inequality: Take two random variables x, y with finite second-order moments and ε an arbitrary positive real number, then (Billingsley, 1995)

$$\mathrm{Prob}(|x-y|>\varepsilon)\le\frac{1}{\varepsilon^2}\mathscr{E}\{(x-y)^2\} \tag{14-20}$$

Markov's inequality: take two random variables x, y with finite absolute moments of order $p>0$ and ε an arbitrary positive real number, then (Stuart and Ord, 1987)

$$\mathrm{Prob}(|x-y|>\varepsilon)\le\frac{1}{\varepsilon^p}\mathscr{E}\{|x-y|^p\} \tag{14-21}$$

14.2 THE COVARIANCE MATRIX OF A FUNCTION OF A RANDOM VARIABLE

Let $x\in\mathbb{R}^n$ be a random vector with mean value μ_x. In general it is impossible to calculate the covariance matrix of the function $f(x)\in\mathbb{R}^m$. An approximation can be calculated through linearization of the function $f(x)$. Assuming that $f(x)$ has continuous derivative w.r.t. x, we find

$$f(x)\approx f(\mu_x)+\left.\frac{\partial f(x)}{\partial x}\right|_{x=\mu_x}(x-\mu_x) \tag{14-22}$$

and

$$\mathrm{Cov}(f(x))\approx\left.\frac{\partial f(x)}{\partial x}\right|_{x=\mu_x}\mathrm{Cov}(x)\left(\left.\frac{\partial f(x)}{\partial x}\right|_{x=\mu_x}\right)^T \tag{14-23}$$

Under some conditions on x and $\partial f(x)/\partial x|_{x=\mu_x}$, the right-hand side of (14-22) is approximately Gaussian distributed (see Section 14.10). Then, (14-23) is very useful for calculating confidence levels and uncertainty bounds on $f(x)$, even if the second-order moments of $f(x)$ do not exist.

In general, if $x\in\mathbb{C}^n$ and/or $f(x)\in\mathbb{C}^m$, then x and/or $f(x)$ in (14-23) should be replaced, respectively, by x_{re} and/or $f_{\mathrm{re}}(x)$. Some important special complex cases lead to simplified formulae. If $x\in\mathbb{C}^n$ is circular complex distributed and $f(x)\in\mathbb{C}^m$ is an analytic function of x, then the right-hand side of (14-22) is circular complex distributed and (14-23) becomes

$$\mathrm{Cov}(f(x))\approx\left.\frac{\partial f(x)}{\partial x}\right|_{x=\mu_x}\mathrm{Cov}(x)\left(\left.\frac{\partial f(x)}{\partial x}\right|_{x=\mu_x}\right)^H \tag{14-24}$$

If $x\in\mathbb{C}^n$ is circular complex distributed and $f(x,\bar{x})\in\mathbb{R}^m$, then

$$\mathrm{Cov}(f(x,\bar{x}))\approx2\mathrm{Re}\left(\left.\frac{\partial f(x,\bar{x})}{\partial x}\right|_{x=\mu_x}\mathrm{Cov}(x)\left(\left.\frac{\partial f(x,\bar{x})}{\partial x}\right|_{x=\mu_x}\right)^H\right) \tag{14-25}$$

(Proof: see Exercise 14.15).

14.3 SAMPLE VARIABLES

A particular realization of a stochastic process $x(t) \in \mathbb{C}^n$ is denoted by $x^{[k]}(t)$ where k can be a random variable itself. It indicates that the outcome of the process $x(t)$ for each value of t depends on the particular value of k. In most cases k is a positive integer number corresponding to the index of the realization.

The *sample mean* $\hat{x}(t)$ and *sample (cross-)covariance matrices* $\hat{C}_{xy}(t)$ and $\hat{C}_x(t)$ of R realizations of the stochastic processes $x(t), y(t) \in \mathbb{C}^n$ are defined as

$$\hat{x}(t) = \frac{1}{R}\sum_{k=1}^{R} x^{[k]}(t) \tag{14-26}$$

$$\hat{C}_x(t) = \frac{1}{R-1}\sum_{k=1}^{R} (x^{[k]}(t) - \hat{x}(t))(x^{[k]}(t) - \hat{x}(t))^H \tag{14-27}$$

$$\hat{C}_{xy}(t) = \frac{1}{R-1}\sum_{k=1}^{R} (x^{[k]}(t) - \hat{x}(t))(y^{[k]}(t) - \hat{y}(t))^H \tag{14-28}$$

The sample mean $\hat{x}(t)$ and sample (cross-)covariance matrices $\hat{C}_{xy}(t)$, $\hat{C}_x(t)$ of *independent realizations* are unbiased estimates of the mean $\mu_x(t)$ and (cross-)covariance matrices $C_{xy}(t)$, $C_x(t)$.

For real and circular complex normally distributed processes the sample mean $\hat{x}(t)$ and sample covariance matrix $\hat{C}_x(t)$ are independently distributed (see Anderson, 1958 for the real case; see Giri, 1965 for the circular complex case). Because $\hat{C}_x(t)$ and $\hat{x}(t)$ are only functions of, respectively, $\hat{C}_{x_{\mathrm{re}}}(t)$ and $x_{\mathrm{re}}(t)$, it can be seen that this result is also valid for noncircular complex normal processes. For real and circular complex normal processes we also have

$$x(t) \in N_n(\mu_x(t), C_x(t)) \Rightarrow \begin{cases} \hat{x}(t) \in N_n(\mu_x(t), C_x(t)/R) \\ \hat{C}_x(t) \in W_n((R-1), C_x(t)/(R-1)) \end{cases} \tag{14-29}$$

$$x(t) \in N_n^c(\mu_x(t), C_x(t)) \Rightarrow \begin{cases} \hat{x}(t) \in N_n^c(\mu_x(t), C_x(t)/R) \\ \hat{C}_x(t) \in W_n^c((R-1), C_x(t)/(R-1)) \end{cases} \tag{14-30}$$

(see Anderson, 1958 for the real case; see Giri, 1965 for the circular complex case).

The statistical performance of estimators is often compared through Monte Carlo simulations. In such simulations the true parameter value is known, while the true covariance of the estimates is unknown. Because the parameter estimates are mostly asymptotically normal, we may use Hotelling's T^2-statistic to test the bias of the estimates. For real and circular complex normally distributed, it becomes

$$b = (\hat{x} - \mu_x)^H(\hat{C}_x/R)^{-1}(\hat{x} - \mu_x) \tag{14-31}$$

with

$$\begin{aligned} x \in N_n(\mu_x, C_x) &\Rightarrow b \in (R-1)\frac{n}{R-n}F(n, R-n) && \text{(a)} \\ x \in N_n^c(\mu_x, C_x) &\Rightarrow b \in (R-1)\frac{2n}{2(R-n)}F(2n, 2(R-n)) && \text{(b)} \end{aligned} \qquad (14\text{-}32)$$

(see Anderson, 1958 for the real case; see Giri, 1965 for the circular complex case). If for all realizations the same dependence exists between the estimated parameters, then $\hat{C}_x$ is rank deficient and the inverse in (14-31) should be replaced by the pseudoinverse. The statistic (14-32) is still valid if n is replaced by rank$(\hat{C}_x)$ (see Exercise 14.18). This rank deficiency problem often occurs when identifying models with a redundant number of parameters. For noncircular complex normal parameters $x \in \mathbb{C}^n$ the bias test (14-31), (14-32) is performed on x_{re} and $\hat{C}_{x_{\text{re}}}$ where n is replaced by $2n$ in (14-32-a).

14.4 MIXING RANDOM VARIABLES

14.4.1 Definition

An important requirement that will be imposed on the perturbing noise after sampling is that it has a limited span of dependence. This requirement is formalized by the mixing assumption for discrete time noise. The definition for mixing random variables used throughout this book is slightly more general than the classical definition given in Brillinger (1981) in the sense that it is also valid for a certain class of nonstationary signals.

The real random vectors $x(t) \in \mathbb{R}^r$, $t = 0, 1, 2, \ldots$, are called *mixing of order P* if

$$\max_{t_k} \sum_{t_1, t_2, \ldots, t_{k-1} = 0}^{\infty} \left|\text{cum}(x_{[a_1]}(t_1), x_{[a_2]}(t_2), \ldots, x_{[a_{k-1}]}(t_{k-1}), x_{[a_k]}(t_k))\right| < \infty \qquad (14\text{-}33)$$

for every $a_i \in \{1, 2, \ldots, r\}$, $i = 1, 2, \ldots, k$, and $k = 1, 2, \ldots, P$. Because cumulants are symmetric in their arguments (see Section 14.1, property 1), this definition is independent of the particular choice of the index t_i, $i = 1, 2, \ldots k$, used to take the maximum. A mixing condition of order P assumes that all moments of order $k = 1, 2, \ldots, P$ exist and are uniformly bounded. The random vectors $x_1(t) \in \mathbb{R}^{r_1}$, $x_2(t) \in \mathbb{R}^{r_2}$, $\ldots$ $x_q(t) \in \mathbb{R}^{r_q}$ are *jointly mixing of order P* if the random vector

$$y(t) = \left[x_1^T(t)\ x_2^T(t)\ \ldots\ x_q^T(t)\right]^T \in \mathbb{R}^{r_1 + r_2 + \ldots + r_q} \qquad (14\text{-}34)$$

is mixing of order P. For complex random vectors $x(t) \in \mathbb{C}^r$ definition (14-33) is applied to $x_{\text{re}}(t) \in \mathbb{R}^{2r}$. For strictly stationary random variables $x(t) \in \mathbb{R}^r$, $t \in \mathbb{Z}$, definition (14-33) simplifies to the classical definition given in Brillinger (1981)

$$\begin{aligned} &\max_{t_k} \sum_{t_1, t_2, \ldots, t_{k-1} = -\infty}^{\infty} \left|\text{cum}(x_{[a_1]}(t_1), x_{[a_2]}(t_2), \ldots, x_{[a_{k-1}]}(t_{k-1}), x_{[a_k]}(t_k))\right| \\ &= \sum_{u_1, u_2, \ldots, u_{k-1} = -\infty}^{\infty} \left|\text{cum}(x_{[a_1]}(u_1), x_{[a_2]}(u_2), \ldots, x_{[a_{k-1}]}(u_{k-1}), x_{[a_k]}(0))\right| \end{aligned} \qquad (14\text{-}35)$$

where $u_i = t_i - t_k$, $i = 1, 2, \ldots, k-1$ (see Section 14.1, property 7). Note that the definition of mixing random variables (14-33) implies that

$$\sum_{t_1, t_2, \ldots, t_k = 1}^{K} \left|\mathrm{cum}(x_{[a_1]}(t_1), x_{[a_2]}(t_2), \ldots, x_{[a_{k-1}]}(t_{k-1}), x_{[a_k]}(t_k))\right| = O(K) \tag{14-36}$$

for $K \to \infty$. The converse is true only for strictly stationary random variables.

For strictly stationary random variables $x(t)$, the mixing condition (14-33) implies that the span of dependence over t is sufficiently small. That is, the random variables $x(t_1)$ and $x(t_2)$ become uncorrelated sufficiently fast (mixing of order 2) or independent (mixing of order ∞) as $t_2 - t_1 \to \infty$. Indeed, if the zero mean noise $x(t)$ is mixing of order 2, then $\mathrm{covar}(x(t_1), x(t_2)) = \mathrm{cum}(x(t_1), \bar{x}(t_2)) \to 0$ as $t_2 - t_1 \to \infty$. As $\mathrm{var}(x(t_1)) = \mathrm{var}(x(t_2)) > 0$ are independent of t_1, t_2, we have

$$\frac{\mathrm{covar}(x(t_1), x(t_2))}{\sqrt{\mathrm{var}(x(t_1))\mathrm{var}(x(t_2))}} \to 0 \text{ as } t_2 - t_1 \to \infty \tag{14-37}$$

and thus $x(t_1)$ and $x(t_2)$ are asymptotically uncorrelated. Similarly, if the noise is mixing of order ∞, then all the higher order correlations tend to zero as $t_2 - t_1 \to \infty$ and, hence, $x(t_1)$ and $x(t_2)$ are asymptotically independent. Note that the converse is not necessarily true: if $x(t_1)$ and $x(t_2)$ are asymptotically ($t_2 - t_1 \to \infty$) independent, then this does not imply that $x(t)$ is mixing of order ∞. Indeed, the mixing condition (14-33) also imposes conditions on how fast the random variables should become independent. For general nonstationary noise, the mixing condition (14-33) is not sufficient to ensure that $x(t_1)$ and $x(t_2)$ are asymptotically uncorrelated or independent. To be asymptotically uncorrelated, we must assume, in addition, that the variance of the noise does not decrease to zero or decreases to zero more slowly than the covariance, so that (14-37) is fulfilled.

14.4.2 Properties

Besides its generality, the power of the mixing assumption lies in the property that a linear combination of powers of mixing variables is also mixing. These properties are formalized in the following lemmas.

Lemma 14.3: Let $x(t) \in \mathbb{C}$ be mixing of order P and $\alpha(t) \in \mathbb{C}$ nonrandom numbers. If $\max_t |\alpha(t)| \le c < \infty$, then $\alpha(t)x(t)$ is mixing of order P.

Proof. Follows directly from definition (14-33). □

Lemma 14.4: Let $x_k(t) \in \mathbb{C}^r$, $k = 1, 2, \ldots, q$, be jointly mixing (over t) random variables of order P. The linear combination $y(t) = \sum_{k=1}^{q} \alpha_k z_k(t)$ with $z_k(t) = x_k(t)$ and/or $z_k(t) = \bar{x}_k(t)$ is mixing of order P.

Proof. Apply the multilinearity property 2 of the cumulants (see Section 14.1). □

Lemma 14.5: Let $x(t) \in \mathbb{C}$ be mixing of order P and $y(t) = \sum_{u=0}^{\infty} h(t, u)x(u)$. If $h(t, u) \in \mathbb{C}$ is absolutely summable with respect to t and u

$$\begin{aligned} &\forall u\colon \sum_{t=0}^{\infty} |h(t, u)| \le C < \infty \qquad &\text{(a)} \\ &\forall t\colon \sum_{u=0}^{\infty} |h(t, u)| \le C < \infty \qquad &\text{(b)} \end{aligned} \tag{14-38}$$

with C a constant independent of t, u then, $y(t)$ is mixing of order P.

Proof. See Appendix 14.B. □

Lemma 14.5 says that mixing noise, filtered by a linear time-variant system with absolutely summable impulse response, remains mixing of the same order. Extension of Lemma 14.5 to the multivariable case is direct (Exercise 14.19). The real multivariable time-invariant version of Lemma 14.5 can be found in Brillinger (1981).

Example 14.6: Let $y(t)$ be generated by independent, identically distributed noise $e(t)$ passing through a stable discrete time filter. If $e(t)$ has bounded Pth order moments, then $y(t)$ is mixing of order P. If $e(t)$ is uniformly bounded or belongs to the exponential family of distributions (for example, exponential, gamma, Laplace, Gaussian), then $y(t)$ is mixing of order ∞. Indeed, for independent noise, the mixing condition (14-33)

$$\max_{t_k} \sum_{t_1, t_2, \ldots, t_{k-1} = 0}^{\infty} |\text{cum}(e(t_1), \ldots, e(t_{k-1}), e(t_k))| = |\text{cum}(e(t_k), \ldots, e(t_k))| < \infty \tag{14-39}$$

boils down to the existence of the kth order moment. All higher order moments exist for uniformly bounded noise and noise belonging to the exponential family of distributions (Stuart and Ord, 1987). □

Corollary 14.7: Let $x \in \mathbb{C}^N$, $H \in \mathbb{C}^{N \times N}$, and $y = Hx$. If $x_{[t]}$ is mixing over t of order P for $N = 1, 2, \ldots, \infty$ and $\|H\|_1 \le c < \infty$, $\|H\|_\infty \le c < \infty$ for $N = 1, 2, \ldots, \infty$, with c a constant independent of N, then $y_{[t]}$ is mixing over t of order P for $N = 1, 2, \ldots, \infty$.

Proof. Follows directly from Lemma 14.5. □

Lemma 14.8: Let $x(t) \in \mathbb{C}$ be mixing of order qP and $y(t) \in \mathbb{C}$ defined as

$$y(t) = \sum_{m=0}^{q} \sum_{u_1, u_2, \ldots, u_m = 0}^{t} h_m(t - u_1, t - u_2, \ldots, t - u_m)x(u_1)x(u_2)\ldots x(u_m) \tag{14-40}$$

If the $h_m(t_1, t_2, \ldots, t_m) \in \mathbb{C}$, $m = 0, 1, \ldots, q$, are absolutely summable

$$\sum_{t_1, t_2, \ldots, t_m = 0}^{\infty} |h_m(t_1, t_2, \ldots, t_m)| \le C_m < \infty \tag{14-41}$$

and $q < \infty$, then $y(t)$ is mixing of order P.

Proof. See Appendix 14.C. □

Lemma 14.8 says that mixing noise of order qP passing through a time-invariant nonlinear system that can be described by a Volterra functional of finite degree q with absolutely summable impulse responses is mixing of order P. Extension of Lemma 14.8 to the multivariable systems is direct. As a special case of the multivariable result, we have the following lemma.

Lemma 14.9: Let $x_k(t) \in \mathbb{C}$, $k = 1, 2, \dots, q$, be jointly mixing (over t) of order qP. The product $y(t) = \prod_{k=1}^{q} x_k(t)$ is mixing of order P.

Special cases of Lemma 14.9 are $x^q(t)$, $x^r(t)\bar{x}^s(t)$ with $r + s = q$, and more generally, the product $\prod_{k=1}^{q} z_k(t)$ of q delayed and/or complex conjugate factors $z_k(t) = x(t - t_k)$ and/or $z_k(t) = \bar{x}(t - t_k)$.

14.5 PRELIMINARY EXAMPLE

Consider, again, the resistance measurement problem of Section 1.2. Assume that M independent experiments are made of N current and voltage measurements each,

$$i^{[r]}(k) = i_0 + n_i^{[r]}(k) \text{ and } u^{[r]}(k) = u_0 + n_u^{[r]}(k) \tag{14-42}$$

$k = 1, 2, \dots, N$, $r = 1, 2, \dots, M$. Assume, furthermore, that $i_0 = 1$ A, $u_0 = 1$ V, and that $n_i^{[r]}(k)$, $n_u^{[r]}(k)$ are independent (over the measurements k and the experiments r) uniformly distributed random variables in the intervals $[-0.5 \text{ A}, 0.5 \text{ A}]$ and $[-0.5 \text{ V}, 0.5 \text{ V}]$, respectively. Invoke the simple approach (1-1), least squares (1-2), and errors-in-variables (1-3) estimators, proposed in Section 1.2, for each experiment $r = 1, 2, \dots, M$, giving M independent realizations of the estimates

$$\hat{R}_{\text{SA}}^{[r]}(N) = \frac{1}{N}\sum_{k=1}^{N} \frac{u^{[r]}(k)}{i^{[r]}(k)} \tag{14-43}$$

$$\hat{R}_{\text{LS}}^{[r]}(N) = \frac{\frac{1}{N}\sum_{k=1}^{N} u^{[r]}(k) i^{[r]}(k)}{\frac{1}{N}\sum_{k=1}^{N} (i^{[r]}(k))^2} \tag{14-44}$$

$$\hat{R}_{\text{EV}}^{[r]}(N) = \frac{\frac{1}{N}\sum_{k=1}^{N} u^{[r]}(k)}{\frac{1}{N}\sum_{k=1}^{N} i^{[r]}(k)} \tag{14-45}$$

Figure 14-1(a) shows the evolution of errors-in-variables resistance estimate (14-45) as a function of the number of measurements N for the first experiment ($r = 1$). The basic question that arises now is: "Does the estimate converge along this particular realization?" That is,

$$\lim_{N \to \infty} \hat{R}_{\text{EV}}^{[1]}(N) = R_{\text{EV}}^{[1]} \tag{14-46}$$

Scrutiny of the errors-in-variables estimates of the first five experiments (Figure 14-1(b)) may cause one to wonder to which realizations the estimates converge (also called *pointwise convergence*)

$$\lim_{N \to \infty} \hat{R}_{\text{EV}}^{[r]}(N) = R_{\text{EV}}^{[r]} \qquad r = 1, 2, \dots, \tag{14-47}$$

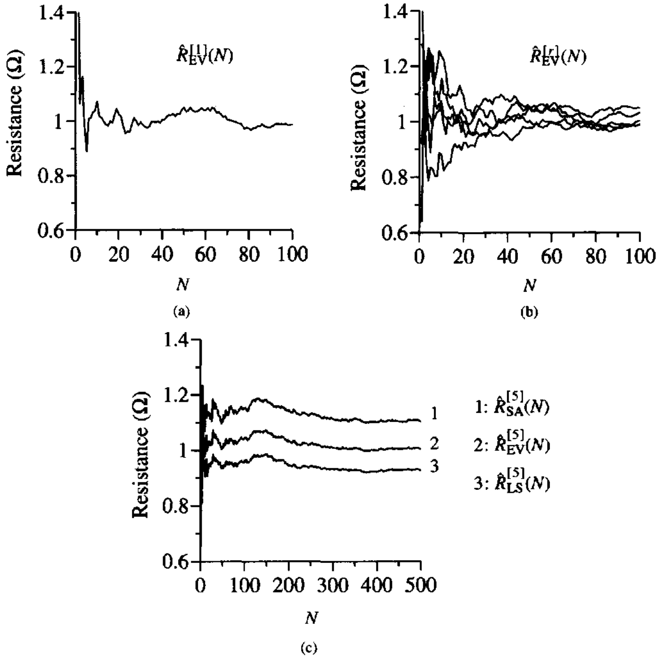


Figure 14-1. Measurement of a resistance: (a) errors-in-variables estimates of the first experiment, (b) errors-in-variables estimates of the first five experiments ($r = 1, 2, \ldots, 5$), (c) simple approach, errors-in-variables and least squares estimate of the fifth experiment.

whether they converge to the same value,

$$R^{[r]}_{EV} = R_{EV} \qquad r = 1, 2, \ldots \tag{14-48}$$

and whether this value equals the true value R_0

$$R_{EV} = R_0 \tag{14-49}$$

If (14-47) is true for "almost all realizations," then we have *stochastic convergence* to a random number R_{EV} and it makes sense to write

$$\underset{N \to \infty}{\text{"lim"}} \hat{R}_{EV}(N) = R_{EV} \tag{14-50}$$

If (14-48) is true for "almost all realizations," then R_{EV} in (14-50) is a deterministic (nonrandom) number. Several definitions of "lim" can be given according to what is meant by "almost all realizations." The precise definitions of the stochastic limits, their properties, and their interrelations will be discussed later on in this chapter. If (14-49) is true, then the resistance estimates are called *consistent*, which means that they converge to the true value as the number of processed measurements N tends to infinity.

Figure 14-1(c) shows the simple approach (14-43), least squares (14-44), and errors-in-variables (14-45) estimates of the fifth experiment ($r = 5$). Clearly, the simple approach and least squares estimates seem to converge to a value that deviates significantly from the true resistance value of 1 Ω. Referring also to Figure 1-4 on page 5, we may ask the following questions: "Is the mean value of the estimates asymptotically different from the true value?", "Is the uncertainty of the estimates (asymptotically) minimal?", and "What is the asymptotic distribution of the estimates?" The first question is strongly related to the consistency problem but is not completely equivalent (see Section 14.14). The second question is handled in Section 14.12, where it is shown that the uncertainty is bounded below by the Cramér-Rao bound. The central limit theorems of Section 14.10 will be helpful to answer the third question. The knowledge of the asymptotic distribution makes it possible to calculate confidence intervals on the estimates.

14.6 DEFINITIONS OF STOCHASTIC LIMITS

Let $x(N)$, $N = 1, 2, \ldots$ be a scalar random sequence. There are several ways in which the sequence might converge to a (random) number x as $N \to \infty$. We will define four modes of stochastic convergence.

1. The sequence $x(N)$, $N = 1, 2, \ldots$ converges to x *in mean square* if, $\mathcal{E}\{|x|^2\} < \infty$, $\mathcal{E}\{|x(N)|^2\} < \infty$ for all N, and $\lim_{N \to \infty} \mathcal{E}\{|x(N) - x|^2\} = 0$. We write

$$\underset{N \to \infty}{\text{l.i.m.}} x(N) = x \Leftrightarrow \lim_{N \to \infty} \mathcal{E}\{|x(N) - x|^2\} = 0 \tag{14-51}$$

2. The sequence $x(N)$, $N = 1, 2, \ldots$ converges to x *with probability 1* (w.p. 1) or *almost surely* if, $\lim_{N \to \infty} x^{[\omega]}(N) = x^{[\omega]}$ for almost all realizations ω, except those $\omega \in \mathbb{A}$ such that $\text{Prob}(\mathbb{A}) = 0$. We write

$$\underset{N \to \infty}{\text{a.s.lim}} x(N) = x \Leftrightarrow \text{Prob}(\lim_{N \to \infty} x(N) = x) = 1 \tag{14-52}$$

This definition is equivalent to (Theorem 2.1.2 of Lukacs, 1975)

$$\underset{N \to \infty}{\text{a.s.lim}} x(N) = x \Leftrightarrow \forall \varepsilon > 0\colon \lim_{N \to \infty} \text{Prob}(\sup_{k \geq N} |x(k) - x| \leq \varepsilon) = 1 \tag{14-53}$$

3. The sequence $x(N)$, $N = 1, 2, \ldots$ converges to x *in probability* if, for every $\varepsilon, \delta > 0$ there exists an N_0 such that for every $N > N_0$: $\text{Prob}(|x(N) - x| \leq \varepsilon) > 1 - \delta$. We write

$$\underset{N \to \infty}{\text{plim}}\, x(N) = x \Leftrightarrow \forall \varepsilon > 0\colon \lim_{N \to \infty} \text{Prob}(|x(N) - x| \leq \varepsilon) = 1 \tag{14-54}$$

4. Let $F_N(x)$ and $F(x)$ be the distribution functions of, respectively, $x(N)$ and x. The sequence $x(N)$, $N = 1, 2, \ldots$ converges to x *in law* or *in distribution* if $F_N(x)$ converges weakly[1] to $F(x)$. We write

$$\underset{N \to \infty}{\text{Lim}}\, x(N) = x \Leftrightarrow \underset{N \to \infty}{\text{Lim}}\, F_N(x) = F(x) \tag{14-55}$$

[1] This means at all continuity points of the limiting function and is denoted by "Lim."

14.7 INTERRELATIONS BETWEEN STOCHASTIC LIMITS

In the previous section we defined several modes of stochastic convergence. The connections between these concepts are

1. Almost sure convergence implies convergence in probability; the converse is not true (Theorem 2.2.1 of Lukacs, 1975; see also Appendix 14.D).
2. Convergence in mean square implies convergence in probability; the converse is not true (Theorem 2.2.2 of Lukacs, 1975; see also Appendix 14.E).
3. Convergence in probability implies convergence in law; the converse is not true (Theorem 2.2.3 of Lukacs, 1975).
4. There is no implication between almost sure and mean square convergence.
5. A sequence $x(N)$ converges in probability to x if and only if every subsequence $x(N_k)$ contains a sub-subsequence $x(N_{k_i})$ that converges $(i \rightarrow \infty)$ almost surely to x (Theorem 2.4.4 of Lukacs, 1975).
6. A sequence converges in probability to a constant if and only if it converges in law to a degenerate distribution[1] (Corollary to Theorem 2.2.3 of Lukacs, 1975).

A graphical representation of the convergence area of the different stochastic limits is given in Figure 14-2. The interrelations between the concepts are summarized in Figure 14-3. As these allow a better understanding of the stochastic limits, some proofs are given in the appendixes. The importance of interrelation 5 is that any theorem proved for the almost sure limit is also valid for the limit in probability. Before illustrating some of the interrelations by (counter) examples, we cite the Borel-Cantelli and the Fréchet-Shohat lemmas, which are useful to establish, respectively, convergence w.p. 1 and convergence in distribution. The

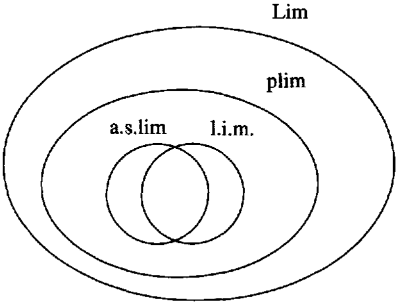


Figure 14-2. Convergence area of the stochastic limits.

$$\begin{tikzcd}
\text{a.s.lim} \arrow[r, Rightarrow, shift right] & \text{plim} \arrow[l, Rightarrow, shift right, "\text{On a subsequence}"'] \arrow[d, Rightarrow, shift right] & \text{l.i.m.} \arrow[l, Rightarrow] \\
 & \text{Lim} \arrow[u, Rightarrow, shift right, "\text{Conv. to a degenerate distribution}"'] &
\end{tikzcd}$$

Figure 14-3. Interrelations between the stochastic limits.

[1] $F(x)$ is degenerate if there exists an x_0 such that $F(x) = 0$ for $x < x_0$ and $F(x) = 1$ for $x \geq x_0$.

Borel-Cantelli lemma roughly says that if the convergence in probability or in mean square is sufficiently fast, this implies convergence with probability 1.

Lemma 14.10 (Borel-Cantelli Lemma): If

$$\sum_{N=1}^{\infty} \mathrm{Prob}(|x(N) - x| > \varepsilon) < \infty \text{ or } \sum_{N=1}^{\infty} \mathcal{E}\{|x(N) - x|^2\} < \infty \tag{14-56}$$

then $x(N)$ converges to x w.p. 1.

Proof. Theorems 2.1.1 and 2.1.3 of Stout (1974); see also Appendix 14.F. □

Lemma 14.11 (Fréchet-Shohat Lemma): Let x have a distribution function $F(x)$ that is uniquely determined by its moments (cumulants). If the moments (cumulants) of the sequence $x(N)$ converge for $N \to \infty$ to the moments (cumulants) of x, then $x(N)$ converges in distribution to x.

Proof. Theorem 1, Section 8.2 of Chow and Teicher (1988). □

Example 14.12: Convergence w.p. 1 and convergence in probability do not imply convergence in mean square (Example 2.1.1 of Stout). Take ω to be uniform in $[0, 1]$, and build the sequence $x(N)$ such that

$$x^{[\omega]}(N) = \begin{cases} N & \omega \in [0, 1/N) \\ 0 & \omega \in [1/N, 1] \end{cases} \tag{14-57}$$

Two realizations of the sequence are, for example,

$$\{x^{[0.3]}(N)\} = \{1, 2, 3, 0, 0, 0, 0, 0, \ldots\}$$
$$\{x^{[0.15]}(N)\} = \{1, 2, 3, 4, 5, 6, 0, 0, \ldots\}$$

We see that $x^{[\omega]}(N)$ is zero for N sufficiently large, which suggests that it will converge to zero. Formally, $\underset{N\to\infty}{\mathrm{plim}}\, x(N) = \underset{N\to\infty}{\mathrm{a.s.lim}}\, x(N) = 0$ since

$$\mathrm{Prob}(\sup_{k \geq N} |x(k)| \leq \varepsilon) = \mathrm{Prob}(|x(N)| \leq \varepsilon) = \mathrm{Prob}(x(N) = 0) = 1 - 1/N$$

is arbitrarily close to 1 for N sufficiently large. There is just one sequence, $x^{[0]}(N)$, that does not converge. This is not in contradiction with the previous results because the probability of getting this particular realization is zero: $\mathrm{Prob}(\omega = 0) = 0$. The mean square limit $\underset{N\to\infty}{\mathrm{l.i.m.}}\, x(N)$ does not exist because $\mathcal{E}\{x^2(N)\} = N$ is unbounded. Note that the Borel-Cantelli lemma cannot be used in this example to establish the almost sure convergence from the convergence in probability. Indeed, $\sum_{N=1}^{\infty} \mathrm{Prob}(|x(N) > \varepsilon|) = \sum_{N=1}^{\infty} 1/N = \infty$. □

Example 14.13: Convergence in probability and convergence in mean square do not imply convergence w.p. 1 (Example 2.1.2 of Stout, 1974). Take ω to be uniform in $[0, 1)$, and build the sequence $T(n, k)$ such that

$$T^{[\omega]}(n, k) = \begin{cases} 1 & \omega \in [(k-1)/n, k/n) \\ 0 & \text{elsewhere} \end{cases}$$

for $k = 1, 2, \ldots, n$ and $n \geq 1$. Let

$$\{x(N)\} = \{\{T(1,k)\}, \{T(2,k)\}, \{T(3,k)\}, \ldots\}$$

with $\{T(n,k)\} = \{T(n,1), T(n,2), \ldots, T(n,n)\}$ and $N = n(n-1)/2 + k$. Two realizations of the sequence are, for example,

$$\begin{aligned}\{x^{[0.27]}(N)\} &= \{\{1\}, \{1,0\}, \{1,0,0\}, \{0,1,0,0\}, \ldots\}\\ \{x^{[0.85]}(N)\} &= \{\{1\}, \{0,1\}, \{0,0,1\}, \{0,0,0,1\}, \ldots\}\end{aligned}$$

We see that the length of each subsequence $\{T(n,k)\}$ of $\{x(N)\}$ increases with n and that it contains exactly one nonzero term. This suggests that $x(N)$ will converge in probability (the probability to get a 1 goes to zero), but not w.p. 1 (the supremum is 1 for any value of N). Formally, $\underset{N\to\infty}{\text{plim}}\, x(N) = 0$ since

$$\lim_{N\to\infty} \text{Prob}(|x(N)| \leq \varepsilon) = \lim_{N\to\infty} \text{Prob}(T(n,k) = 0) = \lim_{N\to\infty} (1 - 1/n) = 1$$

and $\underset{N\to\infty}{\text{l.i.m.}}\, x(N) = 0$ because

$$\lim_{N\to\infty} \mathcal{E}\{x^2(N)\} = \lim_{N\to\infty} \mathcal{E}\{T^2(n,k)\} = \lim_{N\to\infty} 1/n = 0$$

The almost sure limit $\underset{N\to\infty}{\text{a.s.lim}}\, x(N)$ does not exist since $\text{Prob}(\sup_{r\geq N} |x(r)| > \varepsilon) = 1$. Note that the subsequence $T(n,k)$, with k fixed and $n \geq 1$, converges w.p. 1 to zero. This is an illustration of interrelation 5. □

Example 14.14: Convergence in mean square and convergence w.p. 1 are compatible (Example 2.2.3 of Lukacs, 1975). Let $x(N)$ be a random variable that assumes only the values $1/N$ and $-1/N$ with equal probability. We find $\underset{N\to\infty}{\text{l.i.m.}}\, \mathcal{E}\{x^2(N)\} = 0$ since

$$\lim_{N\to\infty} \mathcal{E}\{x^2(N)\} = \lim_{N\to\infty} 1/N^2 = 0$$

Also $\underset{N\to\infty}{\text{a.s.lim}}\, x(N) = 0$ because $|x(k)| < |x(N)|$ for any $k > N$ so that

$$\text{Prob}(\sup_{k\geq N} |x(k)| \leq \varepsilon) = \text{Prob}(|x(N)| \leq \varepsilon)\big|_{N > 1/\varepsilon} = 1 \qquad \square$$

Example 14.15: Convergence in distribution does not imply convergence in probability (Example 2.2.4 of Lukacs, 1975). Let x be a random variable that can take only the values 0 and 1 with equal probability. Next, construct the sequence $x(N) = 1 - x$. We have $\underset{N\to\infty}{\text{Lim}}\, x(N) = x$ because $x(N)$ and x have the same distribution functions $F_N(x) = F(x)$. However, the limit in probability $\underset{N\to\infty}{\text{plim}}\, x(N)$ does not exist because $|x(N) - x| = 1$ so that $\text{Prob}(|x(N) - x| \leq \varepsilon) = 0$. □

14.8 PROPERTIES OF STOCHASTIC LIMITS

The properties of the stochastic limits are similar to those of the classical (deterministic) limit, but there are some subtle differences. The general properties are

1. A continuous function and the almost sure limit may be interchanged

$$\underset{N\to\infty}{\text{a.s.lim}} f(x(N)) = f(x) \text{ with } x = \underset{N\to\infty}{\text{a.s.lim}} x(N) \tag{14-58}$$

2. The almost sure limit and the expected value may be interchanged for uniformly bounded sequences (Theorem 5.4 of Billingsley, 1995)

$$\lim_{N\to\infty} \mathcal{E}\{x(N)\} = \mathcal{E}\{\underset{N\to\infty}{\text{a.s.lim}} x(N)\} \tag{14-59}$$

A direct consequence of (14-59) is that

$$\mathcal{E}\{O_{\text{a.s.}}(N^{-k})\} = O(N^{-k}) \tag{14-60}$$

3. A continuous function and the limit in probability may be interchanged (Theorem 2.3.3 of Lukacs, 1975)

$$\underset{N\to\infty}{\text{plim}} f(x(N)) = f(x) \text{ with } x = \underset{N\to\infty}{\text{plim}} x(N) \tag{14-61}$$

4. The limit in probability and the expected value may be interchanged for uniformly bounded sequences (Theorem 5.4 of Billingsley, 1995)

$$\lim_{N\to\infty} \mathcal{E}\{x(N)\} = \mathcal{E}\{\underset{N\to\infty}{\text{plim}} x(N)\} \tag{14-62}$$

A direct consequence of (14-62) is that

$$\mathcal{E}\{O_{\text{p}}(N^{-k})\} = O(N^{-k}) \tag{14-63}$$

5. The mean square limit is linear (Theorem 3.1 of Jazwinski, 1970)

$$\underset{N\to\infty}{\text{l.i.m.}}(ax(N) + by(N)) = a\underset{N\to\infty}{\text{l.i.m.}} x(N) + b\underset{N\to\infty}{\text{l.i.m.}} y(N) \tag{14-64}$$

where a and b are deterministic (nonrandom) numbers.

6. The mean square limit and the expected value may be interchanged (Theorem 3.1 of Jazwinski, 1970),

$$\lim_{N\to\infty} \mathcal{E}\{x(N)\} = \mathcal{E}\{\underset{N\to\infty}{\text{l.i.m.}} x(N)\} \tag{14-65}$$

A direct consequence of (14-65) is that

$$\mathcal{E}\{O_{\text{m.s.}}(N^{-k})\} = O(N^{-k}) \tag{14-66}$$

7. If $\underset{N\to\infty}{\text{l.i.m.}}x(N) = x$ and $\mathcal{E}\{(x(N)-x)^2\} = O(N^{-k})$, with $k > 0$, then

$$x(N) = x + O_{\text{m.s.}}(N^{-k/2}) \text{ and } x(N) = x + O_{\text{p}}(N^{-k/2}) \tag{14-67}$$

This is a direct consequence of (14-66) and interrelation 2, Section 14.7.

8. If the sequence $x(n)$ is deterministic (nonrandom), then the limit in mean square, the limit w.p. 1, and the limit in probability reduce to the deterministic limits.

Property 1 follows directly from the definition (14-52) of convergence w.p. 1, while property 3 follows from interrelation 5, Section 14.7, and property 1. Properties 1 and 3 require the continuity of the function at ALL values of the limit random variable x. If x is a constant (nonrandom), then continuity in a closed neighborhood of x is sufficient. Note that the limit in mean square and a continuous function may, in general, NOT be interchanged. Note also that the almost sure limit and the limit in probability, in general, do NOT commute with the expected value.

14.9 LAWS OF LARGE NUMBERS

The classical laws of large numbers are used to study the stochastic convergence of the partial sum $S(N) = \sum_{k=1}^{N} x(k)$ of a random sequence $x(k)$, with $x(k)$ independent of N. They state roughly that $S(N)$ converges to its expected value if the span of dependence of $x(k)$ is limited. In this book, we often need the more general case where the sequence $x(k)$ in the partial sum $S(N)$ depends on the number of samples N: $S(N) = \sum_{k=1}^{N} x_N(k)$. According to the stochastic limit used to establish the convergence, we speak about the *weak law of large numbers*,

$$\underset{N\to\infty}{\text{plim}}\,(S(N) - \mathcal{E}\{S(N)\})/N = 0 \tag{14-68}$$

the *strong law of large numbers*,

$$\underset{N\to\infty}{\text{a.s.lim}}(S(N) - \mathcal{E}\{S(N)\})/N = 0 \tag{14-69}$$

and the *law of large numbers*

$$\underset{N\to\infty}{\text{l.i.m.}}(S(N) - \mathcal{E}\{S(N)\})/N = 0 \tag{14-70}$$

Note that the (strong) laws of large numbers (14-69) and (14-70) imply the weak law of large numbers (14-68) (see Section 14.7, interrelations 1 and 2). The analysis of the rate at which $S(N)/N$ converges to its expected value requires some additional assumptions. For the strong law of large numbers (14-69), this rate is given by the *law of the iterated logarithm*

$$S(N)/N = \mathcal{E}\{S(N)\}/N + O_{\text{a.s.}}(N^{-1/2}\sqrt{\ln(\ln(N))}) \tag{14-71}$$

For the law of large numbers (14-70) we have, typically,

$$S(N)/N = \mathcal{E}\{S(N)\}/N + O_{\text{m.s.}}(N^{-1/2}) \tag{14-72}$$

Some interesting versions of the laws of large numbers and their respective convergence rates are listed next. More versions can be found in Chow and Teicher (1988), Lukacs (1975), and Stout (1974).

1. If $x(k)$ is *independent and identically distributed (iid),* then (14-69) applies if and only if $\mathcal{E}\{x(k)\} = x < \infty$ (Theorem 4.3.3 of Lukacs, 1975). If in addition $\text{var}(x(k)) = \sigma^2 < \infty$, then the convergence rate of (14-69) is given by (14-71) (Theorem 3.2.9 of Stout, 1974).
2. If $x(k)$ is *independent,* then

   2a. (14-68) and (14-69) are equivalent (Theorem 2.13.2 of Stout, 1974).

   2b. (14-69) applies if $\text{var}(x(k)) \leq M < \infty$ for any k (Corollary 1 to Theorem 4.3.1 of Lukacs, 1975). If, in addition, $\text{var}(S(N)) = O(N)$ and for some $\delta > 0$,

$$\sum_{N=1}^{\infty} \text{Prob}(|x(N) - \mathcal{E}\{x(N)\}| > \sqrt{\delta N \ln(\ln(N))}) < \infty \tag{14-73}$$

   then the convergence rate of (14-69) is given by (14-71) (Corollary 3, Section 10.2 of Chow and Teicher, 1988)

3. If $x_N(k)$, $N = 1, 2, \ldots, \infty$, is *mixing of order 2* and depends on N, then the law of large numbers (14-70) applies, and its convergence rate is given by (14-72) (proof: see Appendix 14.G). For the strong law of large numbers, the variations of the sequence $x_N(k)$ w.r.t. N should, in addition, be "small enough": if $\text{var}(\sum_{k=1}^{s} x_r(k) - x_s(k)) = O(r-s)$, $r \geq s$, then (14-69) applies (proof: see Appendix 14.G).

The uniformly boundedness condition on the variances in versions 2b and 3 of the law of large numbers is necessary to avoid any increase in the variance of the sequence $x(k)$ to infinity. Otherwise, the uncertainty on the partial sum would not decrease to zero, making it impossible, in general, for $S(N)/N$ to converge to its expected value.

Because the almost sure limit imposes some restrictions on the supremum (see (14-53)), the convergence rate of the strong law of large numbers depends upon the tails of the probability density functions (pdf's) of the random variables $x(k)$. Condition (14-73) dictates that the tails of the pdf's tend sufficiently fast to zero. It is satisfied for uniformly bounded random variables.

Example 14.16: Let $x \in \mathbb{C}^N$, $H \in \mathbb{C}^{N \times N}$, and $y = Hx$, where x, H satisfy the assumptions of Corollary 14.7 with $P = 2$, and where $H_{[i,j]}$ is independent of N for any i, j. The partial sum $S(N) = \sum_{k=1}^{N} y_N(k)$, with $y_N(k) = y_{[k]}$, satisfies the strong law of large numbers (14-69). Indeed,

$$
\begin{aligned}
\operatorname{var}\Big(\sum_{k=1}^{s} y_r(k) - y_s(k)\Big) &= \operatorname{var}\Big(\sum_{k=1}^{s}\sum_{l=s+1}^{r} H_{[k,l]}x_{[l]}\Big) \\
&= \sum_{l_1,l_2=s+1}^{r} \operatorname{cum}(x_{[l_1]}, \bar{x}_{[l_2]}) \sum_{k_1=1}^{s} H_{[k_1,l_1]} \sum_{k_2=1}^{s} \bar{H}_{[k_2,l_2]} \\
&\le \left(\max_{l} \sum_{k=1}^{s} \left|H_{[k,l]}\right|\right)^2 \sum_{l_1,l_2=s+1}^{r} \left|\operatorname{cum}(x_{[l_1]}, \bar{x}_{[l_2]})\right| \\
&\le O(r-s)
\end{aligned}
$$

The last inequality is due to the finite 1-norm of H and the second-order mixing property of $x_{[k]}$ (14-36). □

14.10 CENTRAL LIMIT THEOREMS

The classical central limit theorems pertain to the asymptotic distribution function of the partial sum $S(N) = \sum_{k=1}^{N} x(k)$ of a random sequence $x(k)$. They state, roughly, that $S(N)$ is asymptotically normally distributed

$$\operatorname*{Lim}_{N\to\infty} \frac{S(N) - \mathcal{E}\{S(N)\}}{\sqrt{\operatorname{var}(S(N))}} \in N(0, 1) \tag{14-74}$$

if each $x(k)$ has high probability to be of the same order of magnitude and if the span of dependence of $x(k)$ is limited. Under some additional assumptions, the rate at which the distribution function of $S(N)$ converges to a normal distribution can be established. It is given by the Berry and Esseen theorem

$$\sup_{y}\left|F_N(y) - \Phi(y)\right| \le O(N^{-1/2}) \tag{14-75}$$

with $F_N(y)$ the distribution function of $(S(N) - \mathcal{E}\{S(N)\})/\sqrt{\operatorname{var}(S(N))}$ and $\Phi(y)$ the standard normal distribution function. In this book we often need the more general case where the sequence $x(k)$ in the partial sum $S(N)$ depends on the number of samples N: $S(N) = \sum_{k=1}^{N} x_N(k)$. Some interesting versions of the central limit theorem are listed next. More versions can be found in Billingsley (1995) and Feller (1968).

1. If $x(k)$ is *independent and identically distributed* with finite mean $\mu < \infty$ and finite nonzero variance $0 < \sigma^2 < \infty$, then (14-74) applies with $\mathcal{E}\{S(N)\} = N\mu$ and $\operatorname{var}(S(N)) = N\sigma^2$ (Theorem 27.1 of Billingsley, 1995). If, in addition, $\mathcal{E}\{\left|x(k)\right|^3\} < \infty$, then the convergence rate of (14-74) is given by (14-75) (Theorem 9.1.3 of Chow and Teicher, 1988).

2. $x(k)$ is *independent,* with finite means $\mu_k \le c_1 < \infty$ and finite variances $\sigma_k^2 \le c_2 < \infty$. If for some $\varepsilon > 0$ $x(k)$ has uniformly bounded $2+\varepsilon$ moments $\mathcal{E}\{\left|x(k)\right|^{2+\varepsilon}\} \le C < \infty$ and if $N/(\sqrt{\operatorname{var}(S(N))})^{2+\varepsilon} = o(N^0)$, then (14-74) applies with $\mathcal{E}\{S(N)\} = \sum_{k=1}^{N} \mu_k$ and $\operatorname{var}(S(N)) = \sum_{k=1}^{N} \sigma_k^2$. (Theorem 27.3 of Billingsley, 1995). If, in addition, $\mathcal{E}\{\left|x(k)\right|^3\} \le c_3 < \infty$, then the convergence rate of (14-74) is given by (14-75) (Theorem 9.1.3 of Chow and Teicher, 1988).

3. If $x(k)$ is *m-dependent,* then (14-74) is valid under the same conditions of version 2 of the central limit theorem (Orey, 1958; Rosén, 1967).
4. If $x_N(k)$, $N = 1, 2, \ldots, \infty$, is *mixing of order infinity* and if $\text{var}(S(N)) = O(N)$, then (14-74) and (14-75) apply (proof: see Appendix 14.H for the nonstationary case; see Theorem 4.4.1 of Brillinger (1981) for the stationary case).

The conditions $N/(\sqrt{\text{var}(S(N))})^{2+\varepsilon} = o(N^0)$ and $\text{var}(S(N)) = O(N)$ in, respectively, versions 2 and 3 and version 4 of the central limit theorem are necessary to avoid dominance of a few random variables over the partial sum $S(N)$. If not, the distribution function of $S(N)$ would be determined by the distribution functions of those few dominating random variables and would in general not be normal. Extension of these theorems to the complex and to the (complex) multivariate case is straightforward.

The central limit theorem should be interpreted with some care. The following example will illustrate this.

Example 14.17: Suppose that we take N independent samples of a uniformly distributed random variable $x(k) \in U(0, \sigma^2)$. According to version 1 of the central limit theorem (14-74), the mean value will be asymptotically normally distributed. Figure 14-4 compares the true probability density function of $S(N)/N$ (solid line) with the Gaussian pdf predicted by the central limit theorem (dashed line) for the case $\sigma = 5/\sqrt{3}$. It follows that even for small values of N the Gaussian approximation is remarkably within the interval $[-5, 5]$. Although the mean $S(N)/N$ cannot take values outside the interval $[-\sqrt{3}\sigma, \sqrt{3}\sigma]$, the central limit theorem predicts that this will happen with some (small) probability. Similarly, saying that the weight of newborn babies is normally distributed does not imply that there is a small risk of getting babies with a negative weight! We conclude that the central limit theorem describes, very well, the behavior of the distribution function around its mean value but not at its tails. □

14.11 PROPERTIES OF ESTIMATORS

What kind of properties do we expect from a "good" estimator? It would be nice for the estimated value $\hat{\theta}(N)$ to converge to the true value θ_0 as the number of noisy measurements N tends to infinity. We could also require that the expected value of $\hat{\theta}(N)$ equals the true value or that this is at least asymptotically ($N \to \infty$) valid. "Does the estimator have the smallest possible (asymptotic) mean square error?" and "Is its (asymptotic) distribution function known?" are also important issues. Besides, we would also like that most of (all) these properties remain valid if we do not satisfy some of (all) the basic assumptions made in constructing the estimator. The formal definitions are listed next.

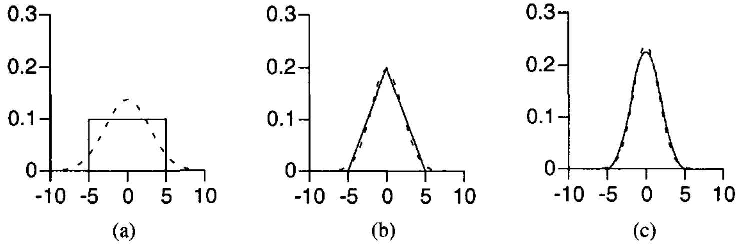


Figure 14-4. Comparison of the true pdf (solid line) and the Gaussian pdf predicted by the central limit theorem (dashed line) of $S(N)/N$ for zero mean, independent uniformly distributed random variables $x(k)$ with $\sigma = 5/\sqrt{3}$: (a) $N = 1$, (b) $N = 2$, and (c) $N = 3$.

1. An estimator $\hat{\theta}(N)$ is *consistent* if it converges to the true value θ_0 as $N \to \infty$. According to the stochastic limit used, we say that $\hat{\theta}(N)$ is *weakly consistent* if

$$\operatorname*{plim}_{N \to \infty} \hat{\theta}(N) = \theta_0 \tag{14-76}$$

strongly consistent if

$$\operatorname*{a.s.lim}_{N \to \infty} \hat{\theta}(N) = \theta_0 \tag{14-77}$$

and *consistent* if

$$\operatorname*{l.i.m.}_{N \to \infty} \hat{\theta}(N) = \theta_0 \tag{14-78}$$

2. An estimator $\hat{\theta}(N)$ is *unbiased* if

$$\mathcal{E}\{\hat{\theta}(N)\} = \theta_0 \tag{14-79}$$

It is *asymptotically unbiased* if (14-79) is valid for $N \to \infty$.

3. An estimator $\hat{\theta}(N)$ is *(statistically) efficient* if, for all θ_0-values, the mean square error matrix of any other estimator $\hat{\psi}(N)$ is not smaller than that of $\hat{\theta}(N)$

$$\text{MSE}(\hat{\psi}(N)) \geq \text{MSE}(\hat{\theta}(N)) \tag{14-80}$$

It is *asymptotically efficient* if (14-80) is valid for $N \to \infty$. For unbiased estimators (14-80) becomes

$$\text{Cov}(\hat{\psi}(N)) \geq \text{Cov}(\hat{\theta}(N)) \tag{14-81}$$

4. The estimator $\hat{\theta}(N)$ is *(asymptotically) normally distributed*

$$\hat{\theta}(N) \to \tilde{\theta}(N) \in N^{n_\theta}(\mathcal{E}\{\tilde{\theta}(N)\}, \text{Cov}(\tilde{\theta}(N))) \tag{14-82}$$

5. An estimator $\hat{\theta}(N)$ is *robust* if one or more of the preceding properties remain unchanged when one or more of the basic assumptions made to construct the estimator are violated.

There is a fundamental difference between asymptotic unbiasedness and (weak or strong) consistency. Indeed, to be asymptotically unbiased, it is, for example, sufficient (but not necessary!) that the asymptotic probability density function $f_{\hat{\theta}}(\hat{\theta})$ of the estimate $\hat{\theta}$ satisfies $f_{\hat{\theta}}(\hat{\theta} - \theta_0) = f_{\hat{\theta}}(\theta_0 - \hat{\theta})$ (see Figure 14-5(a)), while (weak or strong) consistency requires that the asymptotic probability density function is a Dirac function (see Figure 14-5(b)).

The property that a continuous function may be interchanged with the almost sure limit and the limit in probability (see Section 14.8) explains why weak and strong consistency of the estimates are mostly proved. This is not the case for the limit in mean square, which is often used as an intermediate step in the consistency proofs (see Section 14.13). Note, however,

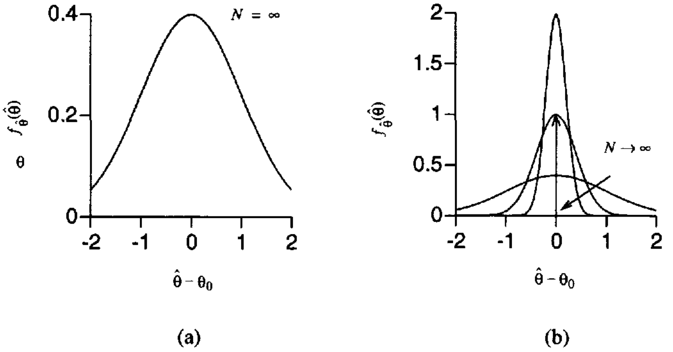


Figure 14-5. Asymptotic pdf of $\hat{\theta}$: (a) asymptotically unbiased estimator, (b) (weakly or strongly) consistent estimator for $N \to \infty$ (the limit pdf is a Dirac function).

that consistency (14-78) implies asymptotic unbiasedness (see Section 14.8, property 6), which is not the case for weak and strong consistency (see Section 14.14).

The practical importance of the efficiency property is that it makes no sense to look for estimators with a lower mean square error matrix. Although (asymptotic) efficiency is a highly desirable property, inefficient estimators with an acceptable accuracy may sometimes be the best practicable candidates (for example, if calculation time is important).

The existence of bias in the estimates is often the reason of the increased mean square error matrix compared with that of the unbiased estimates. However, simple examples of biased minimum mean square estimators exist that are statistically more efficient (have smaller MSE) than any other unbiased estimator (Kendall and Stuart, 1979; Stoica and Moses, 1990; see also Example 14.19 and Exercise 14.24). Minimum mean square error estimators have the following three drawbacks. First, they often require knowledge of the true (unknown) parameter values and, therefore, are not realizable (Kendall and Stuart, 1979; Norton, 1986). Next, if the estimation results are averaged in a second step, then the mean square error can only be reduced to the square of the bias as the number of averages tends to infinity (it can be reduced to zero for unbiased estimates). Finally, minimum mean square estimators are not robust w.r.t. the assumed underlying distribution function of the measurements. This explains why (asymptotically) unbiased estimators are usually preferred over minimum mean square estimators.

One should be very careful when interpreting the asymptotic normality property (14-82). It says that the estimated parameters $\hat{\theta}(N)$ converge in distribution to a random variable $\tilde{\theta}(N)$ that is (asymptotically) normally distributed. This does NOT imply the existence of the moments (expected value, variance, ...) of $\hat{\theta}(N)$ for any finite value of N. However, the asymptotic normality property makes it possible to calculate uncertainty bounds and confidence levels.

14.12 CRAMÉR-RAO LOWER BOUND

Consider the identification of the parameter vector $\theta \in \mathbb{R}^{n_\theta}$ using noisy measurements $z \in \mathbb{R}^N$. The quality of the estimator $\hat{\theta}(z)$ can be represented by its mean square error matrix

$$\mathrm{MSE}(\hat{\theta}(z)) = \mathrm{Cov}(\hat{\theta}(z)) + b_\theta b_\theta^T \tag{14-83}$$

where θ_0 and b_θ denote, respectively, the true value and the bias on the estimates. We may wonder whether there exists a lower limit on the value of the mean square error (14-83) that can be obtained with various estimators. The answer is given by the *generalized Cramér-Rao lower bound.*

Theorem 14.18: Let $f_z(z, \theta_0)$ be the probability density function of the measurements $z \in \mathbb{R}^N$. Assume that $f_z(z, \theta_0)$ and its first- and second-order derivatives w.r.t. $\theta \in \mathbb{R}^{n_\theta}$ exist for all θ_0-values. Assume, furthermore, that the boundaries of the domain of $f_z(z, \theta_0)$ w.r.t. z are θ_0 independent. Then, the *generalized Cramér-Rao lower bound* on the mean square error of any estimator $\hat{G}(z)$ of the function $G(\theta) \in \mathbb{C}^r$ of θ is

$$\mathrm{MSE}(\hat{G}(\hat{\theta}(z))) \geq \left(\frac{\partial G(\theta_0)}{\partial \theta_0} + \frac{\partial b_G}{\partial \theta_0}\right) Fi^{+}(\theta_0) \left(\frac{\partial G(\theta_0)}{\partial \theta_0} + \frac{\partial b_G}{\partial \theta_0}\right)^H + b_G b_G^H \tag{14-84}$$

with $b_G = \mathcal{E}\{\hat{G}(z)\} - G(\theta_0)$ the bias that might be present in the estimate, and $Fi(\theta_0)$ the *Fisher information matrix* of the parameters θ_0

$$Fi(\theta_0) = \mathcal{E}\left\{\left(\frac{\partial \ln f_z(z, \theta_0)}{\partial \theta_0}\right)^T \left(\frac{\partial \ln f_z(z, \theta_0)}{\partial \theta_0}\right)\right\} = -\mathcal{E}\left\{\frac{\partial^2 \ln f_z(z, \theta_0)}{\partial \theta_0^2}\right\} \tag{14-85}$$

Equality holds in (14-84) if and only if there exists a nonrandom matrix Γ such that

$$\hat{G}(\hat{\theta}(z)) - \mathcal{E}\{\hat{G}(\hat{\theta}(z))\} = \Gamma\left(\frac{\partial \ln f_z(z, \theta_0)}{\partial \theta_0}\right)^T \tag{14-86}$$

The expectations in (14-84) and (14-85) are taken w.r.t. the measurements z.

Proof. See Appendix 14.J. □

Note that the calculation of the Cramér-Rao lower bound requires knowledge of the true parameters θ_0, which is often not available (except in simulations). An approximation can be calculated by replacing θ_0 by its estimated value $\hat{\theta}$ in (14-84). Two special cases of the Cramér-Rao inequality are worth mentioning.

If $G(\theta) = \theta$, $b_G = 0$, and $Fi(\theta_0)$ is regular, then we obtain the *Cramér-Rao lower bound for unbiased estimators* (abbreviated as UCRB)

$$\mathrm{Cov}(\hat{\theta}(z)) \geq Fi^{-1}(\theta_0) \tag{14-87}$$

If condition (14-86) is not satisfied, $\hat{\theta}(z) - \theta_0 \neq \Gamma(\partial \ln f_z(z, \theta_0)/\partial \theta_0)^T$, then the lower bound (14-87) is too conservative, and there may still be an unbiased estimator that has smaller variance than any other unbiased estimator. Better (larger) bounds exist when (14-87) is not attainable, but they are often (extremely) difficult to compute. An overview of tighter bounds can be found in Abel (1993).

If $G(\theta) = \theta$, $b_G \neq 0$, and $Fi(\theta_0)$ is regular, then we find the *Cramér-Rao lower bound on the mean square error of biased estimators* (abbreviated as CRB)

$$\mathrm{MSE}(\hat{\theta}(z)) \geq \left(I_{n_\theta} + \frac{\partial b_\theta}{\partial \theta_0}\right) Fi^{-1}(\theta_0) \left(I_{n_\theta} + \frac{\partial b_\theta}{\partial \theta_0}\right)^T + b_\theta b_\theta^T \tag{14-88}$$

It follows that the Cramér-Rao lower bound for asymptotically unbiased estimators ($b_\theta \to 0$ as $N \to \infty$) is asymptotically given by (14-87) only if the derivative of the bias w.r.t. θ_0 is asymptotically zero. Likewise, in the unbiased case, the lower bound (14-88) may be too conservative and tighter bounds exist (Abel, 1993). Note that the first term in the right-hand side of (14-88) can be zero for biased estimators (see Example 14.20).

In general, it is impossible to show that the bias (and its derivative w.r.t. θ) of a weakly or strongly consistent estimator converges to zero as $N \to \infty$. However, the moments of the limiting random variable often exist. The (asymptotic) covariance matrix or mean square error of the limiting random variable is then compared with the UCRB. In this context, the concept of efficiency is also used for weakly or strongly consistent estimators.

Example 14.19: (Stoica and Moses, 1990) Let $z(k)$, $k = 1, 2, \ldots, N$ be zero mean iid Gaussian random variables, $z(k) \in N(0, \sigma^2)$. The sample variance $\hat{\sigma}_z^2 = \sum_{k=1}^{N} z^2(k)/N$ is an unbiased and efficient estimate of σ_z^2 (see Exercise 14.23). Now consider the estimator $\hat{s}^2 = a\hat{\sigma}_z^2$ where $a > 0$ is chosen to minimize the mean square error (14-7) of the estimate $\hat{s}^2$

$$\mathrm{MSE}(\hat{s}^2) = a^2 \mathrm{var}(\hat{\sigma}_z^2) + (a-1)^2 \sigma_z^4 \tag{14-89}$$

Minimizing (14-89) w.r.t. a gives $a = \sigma_z^4/\mathscr{E}\{\hat{\sigma}_z^4\} = N/(N+2)$ with corresponding minimum mean square error

$$\min_a \mathrm{MSE}(a) = 2\sigma_z^4/(N+2) \tag{14-90}$$

This should be compared with the UCRB

$$Fi^{-1}(\sigma_z^2) = \mathrm{var}(\hat{\sigma}_z^2) = 2\sigma_z^4/N \tag{14-91}$$

which is clearly larger than the mean square error (14-90) of the biased estimate $\hat{\sigma}_z^2 N/(N+2)$. It can also be verified that $\hat{\sigma}_z^2 N/(N+2)$ is statistically efficient in the sense that its mean square error reaches the lower bound (14-88). We conclude that the lower bound on the mean square error matrix of biased estimators may be smaller than the lower bound on the covariance matrix of unbiased estimators. □

Example 14.20: Assume that we estimate the weight of a bread from N noisy measurements. The true weight of the bread is 800 g. Regardless of what we measure, we estimate the weight as 100 g. Clearly, the estimator is biased and has zero variance. This is not in contradiction with the lower bound (14-88). Indeed,

$$I_{n_\theta} + \partial b_\theta/\partial\theta = \partial\mathscr{E}\{\hat{\theta}(z)\}/\partial\theta_0 = 0$$

because the estimate $\hat{\theta}(z)$ of the weight is independent of the true value θ_0. □

14.13 HOW TO PROVE ASYMPTOTIC PROPERTIES OF ESTIMATORS?

Ideally, we would like to know everything about the finite sample behavior (N does not increase to infinity) of an estimator. In practice, however, we can prove only a few large sample ($N \to \infty$) properties and hope that an estimator with good asymptotic properties also behaves well for practical sample sizes. The goal of this section is to present the main ideas and techniques without going into the mathematical details. The exact technical conditions and assumptions can be found in Chapter 15. We distinguish two different situations: an explicit (analytic) expression for the estimates $\hat{\theta} \in \mathbb{R}^{n_\theta}$ as a function of the measurements $z \in \mathbb{R}^N$ is available,

$$\hat{\theta}(z) = f(z) \tag{14-92}$$

or $\hat{\theta}$ is implicitly known through the minimization of a cost function $V(\theta, z)$

$$\hat{\theta}(z) = \arg\min_{\theta} V(\theta, z) \tag{14-93}$$

The explicit case (14-92) is illustrated on the resistance measurement problem in Section 14.15, and the implicit case (14-93) is elaborated in Chapter 15 on cost functions that are quadratic-in-the-measurements. We assume that the measurements z are disturbed by additive noise n_z

$$z = z_0 + n_z \tag{14-94}$$

with z_0 the true unknown value.

The following tools are essential in the analysis of the asymptotic properties of an estimator: the law of large numbers (Section 14.9) for the convergence and the consistency, the convergence rate of the law of large numbers (Section 14.9) for the convergence rate of the estimates, the interchangeability of the stochastic limit and the expected value (Section 14.8) for the asymptotic bias, the central limit theorem (Section 14.10) or the Fréchet-Shohat lemma (Lemma 14.11) for the asymptotic normality, and in general the interchangeability of a continuous function and a stochastic limit (Section 14.8).

14.13.1 Convergence—Consistency

In both cases (14-92) and (14-93) the estimate $\hat{\theta}(z)$ converges to some nonrandom number $\tilde{\theta}(z_0)$ by averaging of the disturbing noise n_z. Therefore, we first locate in $f(z)$ and $V(\theta, z)$ the sums that average the measurements z, and next use one of the laws of large numbers of Section 14.9 to prove the convergence of the sums to their expected values. Further analysis is done with the limit in probability or the limit with probability one, as they have the nice property of being interchangeable with a continuous function (see Section 14.8). Putting the stochastic sums in the vector $w \in \mathbb{R}^p$, with p independent of N, we can write this down formally as

$$\begin{aligned} &\hat{\theta}(z) = f(z) = \tilde{f}(z_0, w(n_z, z_0)) && \text{(a)} \\ &\hat{\theta}(z) = \arg\min_{\theta} V(\theta, z) = \arg\min_{\theta} \tilde{V}(\theta, z_0, w(\theta, n_z, z_0)) && \text{(b)} \end{aligned} \qquad \text{(14-95)}$$

The sums w converge for $N \to \infty$ in some sense (mean square, in probability, or w.p. 1) to their expected values

$$\begin{aligned} &w(n_z, z_0) \to \mathcal{E}\{w(n_z, z_0)\} = \mu_w(z_0) \\ &w(\theta, n_z, z_0) \to \mathcal{E}\{w(\theta, n_z, z_0)\} = \mu_w(\theta, z_0) \text{ uniformly in } \theta \end{aligned} \qquad \text{(14-96)}$$

Note that the convergence of $w(\theta, n_z, z_0)$ must be uniform w.r.t. θ, otherwise $\mu_w(\theta, z_0)$ is not a continuous function of θ. The strong or weak convergence then follows directly from the interchangeability of a continuous function and the almost sure limit or the limit in probability

$$\begin{aligned} &\hat{\theta}(z) \to \tilde{\theta}(z_0) = \tilde{f}(z_0, \mu_w(z_0)) && \text{(a)} \\ &\hat{\theta}(z) \to \tilde{\theta}(z_0) = \arg\min_{\theta} \tilde{V}(\theta, z_0, \mu_w(\theta, z_0)) && \text{(b)} \end{aligned} \qquad \text{(14-97)}$$

This is illustrated on the explicit case (14-97.a) using properties 1 and 8 of the almost sure limit (see Section 14.8),

$$\begin{aligned} \underset{N \to \infty}{\text{a.s.lim}}\, \hat{\theta}(z) &= \tilde{f}(\underset{N \to \infty}{\text{a.s.lim}}\, z_0, \underset{N \to \infty}{\text{a.s.lim}}\, w(n_z, z_0)) \\ &= \tilde{f}(\lim_{N \to \infty} z_0, \lim_{N \to \infty} \mu_w(z_0)) \\ &= \lim_{N \to \infty} \tilde{f}(z_0, \mu_w(z_0)) \\ &= \lim_{N \to \infty} \tilde{\theta}(z_0) \end{aligned} \qquad \text{(14-98)}$$

The estimate $\hat{\theta}(z)$ is strongly or weakly consistent if the limit value

$$\theta_* = \lim_{N \to \infty} \tilde{\theta}(z_0) \qquad \text{(14-99)}$$

equals the true value $\theta_* = \theta_0$. In most cases the stronger condition $\tilde{\theta}(z_0) = \theta_0$ is satisfied.

14.13.2 Convergence Rate

Suppose that we have estimated some parameters $\hat{\theta}$ using N data samples. We may wonder now how many additional samples we should measure in order to decrease the uncertainty on $\hat{\theta}$ by a factor of k. The answer is given by the convergence rate of the estimates. It, typically, obeys the so-called $\sqrt{N}$ law; for example, to decrease the uncertainty by a factor of 10 we need 100 times more data samples. The consistency property does not tell how quickly the estimates converge to the true value. An additional analysis is necessary to establish the convergence rate. It starts by analyzing how fast the variance of the stochastic sums w in (14-96) converges to zero. By properties 6 and 7 of Section 14.8, these convergence rates also apply for the limit in mean square and the limit in probability. For mixing sequences, this rate is at least $O_{\text{m.s.}}(N^{-1/2})$ (see Section 14.9, version 3 of the law of large numbers). The almost sure limit is not used in this context because it results in somewhat slower convergence rates (see Section 14.9). Hence, further analysis is done with the limit in probability as it is interchangeable with a continuous function. The main idea is to make a Taylor series expansion of $\hat{\theta}$ (14-95) as a function of the stochastic sums w. The implicit case (14-95.b) is somewhat involved because first an explicit expression of $\hat{\theta}$ as a function of w should be constructed. Therefore, the explicit case (14-95.a) is tackled first. To simplify the notations, in the sequel of the analysis we drop the dependence of w and μ_w on n_z and z_0.

14.13.2.1 Explicit Case. The Taylor series expansion of the kth entry of $\tilde{f}(z_0, w)$ (14-95.a) w.r.t. w at the point $w = \mu_w$ gives

$$\begin{aligned} \tilde{f}_{[k]}(z_0, w) = \tilde{f}_{[k]}(z_0, \mu_w) + \left.\frac{\partial \tilde{f}_{[k]}(z_0, w)}{\partial w}\right|_{w = \mu_w} (w - \mu_w) + \\ \frac{1}{2}(w - \mu_w)^T \left.\frac{\partial^2 \tilde{f}_{[k]}(z_0, w)}{\partial w^2}\right|_{w = \widehat{w}} (w - \mu_w) \end{aligned} \qquad (14\text{-}100)$$

where $\widehat{w}$ is a point on the straight line connecting w to μ_w ($\widehat{w} = tw + (1 - t)\mu_w$ with $t \in [0, 1]$). Suppose now that $\text{Cov}(w) = O(N^{-1})$ so that (property 7 of Section 14.8)

$$w = \mu_w + O_{\text{p}}(N^{-1/2}) \qquad (14\text{-}101)$$

Using the definitions $\hat{\theta}(z) = \tilde{f}(z_0, w)$, $\tilde{\theta}(z_0) = \tilde{f}(z_0, \mu_w)$ and applying result (14-101) to (14-100), taking into account that the matrix dimensions of w, $\tilde{f}$, and the derivatives of $\tilde{f}$ w.r.t. w, are independent of N, gives

$$\begin{aligned} &\hat{\theta}(z) = \tilde{\theta}(z_0) + \delta_\theta(z) + b_\theta(z) && \text{(a)} \\ &\delta_\theta(z) = \left.\frac{\partial \tilde{f}_{[k]}(z_0, w)}{\partial w}\right|_{w = \mu_w} (w - \mu_w) = O_{\text{p}}(N^{-1/2}) && \text{(b)} \\ &b_\theta(z) = O_{\text{p}}(N^{-1}) && \text{(c)} \end{aligned} \qquad (14\text{-}102)$$

From (14-102) it follows directly that the convergence rate of $\hat{\theta}(z)$ to the nonrandom value $\tilde{\theta}(z_0)$ is an $O_{\text{p}}(N^{-1/2})$.

14.13.2.2 Implicit Case. Now we give an approximate analysis for the implicit case (14-93) (see Chapter 15 for the complete analysis). The implicit function that defines $\hat{\theta}(z)$ as a function of w is

$$\tilde{V}'(\hat{\theta}, z_0, w(\hat{\theta})) = 0 \qquad (14\text{-}103)$$

where ' denotes the derivative w.r.t. θ. Taylor series expansion of (14-103) w.r.t. $\hat{\theta}$ at the point $\tilde{\theta}$ gives, neglecting the second and higher order terms,

$$\tilde{V}'^T(\hat{\theta}, z_0, w(\hat{\theta})) = \tilde{V}'^T(\tilde{\theta}, z_0, w(\tilde{\theta})) + \tilde{V}''(\tilde{\theta}, z_0, w(\tilde{\theta}))(\hat{\theta}(z) - \tilde{\theta}(z_0)) \qquad (14\text{-}104)$$

Because $\hat{\theta}(z)$ is the minimizing argument of $\tilde{V}(\hat{\theta}, z_0, w(\hat{\theta}))$ (14-104) reduces to

$$\hat{\theta}(z) - \tilde{\theta}(z_0) = -\tilde{V}''^{-1}(\tilde{\theta}, z_0, w(\tilde{\theta}))\tilde{V}'^{T}(\tilde{\theta}, z_0, w(\tilde{\theta})) \qquad (14\text{-}105)$$

Following the same lines as in the explicit case, the convergence rate of the first- and the second-order derivatives of the cost function in (14-105) is obtained. If (14-101) is valid, then

$$\begin{aligned} &\tilde{V}'(\tilde{\theta}, z_0, w(\tilde{\theta})) = \tilde{V}'(\tilde{\theta}, z_0, \mu_w(\tilde{\theta})) + O_{\mathrm{p}}(N^{-1/2}) = O_{\mathrm{p}}(N^{-1/2}) \quad &\text{(a)} \\ &\tilde{V}''(\tilde{\theta}, z_0, w(\tilde{\theta})) = \tilde{V}''(\tilde{\theta}, z_0, \mu_w(\tilde{\theta})) + O_{\mathrm{p}}(N^{-1/2}) \quad &\text{(b)} \end{aligned} \qquad (14\text{-}106)$$

The last equality in (14-106a) is due to the fact that $\tilde{\theta}(z_0)$ is the minimizing argument of $\tilde{V}(\theta, z_0, \mu_w(\theta))$. Using the interchangeability of a continuous function and the limit in probability, (14-106b) becomes

$$\tilde{V}''^{-1}(\tilde{\theta}, z_0, w(\tilde{\theta})) = \tilde{V}''^{-1}(\tilde{\theta}, z_0, \mu_w(\tilde{\theta})) + O_{\mathrm{p}}(N^{-1/2}) \qquad (14\text{-}107)$$

Collecting (14-105), (14-106), and (14-107) gives Eq. (14-102a) with

$$\begin{aligned} &\delta_{\theta}(z) = -\tilde{V}''^{-1}(\tilde{\theta}, z_0, \mu_w(\tilde{\theta}))\tilde{V}'^{T}(\tilde{\theta}, z_0, w(\tilde{\theta})) = O_{\mathrm{p}}(N^{-1/2}) \quad &\text{(a)} \\ &b_{\theta}(z) = O_{\mathrm{p}}(N^{-1}) \quad &\text{(b)} \end{aligned} \qquad (14\text{-}108)$$

Similarly to the explicit case, it follows from (14-108a) that the convergence rate of $\hat{\theta}(z)$ to $\tilde{\theta}(z_0)$ is an $O_{\mathrm{p}}(N^{-1/2})$.

14.13.3 Asymptotic Bias

The asymptotic bias analysis is done for (weakly or strongly) consistent estimators. No explicit expression for the bias can be found except for some special examples. The best we can hope is to find how the bias behaves as $N \to \infty$. It is derived from the convergence rate analysis (14-102) and (14-108) by making the additional assumption that the disturbing noise is uniformly bounded and using the property that for such noise the expected value and the limit in probability may be interchanged (property 4 of Section 14.8). Taking the expected value of (14-102a) gives

$$\mathcal{E}\{\hat{\theta}(z)\} = \tilde{\theta}(z_0) + \mathcal{E}\{\delta_{\theta}(z)\} + O(N^{-1}) \qquad (14\text{-}109)$$

For the explicit case (14-102b) it is obvious that $\mathcal{E}\{\delta_{\theta}(z)\} = 0$, whereas this is true for the implicit case (14-108) only if

$$\mathcal{E}\{\tilde{V}'(\tilde{\theta}, z_0, w(\tilde{\theta}))\} = \frac{\partial\,\mathcal{E}\{\tilde{V}(\tilde{\theta}, z_0, w(\tilde{\theta}))\}}{\partial\tilde{\theta}} = 0 \qquad (14\text{-}110)$$

Condition (14-110) is often satisfied so that for both cases (14-109) can be written as

$$\mathcal{E}\{\hat{\theta}(z)\} = \tilde{\theta}(z_0) + O(N^{-1}) \tag{14-111}$$

For (weakly or strongly) consistent estimators we almost always have $\tilde{\theta}(z_0) = \theta_0$ so that the bias on $\hat{\theta}(z)$ behaves as an $O(N^{-1})$. If $\tilde{\theta}(z_0) \neq \theta_0$, then, the convergence rate of $\tilde{\theta}(z_0)$ to θ_0 should be added to the $O(N^{-1})$ bias term in (14-109).

14.13.4 Asymptotic Normality

From (14-102) and (14-108) it follows that $\sqrt{N}(\hat{\theta}(z) - \tilde{\theta}(z_0))$ converges in probability, and, hence, also in distribution (interrelation 3, Section 14.7), to $\sqrt{N}\delta_\theta(z)$. Hence, the study of the asymptotic distribution function of $\hat{\theta}(z)$ boils down to the study of the asymptotic distribution of, respectively, $w - \mu_w$, see (14-102.b), and $\tilde{V}'(\tilde{\theta}, z_0, w(\tilde{\theta}))$, see (14-108.a). Thereto we use the central limit theorems (Section 14.10) or the Fréchet-Shohat lemma (Lemma 14.11). If $w - \mu_w$ and $\tilde{V}'(\tilde{\theta}, z_0, w(\tilde{\theta}))$ are asymptotically normally distributed, then $\sqrt{N}\hat{\theta}(z)$ is also asymptotically normally distributed (a linear combination of Gaussian random variables is Gaussian) with mean $\sqrt{N}\tilde{\theta}(z_0)$ and covariance matrix $NE\{\delta_\theta(z)\delta_\theta^T(z)\}$. Note that the analysis assumes only that the moments of $\delta_\theta(z)$ exist, not those of $\hat{\theta}(z)$.

14.13.5 Asymptotic Efficiency

The asymptotic efficiency analysis is done for (weakly or strongly) consistent estimators. It consists of comparing the covariance matrix of the limit random variable $\delta_\theta(z)$ to the Cramér-Rao lower bound for unbiased estimators (14-87). The consistent estimator is asymptotically efficient if

$$\lim_{N \to \infty} N(\mathcal{E}\{\delta_\theta(z)\delta_\theta^T(z)\} - Fi^{-1}(\theta_0)) = 0 \tag{14-112}$$

Note that (14-112) can be true while the moments of $\hat{\theta}(z)$ may not exist.

14.14 PITFALLS

Some erroneous statements such as "strong consistency implies asymptotic unbiasedness" or "the limit in mean square and a continuous function are interchangeable" are tempting to make. Therefore, a list of pitfalls is given, some of which are illustrated by means of counterexamples, namely:

1. Weak and strong consistency do NOT imply asymptotic unbiasedness.
2. Asymptotic unbiasedness does NOT imply any kind of consistency.
3. Weak and strong consistency do NOT imply that the limit of the variance is equal to the variance of the limit.
4. 1 and 3 are special cases of: the limit in probability and the almost sure limit are NOT interchangeable with the expected value. Similar $\mathcal{E}\{O_p(N^{-k})\} \neq O(N^{-k})$ and $\mathcal{E}\{O_{a.s.}(N^{-k})\} \neq O(N^{-k})$.
5. The limit in probability and the limit with probability one ($N \to \infty$) may NOT be interchanged with a continuous matrix function if its matrix dimensions vary with N. For example, let $J \in \mathbb{R}^{N \times p}$ then,

$$\operatorname*{plim}_{N\to\infty} J^T J \neq (\operatorname*{plim}_{N\to\infty} J)^T (\operatorname*{plim}_{N\to\infty} J)$$

Similarly, if $A, B \in \mathbb{R}^{N\times p}$ with $A = O_{\text{a.s.}}(N^{-s})$ and $B = O_{\text{a.s.}}(N^{-r})$ then $A^T B \neq O_{\text{a.s.}}(N^{-(r+s)})$.

6. The supremum (maximum) and the expected value may NOT be interchanged.
7. The limit in mean square and a continuous function are NOT interchangeable.

$$\operatorname*{l.i.m.}_{N\to\infty} f(x(N)) \neq f(\operatorname*{l.i.m.}_{N\to\infty} x(N))$$

Example 14.21: Weak consistency does not imply asymptotic unbiasedness. Let $\hat{\theta}(N)$ be an estimator of $\theta_0 = 1$ that takes the value 1 with probability $1 - 1/N$ and the value N with probability $1/N$. The estimator is weakly consistent,

$$\lim_{N\to\infty} \text{Prob}(|\hat{\theta}(N) - \theta_0| < \delta) = \lim_{N\to\infty} \text{Prob}(|\hat{\theta}(N) - \theta_0| = 0) = \lim_{N\to\infty} (1 - 1/N) = 1$$

and asymptotically biased $\lim_{N\to\infty} \mathcal{E}\{\hat{\theta}(N)\} = \lim_{N\to\infty} (1(1 - 1/N) + N(1/N)) = 2.$ □

Example 14.22: Asymptotic unbiasedness does not imply consistency (14-78). Consider the squared magnitude of the DFT transform of a noise sequence $v(t)$

$$V(k) = \left| \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} v(t) e^{-2\pi j \frac{kt}{N}} \right|^2 \qquad (14\text{-}113)$$

In Kay (1988) it is shown that $V(k)$ is an asymptotically unbiased estimate of the power spectral density $V_0(k)$,

$$\lim_{N\to\infty} \mathcal{E}\{V(k)\} = V_0(k) \qquad (14\text{-}114)$$

and that the variance of $V(k)$ does not decrease to zero as $N \to \infty$

$$\lim_{N\to\infty} \text{var}(V(k)) \approx V_0^2(k) \qquad k \neq 0, k \neq N/2 \qquad (14\text{-}115)$$

(see Appendix 4B of Kay, 1988). Hence, $V(k)$ is an inconsistent estimate. □

14.15 PRELIMINARY EXAMPLE—CONTINUED

We retake the resistance measurement problem of Sections 1.2 and 14.5 and assume that one experiment consisting of N current and voltage measurements is made,

$$i(k) = i_0 + n_i(k) \text{ and } u(k) = u_0 + n_u(k) \qquad (14\text{-}116)$$

$k = 1, 2, \ldots, N$. Unless mentioned otherwise, we assume that the current and voltage errors, $n_i(k)$ and $n_u(k)$, are mutually independent, zero mean iid random variables, $n_u(k) \in U(0, \sigma_u^2)$ and $n_i(k) \in U(0, \sigma_i^2)$. The goal of this section is to predict, theoretically, the behavior of the three resistance estimators, the simple approach (1-1), the least squares method (1-2), and the errors-in-variables approach (1-3). The analysis follows the lines of Section 14.13. In a first step, we rewrite the estimates as a function of the stochastic sums w. We obtain

$$\begin{aligned} \hat{R}_{\mathrm{SA}}(N) &= \frac{1}{N}\sum\nolimits_{k=1}^{N}\frac{u(k)}{i(k)} = w_{[1]} \\ \mathcal{E}\{w_{[1]}\} &= \frac{1}{N}\sum\nolimits_{k=1}^{N}\mathcal{E}\{u(k)\}\mathcal{E}\{1/i(k)\} \\ &= \frac{u_0}{2\sqrt{3}\sigma_i}\int_{-\sqrt{3}\sigma_i}^{\sqrt{3}\sigma_i}(i_0+z)^{-1}dz \\ &= R_0\frac{i_0}{2\sqrt{3}\sigma_i}\ln\left(\frac{1+\sqrt{3}\sigma_i/i_0}{1-\sqrt{3}\sigma_i/i_0}\right) \qquad (\sqrt{3}\sigma_i < i_0) \end{aligned} \tag{14-117}$$

for the simple approach,

$$\begin{aligned} \hat{R}_{\mathrm{LS}}(N) &= \frac{\frac{1}{N}\sum\nolimits_{k=1}^{N}u(k)i(k)}{\frac{1}{N}\sum\nolimits_{k=1}^{N}i^2(k)} = \frac{u_0 i_0 + w_{[1]}}{i_0^2 + w_{[2]}} \\ w_{[1]} &= \frac{1}{N}\sum\nolimits_{k=1}^{N}(u_0 n_i(k) + i_0 n_u(k) + n_i(k)n_u(k)) \text{ with } \mathcal{E}\{w_{[1]}\} = 0 \\ w_{[2]} &= \frac{1}{N}\sum\nolimits_{k=1}^{N}(2i_0 n_i(k) + n_i^2(k)) \text{ with } \mathcal{E}\{w_{[2]}\} = \sigma_i^2 \end{aligned} \tag{14-118}$$

for the least squares method, and

$$\begin{aligned} \hat{R}_{\mathrm{EV}}(N) &= \frac{\frac{1}{N}\sum\nolimits_{k=1}^{N}u(k)}{\frac{1}{N}\sum\nolimits_{k=1}^{N}i(k)} = \frac{u_0 + w_{[1]}}{i_0 + w_{[2]}} \\ w_{[1]} &= \frac{1}{N}\sum\nolimits_{k=1}^{N}n_u(k) \text{ with } \mathcal{E}\{w_{[1]}\} = 0 \\ w_{[2]} &= \frac{1}{N}\sum\nolimits_{k=1}^{N}n_i(k) \text{ with } \mathcal{E}\{w_{[2]}\} = 0 \end{aligned} \tag{14-119}$$

for the errors-in-variables approach. Note that under the condition $\sqrt{3}\sigma_i < i_0$, all the moments of the three resistance estimators exist for any N.

14.15.1 Consistency

Because by assumption $n_u(k)$ and $n_i(k)$ are mutually independent, iid random variables, each entry of w in (14-117) to (14-119) consists of the sum of iid random variables and converges to its expected value (see Section 14.9, version 1 of the law of large numbers). Hence, we find

$$\begin{aligned}
\tilde{R}_{\text{SA}}(N) &= \mathcal{E}\{w_{[1]}\} = R_0\frac{i_0}{2\sqrt{3}\sigma_i}\ln\Big(\frac{1+\sqrt{3}\sigma_i/i_0}{1-\sqrt{3}\sigma_i/i_0}\Big) && \text{(a)} \\
\tilde{R}_{\text{LS}}(N) &= \frac{u_0 i_0 + \mathcal{E}\{w_{[1]}\}}{i_0^2 + \mathcal{E}\{w_{[2]}\}} = \frac{R_0}{1+\sigma_i^2/i_0^2} && \text{(b)} \qquad \text{(14-120)} \\
\tilde{R}_{\text{EV}}(N) &= \frac{u_0 + \mathcal{E}\{w_{[1]}\}}{i_0 + \mathcal{E}\{w_{[2]}\}} = R_0 && \text{(c)}
\end{aligned}$$

where $\sqrt{3}\sigma_i < i_0$ for $\tilde{R}_{\text{SA}}(N)$. The values $\tilde{R}(N)$ in (14-120) are independent of the number of samples N so that $R_* = \tilde{R}(N)$ for each estimator. We conclude that the simple approach and the least squares estimates are inconsistent, $R_{\text{SA}*} \neq R_0$, $R_{\text{LS}*} \neq R_0$, while the errors-in-variables estimate is strongly consistent, $R_{\text{EV}*} = R_0$. Note that $\tilde{R}_{\text{SA}}(N)$ and $\tilde{R}_{\text{LS}}(N)$ tend to R_0 as $i_0/\sigma_i \to 0$ (Exercise 14.25). Taking the same numerical example as in Section 14.5 ($i_0 = 1$ A, $\sigma_i = 1/\sqrt{12}$ A) gives $R_{\text{SA}*} = R_0\ln 3 = 1.099R_0$ and $R_{\text{LS}*} = R_0 12/13 = 0.923R_0$.

14.15.2 Convergence Rate

According to Section 14.13.2, the convergence rate of $\hat{R}(N)$ to $\tilde{R}(N)$ equals the convergence rate of w to μ_w. Because the variance of each entry of w exists and is finite ($\sqrt{3}\sigma_i < i_0$ for (14-117)), the convergence rate of w to μ_w equals $O_{\text{a.s.}}(N^{-1/2}\ln(\ln(N)))$ or $O_{\text{p}}(N^{-1/2})$ (see Section 14.9, respectively versions 1 and 3 of the law of large numbers). Hence, for each of the three estimators, we have

$$\begin{aligned}
\hat{R}(N) &= \tilde{R}(N) + \delta(N) + b(N) && \text{(a)} \\
\delta(N) &= \partial\hat{R}(N)/\partial w\big|_{w=\mu_w} w - \mu_w = O_{\text{p}}(N^{-1/2}) && \text{(b)} \qquad \text{(14-121)} \\
b(N) &= O_{\text{p}}(N^{-1}) && \text{(c)}
\end{aligned}$$

with $b_{\text{SA}}(N) = 0$ and

$$\begin{aligned}
\partial\hat{R}_{\text{SA}}(N)/\partial w\big|_{w=\mu_w} &= 1 \\
\partial\hat{R}_{\text{LS}}(N)/\partial w\big|_{w=\mu_w} &= \begin{bmatrix}(i_0^2+\sigma_i^2)^{-1} & -u_0 i_0(i_0^2+\sigma_i^2)^{-2}\end{bmatrix} \qquad \text{(14-122)} \\
\partial\hat{R}_{\text{EV}}(N)/\partial w\big|_{w=\mu_w} &= \begin{bmatrix}1/i_0 & -u_0 i_0^{-2}\end{bmatrix}
\end{aligned}$$

14.15.3 Asymptotic Normality

The vector w consists of the sum of iid random variables with finite mean value and finite nonzero (co)variance matrix ($\sqrt{3}\sigma_i < i_0$ for (14-117)). According to the multivariable version of the central limit theorem (see Section 14.10, version 1), $w - \mu_w$ is asymptotically normally distributed at the rate $O(N^{-1/2})$. Hence, the estimates $\hat{R}(N)$ are asymptotically normally distributed (at the rate $O(N^{-1/2})$) with mean value $\tilde{R}(N)$ and variance $\mathcal{E}\{\delta^T\delta\}$ (see Section 14.13.4). We find

$$\begin{aligned}
&\mathrm{var}(\delta_{\mathrm{SA}}(N)) = \mathrm{var}(w_{[1]}) && \text{(a)}\\
&\mathrm{var}(\delta_{\mathrm{LS}}(N)) = \frac{\mathrm{var}(w_{[1]})}{(i_0^2+\sigma_i^2)^2} + \frac{\mathrm{var}(w_{[2]})}{(i_0^2+\sigma_i^2)^4}u_0^2 i_0^2 - 2\frac{\mathrm{covar}(w_{[1]}, w_{[2]})}{(i_0^2+\sigma_i^2)^3}u_0 i_0 && \text{(b)}\\
&\mathrm{var}(\delta_{\mathrm{EV}}(N)) = \frac{\mathrm{var}(w_{[1]})}{i_0^2} + \frac{u_0^2}{i_0^4}\mathrm{var}(w_{[2]}) && \text{(c)}
\end{aligned} \qquad (14\text{-}123)$$

where the stochastic vectors w in (a), (b), and (c) are defined in, respectively, (14-117), (14-118), and (14-119). For the numerical example of Section 14.5 ($u_0 = 1$ V, $i_0 = 1$ A, $\sigma_u = 1/\sqrt{12}$ V, and $\sigma_i = 1/\sqrt{12}$ A), we obtain $\mathrm{var}(\delta_{\mathrm{SA}}(N)) = 0.237N^{-1}$, $\mathrm{var}(\delta_{\mathrm{LS}}(N)) = 0.132N^{-1}$, and $\mathrm{var}(\delta_{\mathrm{EV}}(N)) = 0.167N^{-1}$. Note that $\mathrm{var}(\delta_{\mathrm{LS}}(N)) < \mathrm{var}(\delta_{\mathrm{EV}}(N))$, as observed in Section 1.2.2.3. Formulas (14-123b) and (14-123c) make it possible to predict the sample variances, obtained by Monte Carlo simulation, of the least squares and errors-in-variables estimates shown in Figure 1-7 on page 11 (Exercise 14.27).

14.15.4 Asymptotic Efficiency

The Cramér-Rao lower bound does not exist for uniformly distributed random variables because the uniform probability density function does not satisfy the regularity conditions of Theorem 14.18 (the derivatives of the pdf do not exist at the boundaries of the domain). Therefore, the asymptotic variance (14-123c) of the consistent estimator $\hat{R}_{\mathrm{EV}}(N)$ is compared with the UCRB for Gaussian distributed errors (note that in opposition to the uniform case, the moments of $\hat{R}_{\mathrm{EV}}(N)$ do not exist for Gaussian distributed errors).

Putting $\mathrm{var}(w_{[1]}) = \sigma_u^2/N$ and $\mathrm{var}(w_{[2]}) = \sigma_i^2/N$ in (14-123c) gives an explicit expression for the variance of the limiting random variable $\delta_{\mathrm{EV}}(N)$

$$\mathrm{var}(\delta_{\mathrm{EV}}(N)) = \frac{R_0^2}{N}\left(\frac{\sigma_i^2}{i_0^2} + \frac{\sigma_u^2}{u_0^2}\right) \qquad (14\text{-}124)$$

To construct the UCRB we need the likelihood function of the measurements $z = [u(1)u(2)\dots u(N)i(1)i(2)\dots i(N)]^T$. As $u(k)$ and $i(k)$ are mutually independent, iid Gaussian random variables, it is given by ($u_0 = R_0 i_0$)

$$\begin{aligned}
f_z(z, i_0, R_0) &= \prod\nolimits_{k=1}^{N} f_{u(k)}(u(k)) f_{i(k)}(i(k))\\
&= \frac{1}{(2\pi\sigma_u\sigma_i)^N}\exp\left(-\frac{1}{2}\sum_{k=1}^{N}\frac{(u(k)-R_0 i_0)^2}{\sigma_u^2} + \frac{(i(k)-i_0)^2}{\sigma_i^2}\right)
\end{aligned} \qquad (14\text{-}125)$$

Two unknowns appear in (14-125), the true values R_0 and i_0 of, respectively, the resistance and the current. This means that, in maximum likelihood sense, the resistance as well as the current must be estimated. Applying (14-85) gives the UCRB on the current and resistance estimates

$$Fi^{-1}(i_0, R_0) = \frac{1}{N}\begin{bmatrix} \sigma_i^2 & -u_0\dfrac{\sigma_i^2}{i_0^2} \\ -u_0\dfrac{\sigma_i^2}{i_0^2} & R_0^2\left(\dfrac{\sigma_i^2}{i_0^2} + \dfrac{\sigma_u^2}{u_0^2}\right) \end{bmatrix} \qquad (14\text{-}126)$$

Finally, entry [2, 2] of $Fi^{-1}(i_0, R_0)$ is the UCRB on the resistance estimate

$$Fi^{-1}(R_0) = \frac{R_0^2}{N}\left(\frac{\sigma_i^2}{i_0^2} + \frac{\sigma_u^2}{u_0^2}\right) \tag{14-127}$$

Comparing (14-127) and (14-124) shows that the errors-in-variable estimate $\hat{R}_{\mathrm{EV}}(N)$ is asymptotically efficient.

14.15.5 Asymptotic Bias

Under the condition $\sqrt{3}\sigma_i < i_0$, the expected values of the three resistance estimators exist for any N. Applying property 4 of the limit in probability (see Section 14.8), it follows from (14-120) and (14-121) that the bias of the simple approach and the least squares estimates is an $O(N^0)$, while that of the errors-in-variables approach is an $O(N^{-1})$. For the numerical example of Section 14.5 we find $b_{\mathrm{SA}} = 0.10\ \Omega$ and $b_{\mathrm{LS}} = -0.08\ \Omega$.

14.15.6 Robustness

The simple approach (14-117) is not robust w.r.t. the underlying distribution function of the errors. Indeed, for Gaussian current measurement errors $n_i(k)$, neither the expected value nor the variance of $1/i(k)$ exists, so that estimate $\hat{R}_{\mathrm{SA}}(N)$ does not converge w.p. 1, nor in probability, nor in mean square sense (see Section 14.9, versions 1 and 3 of the law of large numbers).

The properties of $\hat{R}_{\mathrm{EV}}(N)$ have been analyzed, assuming that the errors $n_u(k)$ and $n_i(k)$ are mutually independent, iid uniform random variables. One may wonder now what happens with these properties when, for example, the measurement errors are no longer independent and/or are no longer identically distributed. Therefore, the analysis is redone, assuming that the errors $n_u(k)$ and $n_i(k)$ are mixing of order 2. For example, filtered independent noise with uniformly bounded variances satisfies this assumption. The errors may be correlated and their distribution function has not been specified. Stationarity is also no longer required. The following properties remain unchanged and are, hence, robust w.r.t. the independence and stationarity assumption.

1. Consistency: applying the strong law of large numbers for mixing sequences (see Section 14.9, version 3) shows that $w_{[1]}$ and $w_{[2]}$ in (14-119) still converge w.p. 1 to zero, which proves the consistency.
2. Convergence rate: applying the convergence rate of the law of large numbers for mixing sequences (see Section 14.9, version 3) to $w_{[1]}$ and $w_{[2]}$ in (14-119) shows that the convergence rate (14-121b) still applies.
3. Asymptotic normality: if the assumption is tightened to a mixing condition of order infinity, then the central limit theorem for mixing sequences (see Section 14.10, version 4) applied to $w_{[1]}$ and $w_{[2]}$ in (14-119) shows that $\hat{R}_{\mathrm{EV}}(N)$ is still asymptotically normally distributed.
4. Asymptotic bias: if the class of allowable disturbances is restricted to uniformly bounded random variables, then the bias is still an $O(N^{-1})$.

The following property is not robust:

1. Asymptotic efficiency: the estimator $\hat{R}_{\text{EV}}(N)$ does not take into account the dependence between the measurement errors and the particular shape of their distribution function so that, in general, $\text{var}(\hat{\delta}_{\text{EV}}(N))$ will not reach the UCRB.

14.16 PROPERTIES OF THE NOISE AFTER A DISCRETE FOURIER TRANSFORM

In this section we discuss the properties of filtered white noise after a discrete Fourier transform (DFT). We first handle the scalar case and afterward generalize the results to the multivariable case. Taking the discrete Fourier transform of $v(t) = H(q)e(t)$ gives (see Section 5.7.3)

$$V(k) = H(z_k^{-1})E(k) + T_H(z_k^{-1}) \tag{14-128}$$

with $H(z^{-1}) = C(z^{-1})/D(z^{-1})$ the noise model, $E(k)$ and $V(k)$ the discrete Fourier transforms of, respectively, $e(t)$ and $v(t)$, and $T_H(z^{-1}) = J(z^{-1})/D(z^{-1})$ the initial and final conditions of the noise process. The transient term $T_H(z_k^{-1})$ is strongly correlated over the frequency and gives a nonmixing contribution to the noise $V(k)$. Fortunately, it can be shown that its influence decreases to zero with probability one.

Lemma 14.23: Consider filtered white noise $H(q)e(t)$, where $H(z^{-1})$ is stable and $e(t)$ has uniformly bounded absolute moments of order $2+\delta$, with $\delta > 0$: $\mathcal{E}\{|e(t)|^{2+\delta}\} \leq c < \infty$, with c independent of t. The discrete Fourier transform of $H(q)e(t)$ converges w.p. 1 to $H(z_k^{-1})E(k)$. The convergence rate in probability is an $O_p(N^{-1/2})$.

Proof. See Appendix 14.K. □

Lemma 14.24: Let $e(t)$ be independent, identically distributed (iid) noise with existing moments of any order. The discrete Fourier transform $E(k)$ is asymptotically ($N \to \infty$) independent, circular complex normally distributed (convergence in law at the rate $O(N^{-1/2})$). $E(k)$ has zero mean except at $k = 0$ (DC).

Proof. See Appendix 14.L. □

Theorem 14.25 (Asymptotic Normality): The discrete Fourier transform $V(k)$ (14-128) of filtered iid noise $v(t) = H(q)e(t)$, where $H(z^{-1})$ is stable and $e(t)$ has existing moments of any order, is asymptotically ($N \to \infty$) independent, circular complex normally distributed (convergence in law at the rate $O(N^{-1/2})$). For any N, $V(k)$ has zero mean except at $k = 0$ (DC).

Proof. According to Lemma 14.24, $E(k)$ is asymptotically independent, circular complex normally distributed. This is also true for $H(z_k^{-1})E(k)$ because $|H(z_k^{-1})|$ is uniformly bounded (see Exercise 14.6). Applying Lemma 14.23 and interrelations 1 and 3 of the stochastic limits (see Section 14.7) proves the theorem. □

Due to the more restrictive noise assumption, this result is stronger than Theorem 4.4.1 of Brillinger (1981), which shows for mixing stationary time domain noise $v(t)$ that the DFT spectral lines $V(\zeta_1 N)$, $V(\zeta_2 N)$, ..., $V(\zeta_J N)$, at a set of FIXED frequencies $f_r = \zeta_r f_s$,

$r = 1, 2, \ldots, J$, are asymptotically independent, circular complex normally distributed. As the number of time domain samples N increases, the number of DFT lines in between two consecutive spectral lines $V(\zeta_r N)$, $V(\zeta_{r+1} N)$, for which Theorem 4.4.1 of Brillinger (1981) applies, increases to infinity. This is not the case in Theorem 14.25.

Lemma 14.26 (Mixing of Order P): Let $e(t)$ be independent, identically distributed noise with finite moments of order $2P$ and discrete Fourier transform $E(k)$. The squared amplitude spectrum $|E(k)|^2$, DC not included, is mixing of order P.

Proof. See Appendix 14.M. □

Lemma 14.26 does not imply that $E(k)$ is mixing of order $2P$. On the contrary, it can only be proved that $E(k)$ is mixing of order 2 (Lemma 14.27).

Lemma 14.27: Let $e(t)$ be independently distributed noise with mean $\mu < \infty$, variance $\sigma^2 < \infty$, and uniformly bounded fourth-order moments. The discrete Fourier transform $E(k)$ of $e(t)$ and its squared amplitude spectrum $|E(k)|^2$, DC not included, are mixing of order 2.

Proof. See Appendix 14.N. □

Note that Lemma 14.27 requires only the stationarity of the first- and second-order moments of $e(t)$.

Theorem 14.28 (Strong Law of Large Numbers): Let $V(k)$ (14-128) be the DFT of filtered noise $v(t) = H(q)e(t)$, where $H(z^{-1})$ is stable, and where $e(t)$ is independently distributed noise with mean $\mu < \infty$, variance $\sigma^2 < \infty$, and uniformly bounded fourth-order moments: $\mu_4(t) \leq c < \infty$ with c independent of t. Consider the partial sums

$$S(F) = \sum_{k \in \mathbb{F}} W_k V(k) \text{ and } S(F) = \sum_{k \in \mathbb{F}} |W_k V(k)|^2 \tag{14-129}$$

with W_k a uniformly bounded deterministic weighting, $\mathbb{F}$ a subset of the DFT frequencies $k = 0, 1, \ldots, N/2$, and $F = O(N)$ the number of frequencies in the set $\mathbb{F}$. If DC ($k = 0$) belongs to the set $\mathbb{F}$, then $\mu = \mathcal{E}\{e(t)\}$ must be zero. The partial sums $S(F)$ in (14-129) satisfy the strong law of large numbers (14-69). The convergence rate of the partial sums $S(F)/F$ is an $O_p(F^{-1/2})$.

Proof. See Appendix 14.O. □

To study the asymptotic distribution of the estimates, we also need the following central limit theorem.

Theorem 14.29 (Central Limit Theorem): Let $V(k)$ (14-128) be the discrete Fourier transform of filtered iid noise $v(t) = H(q)e(t)$, where $H(z^{-1})$ is stable and $e(t)$ has existing moments of any order. Let $X(k)$ be the discrete Fourier transform of the deterministic signal $x(t)$. Define the sum

$$S(N) = \textstyle\sum_{k=0}^{N-1} X(k)\bar{V}(k) \tag{14-130}$$

If $x(t)$ has constant power,

$$\frac{1}{N}\sum_{k=0}^{N-1}|X(k)|^2 = \frac{1}{N}\sum_{t=0}^{N-1}|x(t)|^2 = O(N^0) \tag{14-131}$$

and has uniformly bounded peak value

$$\max_t |x(t)| \leq c < \infty \tag{14-132}$$

for any N, ∞ included, with c a constant independent of N, then $N^{-1/2}S(N)$ is asymptotically normally distributed (convergence in law at the rate $O(N^{-1/2})$).

Proof. See Appendix 14.P. □

Theorem 14.29 is not valid if, for example, $X(k) = 1$, $k = 0, 1, \ldots, N-1$, because the corresponding time signal $x(t)$ is a pulse whose peak value increases as $O(\sqrt{N})$ (see also Exercise 14.28). This result can easily be understood by rewriting (14-130) as a circular DFT convolution

$$N^{-1/2}S(N) = N^{-1/2}\sum_{k=0}^{N-1}X(k)V(N-k) = \text{DFT}(x(t)v(t)) \tag{14-133}$$

It shows that only a very few (independent of N) samples of $v(t)$ contribute to the statistics of $S(N)$ if $x(t)$ is a pulse-like signal, while the central limit theorem requires a large number (increasing with N) of samples.

Theorem 14.29 requires that the energy of the signal $x(t)$ is more or less equally distributed over all time samples and not concentrated in a few points. This is the case for the following classes of signals:

1. Periodic signals with a fixed number F of frequencies. The DFT spectrum $X(k)$ increases as $N^{1/2}$ at the F excited frequencies (power per frequency is an $O(N^0)$) and is zero (if an integer number of periods is observed) or decreases as $N^{-1/2}$ at the other frequencies. The peak value is independent of N.
2. Peak value optimized periodic signals with flat amplitude spectrum that excite all DFT lines $k = 0, 1, \ldots, N-1$ with $X(k) = \overline{X}(N-k)$ and $|X(k)| = 1$ (power per frequency is an $O(N^{-1})$). The phases of $X(k)$ can always be chosen such that the peak value of $x(t)$ is an $O(N^0)$ close to 1 (Kahane, 1980).
3. Peak value optimized periodic signals with flat amplitude spectrum that excite only DFT lines $r(k+1)$, $k = 0, 1, \ldots, N/r-2$ with $r \in \mathbb{N}_0$, $X(k) = \overline{X}(N-k)$ and $|X(k)| = 1$ (power per excited frequency is an $O(N^{-1})$). The signal $x(t)$ is an r-times periodic repetition of the signal based on N/r samples that excites all DFT lines (see signal class 2). The peak value is an $O(N^0)$ close to 1 because this is also the case for signal class 2.
4. Peak value optimized periodic signals with flat amplitude spectrum that excite K (independent of N) frequency bands. For example, $K = 2$ frequency bands $[N/16, N/8]f_0$ and $[N/5, N/3]f_0$ with $f_0 = f_s/N$ the DFT resolution. The signal $x(t)$ can be written as a linear combination of K modulated versions of class 2 and/or 3 signals, each with an $O(N^0)$ peak value (Schoukens et al., 1996b). Hence, the peak value of $x(t)$ is an $O(N^0)$ because the peak value of a linear combination of K signals with uniformly bounded peak value is uniformly bounded.

Only upper bounds on the peak value are available for periodic signals with nonflat amplitude spectra. The same is true for periodic signals that excite only logarithmically spaced DFT lines (lacunar multisines), for example, 2^k $(k = 0, 1, \ldots, \ln(N/2) - 1,$ $X(k) = \bar{X}(N-k)$ and $|X(k)|^2 = O(N/\ln N))$. Fortunately, the upper bound increases slowly with N: for signals satisfying (14-131) the phases of $X(k)$ can always be chosen such that the peak value is bounded by $O(\sqrt{\ln N})$ (Theorem 4, Chapter 6 of Kahane, 1985), while for arbitrary phases the peak value is w.p. 1 an $O(\sqrt{\ln N})$

$$\operatorname{Prob}(\max_t |x(t)| \geq O(\sqrt{\ln N})) \leq N^{-2} \tag{14-134}$$

(Theorem 1 and Exercise 5, Chapter 6 of Kahane, 1985). This means that the risk of selecting phases with a peak value increasing faster than $O(\sqrt{\ln N})$ is very small. The following corollary can then be used.

Corollary 14.30: Let $V(k)$ (14-128) satisfy the assumptions of Lemma 14.31. If $x(t)$ satisfies (14-131) and has a peak value $\max_t |x(t)| = O(\sqrt{\ln N})$, then $N^{-1/2}S(N)$ (14-130) is asymptotically normally distributed (convergence in law at the rate $O\sqrt{\ln N/N}$).

Proof. See Appendix 14.Q. □

The preceding theorems are useful for studying the asymptotic behavior of frequency domain estimators when using deterministic excitation signals. For signals with a stochastic behavior, such as filtered white noise, periodic noise, and random multisines, we need the following theorems.

Lemma 14.31 (Mixing of Order P): Let $H_1(z^{-1})$, $H_2(z^{-1})$ be stable filters and $E_1(k)$, $E_2(k)$ the DFT spectra of iid random variables $e_1(t)$, $e_2(t)$ with existing moments of order P. Let $X(k)$ be one of the following spectra, $H_2(z_k^{-1})E_2(k)$, or the DFT spectrum of an integer number of periods of a normalized random multisine (see Definition 3.2) or normalized periodic noise (see Definition 3.3), all with existing Pth order moments. If $X(k)$ is independent of $E_1(k)$, then $X(k)H_1(z_k^{-1})E_1(k)$ is mixing of order P.

Proof. See Appendix 14.R. □

Theorem 14.32 (Strong Law of Large Numbers): Let $V(k)$ (14-128) be the DFT of filtered noise $v(t) = H(q)e(t)$, where $H(z^{-1})$ is stable, and where $e(t)$ is independently distributed noise with uniformly bounded absolute moments of order $2 + \delta$, with $\delta > 0$. Let $X(k)$ be the DFT spectrum of one of the following signals, filtered white noise, or an integer number of periods of a normalized random multisine (see Definition 3.2) or normalized periodic noise (see Definition 3.3), all with uniformly bounded fourth-order moments. Assume, furthermore, that $X(k)$ is independent of $V(k)$. Consider the partial sums

$$S(F) = \sum_{k \in \mathbf{F}} W_k X(k) \bar{V}(k) \text{ and } S(F) = \sum_{k \in \mathbb{F}} |W_k X(k)|^2 \tag{14-135}$$

with W_k a uniformly bounded deterministic weighting, $\mathbb{F}$ a subset of the DFT frequencies $k = 0, 1, \ldots, N/2$, and $F = O(N)$ the number of frequencies in the set $\mathbb{F}$. If DC $(k = 0)$ belongs to the set $\mathbb{F}$, then $\mu = \mathcal{E}\{e(t)\}$ must be zero. The partial sums $S(F)$ (14-135) satisfy the strong law of large numbers (14-69). The convergence rate of the partial sums $S(F)/F$ is an $O_p(F^{-1/2})$.

Proof. See Appendix 14.S. □

Theorem 14.33 (Central Limit Theorem): Let $V(k)$ (14-128) be the DFT of filtered iid noise $v(t) = H(q)e(t)$, where $H(z^{-1})$ is stable and $e(t)$ has existing moments of any order. Let $X(k)$ be the DFT spectrum of one of the following signals, filtered iid noise, an integer number of periods of a normalized random multisine (see Definition 3.2), or normalized periodic noise (see Definition 3.3), all with existing moments of any order. Assume, furthermore, that $X(k)$ is independent of $V(k)$. Consider the partial sums

$$S(F) = \sum_{k \in \mathbb{F}} W_k X(k)\bar{V}(k),\ S(F) = \sum_{k \in \mathbb{F}} |W_k V(k)|^2 \quad \text{and} \quad S(F) = \sum_{k \in \mathbb{F}} |W_k X(k)|^2 \quad (14\text{-}136)$$

with W_k a uniformly bounded deterministic weighting, $\mathbb{F}$ a subset of the DFT frequencies $k = 0, 1, \ldots, N/2$, and $F = O(N)$ the number of frequencies in the set $\mathbb{F}$. If DC ($k = 0$) belongs to the set $\mathbb{F}$, then $\mu = \mathcal{E}\{e(t)\}$ must be zero. The sums $F^{-1/2}S(F)$ (14-136) are asymptotically normally distributed (convergence in law at the rate $O(F^{-1/2})$).

Proof. See Appendix 14.T. □

The preceding theorems can easily be generalized to the multivariable case $v(t) = H(q)e(t)$ with $e(t) \in \mathbb{R}^p$, $v(t) \in \mathbb{R}^q$ and $H(q)$ a stable transfer function matrix. Formula (14-128) is still valid with $E(k) \in \mathbb{C}^p$, $V(k) \in \mathbb{C}^q$, $H(z^{-1}) = D^{-1}(z^{-1})C(z^{-1})$, $T_H(z^{-1}) = D^{-1}(z^{-1})J(z^{-1})$, $D(z^{-1})$ a q by q matrix, $C(z^{-1})$ a q by p matrix, and $J(z^{-1})$ a q by 1 vector (see Section 5.6).

14.17 EXERCISES

14.1. Let $x, y \in \mathbb{R}$ have finite second-order moments. Prove that $\text{var}(x + y) \leq (2\text{var}(x) + 2\text{var}(y))$ (hint: use $(a + b)^2 \leq (a + b)^2 + (a - b)^2 = 2a^2 + 2b^2$).

14.2. Show that conditions (14-13) are equivalent to condition (14-12).

14.3. Let $x \in \mathbb{C}^n$ be circular complex distributed and $A \in \mathbb{C}^{m \times n}$. Show that $y = Ax$ is circular complex distributed.

14.4. Show that $\text{Cov}(x_{\text{re}}) = 0.5(\text{Cov}(x))_{\text{Re}}$ for circular complex noise x.

14.5. Show that the probability density function of $x \in N_n^c(\mu_x, C_x)$ is given by (14-14) (hint: use $x_{\text{re}} \in N_{2n}((\mu_x)_{\text{re}}, 0.5(C_x)_{\text{Re}})$ and apply Lemma 13.3).

14.6. Let $x \in N^c(\mu, \sigma_x^2)$ and $a \in \mathbb{C}$ with $|a| < \infty$. Show that $ax \in N^c(a\mu, |a|^2\sigma_x^2)$.

14.7. Let $x \in N^c(0, \sigma_x^2)$. Show that $\mathcal{E}\{|x|^4\} = 2\sigma_x^4$.

14.8. Let $x \in N^c(0, \sigma_x^2)$. Show that $\mathcal{E}\{x^n\} = 0$ (hint: use $\mathcal{E}\{u^{2n}\} = \sigma_u^{2n}(2n)!/(2^n n!)$ for $u \in N(0, \sigma_u^2)$ (Stuart and Ord, 1987) and $\sum_{k=0}^{n \text{ div } 2} C_n^{2k} = \sum_{k=0}^{n \text{ div } 2} C_n^{2k+1} = 2^{n-1}$ (Gradshteyn and Ryzhik, 1980)).

14.9. Show that $x \in N_n^c(\mu_x, C_x)$ can be written as $x = \mu_x + Ay$ with $y \in N_p^c(0, I_p)$ and $p = \text{rank}(C_x)$ (hint: use $C_x = AA^H$ with $A \in \mathbb{C}^{n \times p}$ and $\text{rank}(A) = p$).

14.10. Let $x \in N_n(0, I_n)$ $(x \in N_n^c(0, I_n))$ and $P \in \mathbb{R}^{n \times n}$ $(P \in \mathbb{C}^{n \times n})$ a symmetric (Hermitian) idempotent matrix of rank p. Show that $x^T P x \in \chi^2(p)$ $(2x^H P x \in \chi^2(2p))$ (hint: use $P = U \Lambda U^H$ with $U^{-1} = U^H$ and $\Lambda = \text{diag}(I_p, O_{n-p})$ with $p = \text{rank}(P)$).

14.11. Let $x \in \chi^2(n)$. Show that $\mathcal{E}\{1/x\} = 1/(n-2)$ and $\text{var}(1/x) = 2/((n-2)^2(n-4))$ (hint: take $y \in F(n_1, n_2)$ and use the rules $\mathcal{E}\{y\} = \mathcal{E}\{x_1/n_1\}E\{n_2/x_2\}$ and $\text{var}(x_1 x_2) = \sigma_{x_1}^2 \sigma_{x_2}^2 + \sigma_{x_1}^2 \mu_{x_1}^2 + \sigma_{x_2}^2 \mu_{x_2}^2$ with $\mathcal{E}\{x\} = n$, $\mathcal{E}\{y\} = n_2/(n_2-1)$ and $\text{var}(y) = (2n_2^2(n_1 + n_2 - 2))/(n_1(n_2-2)^2(n_2-4))$ (Stuart and Ord, 1987)).

14.12. Let $x \in \mathbb{C}^n$ be circular complex distributed and $f(x) \in \mathbb{C}^m$ an analytic function. Show that $\text{Cov}(f(x))$ is given by (14-24).

14.13. Let $y \in W_p^c(n, C_x)$ with $\text{Im}(C_x) = 0$. Show that $\text{Re}(y) \in W_p(2n, C_x/2)$.

14.14. Show using (14-16) that $\text{cum}(x_k) = \mathcal{E}\{x_k\}$, $\text{cum}(x_k, \bar{x}_k) = \text{var}(x_k)$, and $\text{cum}(x_k, \bar{y}_l) = \text{covar}(x_k, y_l)$.

14.15. Let $x \in \mathbb{C}^n$ be circular complex distributed. Show that the covariance matrix of $f(x, \bar{x}) \in \mathbb{R}^m$ is given by (14-25) (hint: start from (14-23) with x replaced by x_{re}, and use $C_{x_{\text{re}}} = 0.5(C_x)_{\text{Re}}$, Lemma 13.4, and (13-60)).

14.16. Show that the sample mean (14-26) and sample (cross-)covariance matrices (14-27), (14-28) of independent realizations are unbiased estimates of the mean and (cross-)covariance matrices.

14.17. Let $x \in N(0, \sigma_x^2)$ $(x \in N^c(0, \sigma_x^2))$ and calculate the sample variance $\hat{\sigma}_x^2$ of R independent realizations. Show that $(R-1)\hat{\sigma}_x^2/\sigma_x^2 \in \chi^2(R-1)$ $(2(R-1)\hat{\sigma}_x^2/\sigma_x^2 \in \chi^2(2(R-1)))$ (hint: first show that $(R-1)\hat{\sigma}_x^2 = X^H P X$ with $X_{[k]} = x^{[k]}$, $P = I_R - U/R$ and $U_{[k,l]} = 1$, $k, l = 1, 2, \ldots, R$; next prove that P is a symmetric idempotent matrix of rank $R-1$ and apply the results of Exercise 14.10).

14.18. Let $x \in N_n(\mu_x, C_x)$ with $\text{rank}(C_x) = p < n$. Show that $b = (\hat{x} - \mu_x)^T (\hat{C}_x/R)^+(\hat{x} - \mu_x)$ is $\dfrac{p(R-1)}{(R-p)}F(p, R-p)$ distributed (hint: apply the results of Exercise 14.9 on $\hat{x}$ and $\hat{C}_x$, and use Theorem 13.1).

14.19. Let $x(t) \in \mathbb{C}^r$ be mixing of order P and $h(t, u) \in \mathbb{C}^{s \times r}$ the impulse response of a stable linear time-variant multivariable system. Show that $y(t) = \sum_{u=0}^{\infty} h(t, u) x(u)$ is mixing of order P.

14.20. Prove the linearity of the limit in mean square (14-64) (hint: use $(x+y)^2 \leq (x+y)^2 + (x-y)^2 = 2x^2 + 2y^2$).

14.21. Prove that the limit in mean square and the expected value commute (hint: first show that $\mathcal{E}\{x^2\} \geq (\mathcal{E}\{x\})^2$).

14.22. Prove that $\mathcal{E}\{O_{\text{m.s.}}(N^{-k})\} = O(N^{-k})$ (hint: show that $\lim\limits_{N \to \infty} N^k \mathcal{E}\{O_{\text{m.s.}}(N^{-k})\} < \infty$).

14.23. Let $z(k) \in N(0, \sigma_z^2)$, $k = 1, 2, \ldots, N$. Show that the sample variance $\hat{\sigma}_z^2 = \sum_{k=1}^{N} z^2(k)/N$ is an unbiased and efficient estimate of σ_z^2.

14.24. Let $z(k) \in E(\mu_z, \mu_z^2)$, $k = 1, 2, \ldots, N$. Show that the sample mean $\hat{\mu}_z = \sum_{k=1}^{N} z(k)/N$ is an unbiased and efficient estimate of μ_z, with $\text{var}(\hat{\mu}_z) = \mu_z^2/N$. Now consider the estimator $\hat{m} = a\hat{\mu}_z$, where a is chosen to minimize the mean square error $\text{MSE}(\hat{m})$. Show that $a = N/(N+1)$ and that $\text{MSE}(\hat{m}) = \mu_z^2/(N+1)$. Conclude that the MSE of the biased estimate $\hat{m}$ is smaller than that of the unbiased estimate $\hat{\mu}_z$.

14.25. Show that the simple approach (14-117) and the least squares estimates (14-118) tend to the true value R_0 as the signal-to-noise ratio of the current measurements tends to infinity (hint: use (14-120) and let $i_0/\sigma_i \to \infty$).

14.26. Illustrate reasoning (14-98) on the errors-in-variables estimate (14-119).

14.27. Run a Monte Carlo simulation for the resistance measurement problem (14-116) with $u_0 = 1$ V, $i_0 = 1$ A, and iid errors $n_u(k) \in U(0, 12^{-1}\ \text{V}^2)$, $n_i(k) \in U(0, 12^{-1}\ \text{A}^2)$. Calculate the least squares (14-118) and errors-in-variables (14-119) estimates for increasing values of N. Compare the sample variance of the estimates with the values predicted by (14-123b) and (14-123c).

14.28. Let $E(k)$ be the DFT spectrum of iid noise $e(t)$ with existing moments of any order. Consider the sum $S(N) = \sum_{k=0}^{N-1} E(k)$. Show that $N^{-1/2}S(N)$ is not asymptotically normally distributed. Explain. (Hint: use expression (14-182) to show that the cumulants of $N^{-1/2}S(N)$ do not tend to those of a normal random variable; note that $S(N) = e(0)$.)

14.18 APPENDIXES

Appendix 14.A: Indecomposable Sets

Consider a table with I rows and, possibly, a different number of columns per row

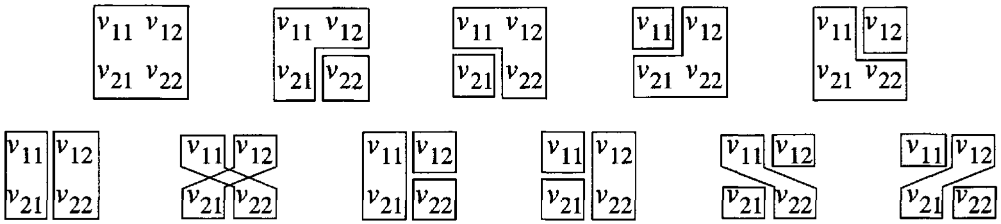
$$\begin{array}{cccc} v_{11} & v_{12} & \cdots & v_{1J_1} \\ v_{21} & v_{22} & \cdots & v_{2J_2} \\ \cdots & \cdots & \cdots & \cdots \\ v_{I1} & v_{I2} & \cdots & v_{IJ_I} \end{array} \tag{14-137}$$

Although some of the entries of this table may take the same numerical value, all elements are considered as distinguishable. Let $\mathbb{P} = \mathbb{P}_1 \cup \mathbb{P}_2 \cup \cdots \cup \mathbb{P}_K$ be a partition of table (14-137). The sets $\mathbb{P}_i$ and $\mathbb{P}_j$ of the partition *hook* if there exist a $v_{i_1 j_1} \in \mathbb{P}_i$ and a $v_{i_2 j_2} \in \mathbb{P}_j$ such that $i_1 = i_2$. Two sets $\mathbb{P}_i$ and $\mathbb{P}_j$ *communicate* if there exists a sequence of sets $\mathbb{P}_i = \mathbb{P}_{m_1}, \mathbb{P}_{m_2}, \ldots, \mathbb{P}_{m_R} = \mathbb{P}_j$ such that $\mathbb{P}_{m_r}$ and $\mathbb{P}_{m_{r+1}}$ hook for $r = 1, 2, \ldots, R-1$. The partition $\mathbb{P}$ is *indecomposable* if all sets of the partition communicate.

Example 14.34: Consider a 2 by 2 table

$$\begin{array}{cc} v_{11} & v_{12} \\ v_{21} & v_{22} \end{array} \tag{14-138}$$

All the indecomposable partitions of (14-138) are given by



□

Lemma 14.35: Let $y_i = \prod_{j=1}^{J_i} v_{ij}$, $i = 1, 2, \ldots, I$, then

$$\text{cum}(y_1, y_2, \ldots, y_I) = \sum_{\mathbb{P}} \text{cum}(v_{ij} \in \mathbb{P}_1)\text{cum}(v_{ij} \in \mathbb{P}_2)\cdots\text{cum}(v_{ij} \in \mathbb{P}_K) \tag{14-139}$$

where the summation is taken over all indecomposable partitions $\mathbb{P} = \mathbb{P}_1 \cup \mathbb{P}_2 \cup \cdots \cup \mathbb{P}_K$ of table (14-137), and with $\text{cum}(v_{ij} \in \mathbb{P}_r)$ the joint cumulant of all the elements of $\mathbb{P}_r$.

Proof. See Leonov and Shiryaev (1959). □

Example 14.36: Suppose we want to calculate $\text{cum}(x_1x_2, x_3x_4)$. Note that the definition $y_i = \prod_{j=1}^{J_i} v_{ij}$ in Lemma 14.35 defines the structure of table (14-137): the index i defines the row and j defines the column. Hence, in this case we have on the first row x_1 and x_2 (the order is not important) and on the second row x_3 and x_4 (the order is not important). Using Lemma 14.35 and the result of Example 14.34 with $v_{11} = x_1$, $v_{12} = x_2$, $v_{21} = x_3$, and $v_{22} = x_4$, we find

$$\begin{aligned}\text{cum}(x_1x_2, x_3x_4) = {} & \text{cum}(x_1, x_2, x_3, x_4) + \\ & \text{cum}(x_1, x_2, x_3)\text{cum}(x_4) + \text{cum}(x_1, x_2, x_4)\text{cum}(x_3) + \\ & \text{cum}(x_2, x_3, x_4)\text{cum}(x_1) + \text{cum}(x_1, x_3, x_4)\text{cum}(x_2) + \\ & \text{cum}(x_1, x_3)\text{cum}(x_2, x_4) + \text{cum}(x_1, x_4)\text{cum}(x_2, x_3) + \\ & \text{cum}(x_1, x_3)\text{cum}(x_2)\text{cum}(x_4) + \text{cum}(x_2, x_4)\text{cum}(x_1)\text{cum}(x_3) + \\ & \text{cum}(x_1, x_4)\text{cum}(x_2)\text{cum}(x_3) + \text{cum}(x_2, x_3)\text{cum}(x_1)\text{cum}(x_4)\end{aligned} \tag{14-140}$$

□

Appendix 14.B: Proof of Lemma 14.5

For $k = 1, 2, \ldots, P$ we have

$$\max_{t_k} \sum_{t_1, t_2, \ldots, t_{k-1} = 0}^{\infty} |\text{cum}(y(t_1), y(t_2), \ldots, y(t_k))| \tag{14-141}$$

$$\leq \max_{t_k} \sum_{u_1, \ldots, u_k = 0}^{\infty} \sum_{t_1, \ldots, t_{k-1} = 0}^{\infty} |h(t_1, u_1)|\ldots|h(t_k, u_k)||\text{cum}(x(u_1), x(u_2), \ldots, x(u_k))| \tag{14-142}$$

$$\leq C^{k-1}\max_{t_k} \sum_{u_k = 0}^{\infty} |h(t_k, u_k)| \sum_{u_1, \ldots, u_{k-1} = 0}^{\infty} |\text{cum}(x(u_1), x(u_2), \ldots, x(u_k))| \tag{14-143}$$

$$\leq C^{k-1}\left(\max_{u_k} \sum_{u_1, \ldots, u_{k-1} = 0}^{\infty} |\text{cum}(x(u_1), x(u_2), \ldots, x(u_k))|\right)\left(\max_{t_k} \sum_{u_k = 0}^{\infty} |h(t_k, u_k)|\right) \tag{14-144}$$

$$\leq C^k \max_{u_k} \sum_{u_1, \ldots, u_{k-1} = 0}^{\infty} \left|\mathrm{cum}(x(u_1), x(u_2), \ldots, x(u_k))\right| \qquad (14\text{-}145)$$

$$< \infty \qquad (14\text{-}146)$$

Inequality (14-143) is obtained by applying $k-1$ times (14-38a), inequality (14-145) uses (14-38b), and inequality (14-146) uses the mixing of order P property of x. □

Appendix 14.C: Proof of Lemma 14.8

The proof follows exactly the same lines of the proof of Theorem 2.9.1 of Brillinger (1981), modified for the more general mixing definition (14-33) as in the proof of Lemma 14.5. Instead of repeating the proof of Brillinger (1981), we will illustrate the differences on the second-order cumulant of a second-degree Volterra system. Background information concerning the new concepts required for this proof can be found in Appendix 14.A. Assuming that the system is causal and that the input is zero for negative times, we have

$$y(t) = \sum_{u_1, u_2 = 0}^{t} h_2(t-u_1, t-u_2)x(u_1)x(u_2) = \sum_{u_1, u_2 = 0}^{t} h_2(u_1, u_2)x(t-u_1)x(t-u_2) \qquad (14\text{-}147)$$

(for noncausal systems and noncausal inputs the sums in (14-147) are extended from $-\infty$ to $+\infty$). The second-order cumulant of $y(t)$ is

$$\mathrm{cum}(y(t_1), y(t_2)) = \sum_{u_{11}, u_{12} = 0}^{t_1} \sum_{u_{21}, u_{22} = 0}^{t_2} h_2(u_{11}, u_{12})h_2(u_{21}, u_{22}) \cdot \\ \mathrm{cum}(x(t_1-u_{11})x(t_1-u_{12}), x(t_2-u_{21})x(t_2-u_{22})) \qquad (14\text{-}148)$$

The mixing condition of order 2 becomes

$$\max_{t_2 \geq 0} \sum_{t_1 = 0}^{\infty} \left|\mathrm{cum}(y(t_1), y(t_2))\right| \leq \left(\sum_{u_{11}, u_{12} = 0}^{\infty} |h_2(u_{11}, u_{12})|\right)\left(\sum_{u_{21}, u_{22} = 0}^{\infty} |h_2(u_{21}, u_{22})|\right) \cdot \\ \max_{\substack{u_{11}, u_{12}, u_{12}, u_{22} \\ t_2 \geq 0}} \left(\sum_{t_1 = 0}^{\infty} \left|\mathrm{cum}(x(t_1-u_{11})x(t_1-u_{12}), x(t_2-u_{21})x(t_2-u_{22}))\right|\right) \qquad (14\text{-}149)$$

$$\leq C_2^2 \max_{u_1, u_2, v_2} \sum_{v_1 = 0}^{\infty} \left|\mathrm{cum}(x(v_1)x(v_1+u_1), x(v_2)x(v_2+u_2))\right|$$

where $v_1 = t_1 - u_{11}$, $u_1 = u_{11} - u_{12}$, $v_2 = t_2 - u_{21}$, and $u_2 = u_{21} - u_{22}$. We will now prove that

$$\max_{u_1, u_2, v_2} \sum_{v_1 = 0}^{\infty} \left|\mathrm{cum}(x(v_1)x(v_1 + u_1), x(v_2)x(v_2 + u_2))\right| \le C < \infty \tag{14-150}$$

which shows that $y(t)$ is mixing of order 2. From Example 14.36 it follows that the cumulant in (14-150) contains 4 four kinds of contributions. For each type of contribution we will show that (14-150) is valid. We find for the type I contribution (fourth-order cumulant);

$$\begin{aligned} &\max_{u_1, u_2, v_2} \sum_{v_1 = 0}^{\infty} \left|\mathrm{cum}(x(v_1), x(v_1 + u_1), x(v_2), x(v_2 + u_2))\right| \le \\ &\max_{v_2} \sum_{v_1, u_1, u_2 = 0}^{\infty} \left|\mathrm{cum}(x(v_1), x(u_1), x(v_2), x(u_2))\right| \le C \end{aligned} \tag{14-151}$$

for the type II contributions (product of third- and first-order cumulants)

$$\begin{aligned} &\max_{u_1, u_2, v_2} \sum_{v_1 = 0}^{\infty} \left|\mathrm{cum}(x(v_1), x(v_2), x(v_2 + u_2))\mathrm{cum}(x(v_1 + u_1))\right| \le \\ &\left(\max_{v_2} \sum_{v_1, u_2 = 0}^{\infty} \left|\mathrm{cum}(x(v_1), x(v_2), x(u_2))\right|\right)\left(\max_{u_1}\left| \mathrm{cum}(x(u_1))\right|\right) \le C \end{aligned} \tag{14-152}$$

and similarly for the three other third-order terms; for the type III contributions (product of two second-order cumulants)

$$\begin{aligned} &\max_{u_1, u_2, v_2} \sum_{v_1 = 0}^{\infty} \left|\mathrm{cum}(x(v_1), x(v_2 + u_2))\mathrm{cum}(x(v_1 + u_1), x(v_2))\right| \le \\ &\left(\max_{u_2} \sum_{v_1 = 0}^{\infty} \left|\mathrm{cum}(x(v_1), x(u_2))\right|\right)\left(\max_{v_2} \sum_{u_1 = 0}^{\infty} \left| \mathrm{cum}(x(u_1), x(v_2))\right|\right) \le C \end{aligned} \tag{14-153}$$

and similarly for the other term; and finally for the type IV contributions (product of second-order cumulant and two first-order cumulants)

$$\begin{aligned} &\max_{u_1, u_2, v_2} \sum_{v_1 = 0}^{\infty} \left|\mathrm{cum}(x(v_1 + u_1), x(v_2 + u_2))\mathrm{cum}(x(v_1))\mathrm{cum}(x(v_2))\right| \le \\ &\left(\max_{u_2} \sum_{v_1 = 0}^{\infty} \left|\mathrm{cum}(x(v_1), x(u_2))\right|\right)\left(\max_{v_1}\left| \mathrm{cum}(x(v_1))\right|\right)\left(\max_{v_2}\left| \mathrm{cum}(x(v_2))\right|\right) \le C \end{aligned} \tag{14-154}$$

and similarly for the three other terms. The last inequalities in (14-151), (14-152), (14-153), and (14-154) are due to the fourth-order mixing property of $x(t)$. The basic reason why all the terms in (14-140) have a finite contribution to (14-150) is that they all stem from indecomposable partitions.

Extending the summations in all the equations from $-\infty$ to $+\infty$ shows that the proof is also valid for noncausal systems and noncausal inputs. □

Appendix 14.D: Almost Sure Convergence Implies Convergence in Probability

$x(N)$ converges w.p. 1 to x, hence, for any $\varepsilon, \delta > 0$ there exists an N such that

$$1-\delta \le \text{Prob}(\sup_{k\ge N}|x(k)-x| \le \varepsilon) = \text{Prob}(\bigcap_{k=N}^{\infty} |x(k)-x| \le \varepsilon) \le \text{Prob}(|x(N)-x| \le \varepsilon)$$

The last inequality shows that $x(N)$ converges in probability to x. □

Appendix 14.E: Convergence in Mean Square Implies Convergence in Probability

$x(N)$ converges in mean square to x, hence, for any $\varepsilon, \delta > 0$ there exists a N_0 such that for every $N > N_0$, $\mathcal{E}\{|x(N)-x|^2\} \le \delta\varepsilon^2$. Using Chebyshev's inequality (14-20) we find

$$\text{Prob}(|x(N)-x| \le \varepsilon) = 1-\text{Prob}(|x(N)-x| > \varepsilon) \ge 1-\frac{1}{\varepsilon^2}\mathcal{E}\{|x(N)-x|^2\} \ge 1-\delta$$

which shows that $x(N)$ converges in probability to x. □

Appendix 14.F: The Borel-Cantelli Lemma

If $\text{Prob}(A \cup B) = 1$, then $\text{Prob}(A \cap B) = 1-\text{Prob}(\overline{A \cap B}) = 1-\text{Prob}(\bar{A} \cup \bar{B})$. Hence,

$$\text{Prob}(\sup_{k\ge N}|x(k)-x| \le \varepsilon) = \text{Prob}(\bigcap_{k=N}^{\infty} |x(k)-x| \le \varepsilon) = 1-\text{Prob}(\bigcup_{k=N}^{\infty} |x(k)-x| > \varepsilon)$$

Using $\text{Prob}(A \cup B) \le \text{Prob}(A) + \text{Prob}(B)$ and Chebyshev's inequality (14-20) we find

$$\text{Prob}(\sup_{k\ge N}|x(k)-x| \le \varepsilon) \ge 1-\sum_{k=N}^{\infty} \text{Prob}(|x(k)-x| > \varepsilon) \ge 1-\frac{1}{\varepsilon^2}\sum_{k=N}^{\infty} \mathcal{E}\{|x(k)-x|^2\} \quad (14\text{-}155)$$

From (14-56) it follows that for any $\varepsilon, \delta_1, \delta_2 > 0$ there exists an N such that for every $k \ge N$

$$\textstyle\sum_{k=N}^{\infty}\text{Prob}(|x(k)-x| > \varepsilon) < \delta_1 \quad \text{and} \quad \sum_{k=N}^{\infty}\mathcal{E}\{|x(k)-x|^2\} < \delta_2\varepsilon^2 \qquad (14\text{-}156)$$

Putting (14-156) in (14-155) proves the theorem. □

Appendix 14.G: Proof of the (Strong) Law of Large Numbers for Mixing Sequences

We will give the proof for the general case where the sequence in the partial sum depends on N. To simplify the notations, we introduce the zero mean variables $y(N) = (S(N) - \mathcal{E}\{S(N)\})/N$ and $z_N(k) = x_N(k) - \mathcal{E}\{x_N(k)\}$. The law of large numbers then becomes $\underset{N \to \infty}{\text{l.i.m.}} y(N) = 0$. The calculation of this limit in mean square requires an expression for $\text{cum}(y(N), y(N))$ (see (14-51)). Using the multilinearity of the cumulant (see Section 14.1, property 2) and the second-order mixing condition of $x_N(k)$, we find

$$\text{cum}(y(N), y(N)) = \frac{1}{N^2}\sum_{k,\,l=1}^{N}\text{cum}(x_N(k), x_N(l)) = O(N^{-1}) \tag{14-157}$$

where the last equality is due to (14-36). From (14-157), it follows that $\lim_{N \to \infty}\mathcal{E}\{y^2(N)\} = 0$, which already proves the law of large numbers for mixing sequences, and that $\text{var}(y(N)) = O(N^{-1})$. Using properties 6 and 7 of Section 14.8, it follows that the convergence rate of the limit in mean square is an $O_{\text{m.s.}}(N^{-1/2})$.

Result (14-157) is not sufficient to prove the strong convergence of $y(N)$ via the Borel-Cantelli Lemma 14.10 ($\sum_{N=1}^{\infty} 1/N = \infty$). Therefore, we will first prove that the subsequence $y(N^2)$ converges w.p. 1 to zero. Next, we will show that the deviation of any element in the main sequence $y(N)$ with a nearby element in the subsequence $y(N^2)$ converges to zero w.p. 1. It is easy to see that $\text{cum}(y(N^2), y(N^2)) = O(N^{-2})$ and therefore

$$\sum_{N=1}^{\infty}\mathcal{E}\{y^2(N^2)\} = \sum_{N=1}^{\infty}O(N^{-2}) < \infty \tag{14-158}$$

Applying the Borel-Cantelli Lemma 14.10 to (14-158) shows that $\underset{N \to \infty}{\text{a.s.lim}}\, y(N^2) = 0$. It remains to be proved that this implies the strong convergence of the whole sequence $y(N)$. Therefore, it is sufficient to show that the maximal difference between the subsequence and the complete sequence

$$\sup_{N^2 < k \leq (N+1)^2}|y(k) - y(N^2)| \tag{14-159}$$

converges to zero w.p. 1. The difference $y(k) - y(N^2)$ can be rewritten as the sum of three contributions

$$y(k) - y(N^2) = \Delta_1(k) + \Delta_2(k) + \Delta_3(k) \tag{14-160}$$

with

$$\Delta_1(k) = -\frac{(k - N^2)}{k}y(N^2),\ \Delta_2(k) = \frac{1}{k}\sum_{r=N^2+1}^{k} z_k(r) \text{ and } \Delta_3(k) = \frac{1}{k}\sum_{r=1}^{N^2}(z_k(r) - z_{N^2}(r))$$

Using $\sup_k|a(k) + b(k)| \leq \sup_k|a(k)| + \sup_k|b(k)|$, (14-159) is bounded above by

$$\sup_{N^2 < k \leq (N+1)^2}|\Delta_1(k)| + \sup_{N^2 < k \leq (N+1)^2}|\Delta_2(k)| + \sup_{N^2 < k \leq (N+1)^2}|\Delta_3(k)| \tag{14-161}$$

The strong convergence to zero of the first term in (14-161) is first established. Using $\sup_{r<k\le s} |a(k)| \le \sum_{k=r+1}^{s} |a(k)|$, it is bounded above by

$$\sum_{k=N^2+1}^{(N+1)^2} |\Delta_1(k)| \le |y(N^2)|\left(\max_{N^2<k\le(N+1)^2}\left|\frac{k-N^2}{k}\right|\right)((N+1)^2-N^2) \tag{14-162}$$

Using $\max_{N^2<k\le(N+1)^2}\left|\frac{k-N^2}{k}\right| = \frac{(N+1)^2-N^2}{(N+1)^2}$ for $N \ge 1$, (14-162) becomes

$$\sum_{k=N^2+1}^{(N+1)^2} |\Delta_1(k)| \le |y(N^2)|\left(\frac{2N+1}{N+1}\right)^2 = |y(N^2)|O(N^0) \tag{14-163}$$

Because the subsequence $y(N^2)$ converges strongly to zero, so does (14-163).

The strong convergence to zero of the second term in (14-161) will be established by application of the Borel-Cantelli Lemma 14.10. This requires that its variance decreases sufficiently rapidly to zero as $N \to \infty$. Using $\mathrm{var}(a(k)) \le \mathcal{E}\{a^2(k)\}$, $(\sup_k |a(k)|)^2 = \sup_k |a(k)|^2$, and $\sup_{r<k\le s} |b(k)| \le \sum_{k=r+1}^{s} |b(k)|$, we find

$$\begin{aligned}\mathrm{var}(\sup_{N^2<k\le(N+1)^2}|\Delta_2(k)|) &\le \mathcal{E}\{\sup_{N^2<k\le(N+1)^2}|\Delta_2(k)|^2\}\\ &\le \sum_{k=N^2+1}^{(N+1)^2} \mathcal{E}\{|\Delta_2(k)|^2\}\\ &\le \sum_{k=N^2+1}^{(N+1)^2} \frac{1}{k^2} \sum_{r,s=N^2+1}^{k} \mathrm{cum}(z_k(r), z_k(s))\end{aligned} \tag{14-164}$$

As $z_k(r)$ is mixing of order 2, (14-164) is bounded above by (see (14-36))

$$\begin{aligned}\mathrm{var}(\sup_{N^2<k\le(N+1)^2}|\Delta_2(k)|) &\le \sum_{k=N^2+1}^{(N+1)^2} \frac{1}{k^2}O(k-N^2)\\ &\le C\sum_{k=N^2+1}^{(N+1)^2} \frac{k-N^2}{k^2}\end{aligned} \tag{14-165}$$

Using $\max_{N^2<k\le(N+1)^2}\left|\frac{k-N^2}{k^2}\right| = \frac{(N+1)^2-N^2}{(N+1)^4}$ for $N \ge 3$, (14-165) becomes

$$\begin{aligned}\mathrm{var}(\sup_{N^2<k\le(N+1)^2}|\Delta_2(k)|) &\le C\left(\max_{N^2<k\le(N+1)^2}\left|\frac{k-N^2}{k^2}\right|\right)((N+1)^2-N^2)\\ &\le O(N^{-2})\end{aligned} \tag{14-166}$$

Applying the Borel-Cantelli lemma to (14-166) shows that the second term in (14-161) converges to zero w.p. 1.

The strong convergence to zero of the third term in (14-161) is shown following exactly the same lines as that of the second term. Similar to (14-164), we find

$$\operatorname{var}(\sup_{N^2 < k \leq (N+1)^2} |\Delta_3(k)|) \leq \sum_{k = N^2+1}^{(N+1)^2} \frac{1}{k^2} \operatorname{var}(\sum_{r=1}^{N^2} (z_k(r) - z_{N^2}(r))) \tag{14-167}$$

Using the assumption $\operatorname{var}(\sum_{r=1}^{N^2} (z_k(r) - z_{N^2}(r))) = O(k - N^2)$ and following the lines of (14-165) and (14-166), (14-167) becomes

$$\operatorname{var}(\sup_{N^2 < k \leq (N+1)^2} |\Delta_3(k)|) \leq O(N^{-2}) \tag{14-168}$$

Applying the Borel-Cantelli lemma to (14-168) shows that the third term in (14-161) converges to zero w.p. 1. Finally, it follows that $\text{a.s.}\lim_{N \to \infty} y(N) = 0$, which proves the strong law of large numbers for mixing sequences. □

Appendix 14.H: Proof of the Central Limit Theorem for Mixing Sequences

We will show that the cumulants of $S(N)/\sqrt{N}$ converge for $N \to \infty$ to those of a normal distribution. This concludes the proof because the normal distribution is uniquely determined by its moments (see the Fréchet-Shohat Lemma 14.11).

The proof will be given for the general case where the sequence in the partial sum depends on N. Using the multilinearity of the cumulant (see Section 14.1, property 2), we find for the Jth order cumulant $c_J(N)$ of $S(N)/\sqrt{N}$

$$c_J(N) = \frac{1}{N^{J/2}} \sum_{k_1, k_2, \ldots, k_J = 1}^{N} \operatorname{cum}(x_N(k_1), x_N(k_2), \ldots, x_N(k_J)) \tag{14-169}$$

Because $x_N(k)$ is mixing of order J, $J = 1, 2, \ldots$, the summation in (14-169) is an $O(N)$ (see (14-36)), so that

$$\lim_{N \to \infty} c_J(N) = \lim_{N \to \infty} O(N^{1-J/2}) = 0 \text{ for } J > 2 \tag{14-170}$$

By assumption $\operatorname{var}(S(N)) = O(N)$ so that

$$\lim_{N \to \infty} c_2(N) = C \text{ with } 0 < C < \infty \tag{14-171}$$

We conclude from (14-170) and (14-171) that the $c_J(N)$ converge to the cumulants of a normal distribution (see Example 14.2). The mixing assumption guarantees that the second-order cumulants of $x_N(k)$ are uniformly bounded. This ensures that the number of random variables $x_N(k)$ that have an $O(N^0) \geq C > 0$ contribution to $\operatorname{var}(S(N)) = O(N)$ increases as $O(N)$.

The cumulants are the coefficients in the Taylor series expansion of the logarithm of the characteristic function $\phi(t)$ (Brillinger, 1981). From (14-170), it follows that $\ln(\phi(t))$ corresponding to $S(N)$ equals that of a normal random variable within an $O(N^{-1/2})$, uniformly in t. Because the characteristic function is related to the probability density function by the Fourier integral, it follows that the distribution function $F_N(y)$ of $S(N)$ equals that of a normal random variable within an $O(N^{-1/2})$, uniformly in y. □

Appendix 14.I: Generalized Cauchy-Schwarz Inequality for Random Vectors

Let $U \in \mathbb{C}^n$ and $V \in \mathbb{C}^k$ be complex random vectors. Define the n by n matrix $M = \mathcal{E}\{(U-\Gamma V)(U-\Gamma V)^H\}$ with $\Gamma \in \mathbb{C}^{n\times k}$. By construction, M is positive semidefinite

$$\mathcal{E}\{(U-\Gamma V)(U-\Gamma V)^H\} \geq 0 \tag{14-172}$$

Elaborating inequality (14-172) with $\Gamma = \mathcal{E}\{UV^H\}[\mathcal{E}\{VV^H\}]^+$ using property 2 of the pseudoinverse + (see Section 13.5) gives the *generalized Cauchy-Schwarz inequality* for random vectors

$$\mathcal{E}\{UU^H\} - \mathcal{E}\{UV^H\}[\mathcal{E}\{VV^H\}]^+\mathcal{E}\{VU^H\} \geq 0 \tag{14-173}$$

We will show that (14-173) reaches the lower bound if and only if there exists a nonrandom matrix Γ such that $U = \Gamma V$. Equality holds in (14-172), and, hence, also in (14-173), if and only if all the eigenvalues of the matrix $M = \mathcal{E}\{(U-\Gamma V)(U-\Gamma V)^H\}$ are zero. The positive semidefiniteness of M implies that all its eigenvalues are zero if and only if their sum equals zero. Hence, $\operatorname{tr}(M) = 0$ (see Exercise 13.12.), so that

$$\operatorname{tr}(\mathcal{E}\{(U-\Gamma V)(U-\Gamma V)^H\}) = \mathcal{E}\{(U-\Gamma V)^H(U-\Gamma V)\} = \mathcal{E}\{\|U-\Gamma V\|_2^2\} = 0 \tag{14-174}$$

The last equality in is true if and only if $U = \Gamma V$. □

Appendix 14.J: Proof of the Generalized Cramér-Rao Inequality (Theorem 14.18)

The generalized Cramér-Rao inequality (14-84) follows as a special case of the generalized Cauchy-Schwarz inequality for random vectors (see Appendix 14.I). In what follows, all the expectations are taken w.r.t. the measurements z. Choosing $U = \hat{G}(\hat{\theta}(z)) - \mathcal{E}\{\hat{G}(\hat{\theta}(z))\}$ and $V^H = \partial \ln f_z(z, \theta_0)/\partial\theta_0$ in (14-173), taking into account that

$$\mathcal{E}\left\{\frac{\partial \ln f_z(z, \theta_0)}{\partial\theta_0}\right\} = \frac{\partial}{\partial\theta_0}\int_z f_z(z, \theta_0)dz = \frac{\partial 1}{\partial\theta_0} = 0 \tag{14-175}$$

$$\mathcal{E}\left\{\hat{G}(\hat{\theta}(z))\frac{\partial \ln f_z(z, \theta_0)}{\partial\theta_0}\right\} = \frac{\partial}{\partial\theta_0}\int_z \hat{G}(\hat{\theta}(z))f_z(z, \theta_0)dz = \frac{\partial G(\theta_0)}{\partial\theta_0} + \frac{\partial b_G}{\partial\theta_0} \tag{14-176}$$

gives

$$\operatorname{Cov}(\hat{G}(\hat{\theta}(z))) \geq \left(\frac{\partial G(\theta_0)}{\partial\theta_0} + \frac{\partial b_G}{\partial\theta_0}\right) Fi^+ \left(\frac{\partial G(\theta_0)}{\partial\theta_0} + \frac{\partial b_G}{\partial\theta_0}\right)^H \tag{14-177}$$

Adding $b_G b_G^T$ to both sides of (14-177) gives (14-84). The second equality in (14-85) is obtained by differentiating (14-175) w.r.t. θ_0

$$\begin{gathered}\frac{\partial}{\partial\theta_0}\int_Z \frac{\partial \ln f_z(z, \theta_0)}{\partial\theta_0} f_z(z, \theta_0)dz = 0 \\ \Downarrow \\ \mathcal{E}\left\{\frac{\partial^2 \ln f_z(z, \theta_0)}{\partial\theta_0^2}\right\} + \mathcal{E}\left\{\left(\frac{\partial \ln f_z(z, \theta_0)}{\partial\theta_0}\right)^T\left(\frac{\partial \ln f_z(z, \theta_0)}{\partial\theta_0}\right)\right\} = 0\end{gathered} \tag{14-178}$$

Equations (14-175), (14-176), and (14-178) assume that the necessary regularity conditions to allow for the reversal of the order of differentiation and integration are satisfied (Z is the θ_0-independent range of integration). The suitable regularity conditions for the existence of the expected values and the derivatives can be found in Caines (1988). The necessary and sufficient condition (14-86) to attain the lower bound is a direct consequence of the generalized Cauchy-Schwarz inequality (see Appendix 14.I).

Appendix 14.K: Proof of Lemma 14.23

Applying Eq. (5-97) on page 165 to the polynomial $J(z^{-1}, \theta) = \sum_{m=0}^{n_j} j_m z^{-m}$ in the noise model (14-128) gives

$$J(z^{-1}, \theta) = N^{-1/2}\left(\sum_{m=1}^{n_c}\sum_{t=1}^{m} c_m \Delta_N e(t) z^{t-m} - \sum_{n=1}^{n_d}\sum_{t=1}^{n} d_n \Delta_N v(t) z^{t-n}\right) \tag{14-179}$$

where $\Delta_N x(t) = x(-t) - x(N-t)$ with $x = e, v$. It shows that the coefficients j_m, $m = 0, 1, ..., n_j$, of $J(z^{-1}, \theta)$ depend linearly on $2(n_c + n_d)$ (finite number independent of N) random variables. Because $e(t)$ has uniformly bounded absolute moments of order $2 + \delta$, we also have that $\mathcal{E}\{|v(t)|^{2+\delta}\} \leq c < \infty$ (bounded input–bounded output property of a stable system $H(z^{-1})$). Hence, the coefficients j_m of $J(z^{-1}, \theta)$ can be written as

$$j_m = N^{-1/2}x(N) \text{ with } \mathcal{E}\{x(N)\} = 0 \text{ and } \mathcal{E}\{|x(N)|^{2+\delta}\} \leq c < \infty \tag{14-180}$$

(c is independent of t). Because $\text{var}(j_m) = O(N^{-1})$ it follows that $j_m = O_{\text{m.s.}}(N^{-1/2})$, which implies that $j_m = O_p(N^{-1/2})$ (see Section 14.7, interrelation 2). Applying Markov's inequality (14-21) with $p = 2 + \delta$ to $j_m(N)$ (14-180), we find

$$\sum_{N=1}^{\infty} \text{Prob}(|j_m(N)| > \varepsilon) \leq \sum_{N=1}^{\infty} \frac{1}{\varepsilon^{2+\delta}} \mathcal{E}\{|x(N)|^{2+\delta}\} \mathrm{N}^{-(1+\delta/2)} < \infty \tag{14-181}$$

which shows that $j_m(N)$ converges w.p. 1 to zero (see the Borel-Cantelli Lemma 14.10). Using properties 1 and 3 of the stochastic limits (see Section 14.8), it follows that the results for j_m are also valid for $J(z^{-1}, \theta)$. □

Appendix 14.L: Proof of Lemma 14.24

We will show that the joint cumulants of $E(k)$ tend for $N \to \infty$ to those of an independent, circular complex normally distributed random variable (the joint cumulants of order 3 and larger of a multivariate complex normal random variable are zero, see Example 14.2).

This concludes the proof because the normal distribution is uniquely determined by its moments (see the Fréchet-Shohat Lemma 14.11).

The Jth order joint cumulant of $E(k)$ is

$$\operatorname{cum}(E(k_1), E(k_2), \ldots, E(k_J)) = \frac{1}{N^{J/2}} \sum_{t_1, t_2, \ldots, t_J = 0}^{N-1} e^{-\frac{2\pi}{N} j \sum_{i=1}^{J} k_i t_i} \operatorname{cum}(e(t_1), e(t_2), \ldots, e(t_J))$$

with $k_i = 0, 1, \ldots, N-1$, $i = 1, 2, \ldots, J$, and $J = 1, 2, \ldots$. Because the noise $e(t)$ is iid, $\operatorname{cum}(e(t_1), e(t_2), \ldots, e(t_J))$ is different from zero only when $t_1 = t_2 = \cdots = t_J$ (see Section 14.1, property 3 of the cumulants). Putting $C_J = \operatorname{cum}(e(t), e(t), \ldots, e(t))$ and using $\sum_{k=0}^{N-1} x^k = (1-x^N)/(1-x)$, we find

$$\operatorname{cum}(E(k_1), E(k_2), \ldots, E(k_J)) = \frac{C_J}{N^{J/2-1}} \delta\left(\left(\sum_{i=1}^{J} k_i\right) \bmod N\right) \tag{14-182}$$

with $\delta(k)$ the Kronecker delta. From (14-182) it follows that for any N

$$\begin{array}{llr} \operatorname{cum}(E(k)) = 0 & k \neq 0 & \text{(a)} \\ \operatorname{cum}(E(k_1), E(k_2)) = 0 & (k_1 + k_2) \bmod N \neq 0 & \text{(b)} \\ \operatorname{cum}(E(k_1), E(N-k_1)) = \operatorname{cum}(E(k_1), \bar{E}(k_1)) = C_2 & & \text{(c)} \end{array} \tag{14-183}$$

All the cumulants of order larger than 2 tend to zero as $N \to \infty$ (14-182) and $\operatorname{var}(E(k)) = O(N^0)$ (14-183c). We conclude from (14-182) and (14-183) that the cumulants equal, asymptotically, those of a zero mean (DC not included) independent circular complex normally distributed random variable.

The proof of the convergence rate of the distribution function is similar to that given in Appendix 14.H. □

Appendix 14.M: Proof of Lemma 14.26

Because $|E(k)|^2$ is given by

$$|E(k)|^2 = \frac{1}{N} \sum_{t, u = 0}^{N-1} e(t) e(u) e^{-\frac{2\pi}{N} jk(t-u)} \tag{14-184}$$

we find for the Jth order joint cumulant of $|E(k)|^2$

$$\begin{aligned} c_J(k_1, k_2, \ldots, k_J) &= \operatorname{cum}(|E(k_1)|^2, |E(k_2)|^2, \ldots, |E(k_J)|^2) \\ &= \frac{1}{N^J} \sum_{\substack{t_1, t_2, \ldots, t_J = 0 \\ u_1, u_2, \ldots, u_J = 0}}^{N-1} e^{-\frac{2\pi}{N} j \sum_{i=1}^{J} k_i (t_i - u_i)} \quad c(t_1, \ldots, t_J, u_1, \ldots, u_J) \end{aligned} \tag{14-185}$$

with $c(t_1, \ldots, t_J, u_1, \ldots, u_J) = \text{cum}(e(t_1)e(u_1), e(t_2)e(u_2), \ldots, e(t_J)e(u_J))$. Application of Lemma 14.35 with $v_{i1} = e(t_i)$ and $v_{i2} = e(u_i)$ gives

$$c(t_1, \ldots, t_J, u_1, \ldots, u_J) = \sum_{\mathbb{P}} \text{cum}(e(t_{ij}) \in \mathbb{P}_1)\text{cum}(e(t_{ij}) \in \mathbb{P}_2)\ldots\text{cum}(e(t_{ij}) \in \mathbb{P}_K) \quad (14\text{-}186)$$

with $e(t_{ij})$ an element of table (14-137) with $J_1 = J_2 = \cdots = J_J = 2$ and $I = J$, and where the summation extends over all indecomposable partitions $\mathbb{P} = \mathbb{P}_1 \cup \mathbb{P}_2 \cup \cdots \cup \mathbb{P}_K$ of this table. Since $e(t)$ is iid, $\text{cum}(e(t_{ij}) \in \mathbb{P}_k)$ is different from zero if and only if for all $e(t_{ij}) \in \mathbb{P}_k$, $e(t_{ij}) = e(u_k)$ (see Section 14.1, property 3 of the cumulants), and $\text{cum}(e(t_{ij} = u_k) \in \mathbb{P}_k) = c(\mathbb{P}_k)$ is independent of u_k. Putting these results in (14-185) gives

$$c_J(k_1, \ldots, k_J) = \frac{1}{N^J}\sum_{\mathbb{P}} c(\mathbb{P}_1)\ldots c(\mathbb{P}_K) \sum_{u_1, \ldots, u_K = 0}^{N-1} \exp(-\frac{2\pi}{N}j\sum_{i=1}^{J} k_i(u_{r_i} - u_{s_i})) \quad (14\text{-}187)$$

with $r_i, s_i \in \{1, 2, \ldots, K\}$. Because the partition $\mathbb{P} = \mathbb{P}_1 \cup \mathbb{P}_2 \cup \cdots \cup \mathbb{P}_K$ is indecomposable, all the differences $u_{r_i} - u_{s_i}$, $i = 1, 2, \ldots, J$, are obtained by addition and subtraction of the $K-1$ independent differences $u_K - u_k$, $k = 1, 2, \ldots, K-1$ (see Lemma 2.3.1, p. 20, Brillinger, 1981)

$$u_{r_i} - u_{s_i} = \sum_{k=1}^{K-1} A_{[i,k]}(u_K - u_k) \text{ for } i = 1, 2, \ldots, J \quad (14\text{-}188)$$

with $A \in \{-1, 0, 1\}^{J \times (K-1)}$ and $\text{rank}(A) = K-1$. Using (14-188), the second summation in the right-hand side of (14-187) becomes

$$\begin{aligned}\sum_{u_1, \ldots, u_K = 0}^{N-1} \exp(-\frac{2\pi}{N}j\sum_{i=1}^{J} k_i(u_{r_i} - u_{s_i})) &= \sum_{u_K = 0}^{N-1} \exp(-\frac{2\pi}{N}ju_K\sum_{k=1}^{K-1}\sum_{i=1}^{J} k_iA_{[i,k]}) \cdot \\ &\quad \prod_{k=1}^{K-1}\sum_{u_k=0}^{N-1} \exp(\frac{2\pi}{N}ju_k\sum_{i=1}^{J} k_iA_{[i,k]})\end{aligned} \quad (14\text{-}189)$$

Applying K times $\sum_{k=0}^{N-1} x^k = (1-x^N)/(1-x)$ to (14-189) shows that (14-189) is zero unless $K-1$ linear independent constraints are satisfied $(\text{rank}(A) = K-1)$

$$\sum_{u_1, \ldots, u_K = 0}^{N-1} \exp(-\frac{2\pi}{N}j\sum_{i=1}^{J} k_i(u_{r_i} - u_{s_i})) = N^K \Leftrightarrow (k^TA) \bmod N = 0 \quad (14\text{-}190)$$

with $k^T = [k_1, k_2, \ldots, k_J]$. Using (14-190), (14-187), we find

$$c_J(k_1, \ldots, k_J) = \begin{cases} \frac{1}{N^{J-K}}\sum_{\mathbb{P}} c(\mathbb{P}_1)\ldots c(\mathbb{P}_K) \Leftrightarrow (k^TA) \bmod N = 0 \\ 0 \qquad\qquad\qquad\qquad \text{elsewhere} \end{cases}$$

Hence, the mixing condition of the Jth order joint cumulant becomes

$$\max_{k_J} \sum_{k_1, \ldots, k_{J-1} = 1}^{N-1} |c_J(k_1, \ldots, k_J)| = \left|\sum_{\mathbb{P}} (N-1)^{J-K} N^{K-J} c(\mathbb{P}_1) \ldots c(\mathbb{P}_K)\right| = O(N^0)$$

where the last equality is due to the fact that the number of indecomposable sets is independent of N. □

Appendix 14.N: Proof of Lemma 14.27

Using (14-183b) and (14-183c) we find $\max_{k_2} \sum_{k_1 = 1}^{N-1} |\mathrm{cum}(E(k_1), E(k_2))| = C_2$, so that $E(k)$ is mixing of order 2.

The proof that $|E(k)|^2$ is mixing of order 2 is similar to the proof of Lemma 14.26, except that the third- and fourth-order cumulants of $e(t)$ are now not necessarily stationary. Hence, it is sufficient to study only the contribution of these nonstationary cumulants to the mixing condition. The second-order cumulant of $|E(k)|^2$ is given by (14-185) with $J = 2$. Equation (14-140) with $x_1 = e(t_1)$, $x_2 = e(u_1)$, $x_3 = e(t_2)$, and $x_4 = e(u_2)$ gives an explicit expression for (14-186) with $J = 2$. The eight terms in (14-140) containing a first-order cumulant have a zero contribution to (14-185) because

$$\sum_{t=0}^{N-1} \mathrm{cum}(e(t)) e^{-\frac{2\pi}{N} jkt} = \mu \sum_{t=0}^{N-1} e^{-\frac{2\pi}{N} jkt} = 0 \Leftrightarrow k \neq 0$$

This eliminates all the (nonstationary) third-order cumulants. Because $e(t)$ is independent over t, the contribution of the (nonstationary) fourth-order cumulant to (14-185) becomes

$$\frac{1}{N^2} \sum_{\substack{t_1, t_2 = 0 \\ u_1, u_2 = 0}}^{N-1} \mathrm{cum}(e(t_1), e(u_1), e(t_2), e(u_2)) e^{-\frac{2\pi}{N} j \sum_{i=1}^{2} k_i(t_i - u_i)} = \frac{1}{N^2} \sum_{t=0}^{N-1} C_4(t) \qquad (14\text{-}191)$$

Taking into account that $|C_4(t)|$ is uniformly bounded, it can be seen that (14-191) has a $O(N^0)$ contribution to the mixing condition. □

Appendix 14.O: Proof of Theorem 14.28

Applying Lemma 14.23 to the partial sums (14-129) gives

$$\begin{aligned} &\frac{1}{F} \sum_{k \in \mathbb{F}} W_k V(k) \rightarrow \frac{1}{F} \sum_{k \in \mathbb{F}} W_k H(z_k^{-1}) E(k) && \text{w.p. 1} \\ &\frac{1}{F} \sum_{k \in \mathbb{F}} |W_k V(k)|^2 \rightarrow \frac{1}{F} \sum_{k \in \mathbb{F}} |W_k H(z_k^{-1}) E(k)|^2 && \text{w.p. 1} \end{aligned} \qquad (14\text{-}192)$$

at the rate $O_p(F^{-1/2})$. Hence, it is sufficient to study the following partial sums:

$$S(F) = \sum_{k \in \mathbb{F}} w_k E(k) \quad \text{and} \quad S(F) = \sum_{k \in \mathbb{F}} |w_k E(k)|^2 \qquad (14\text{-}193)$$

where $w_k = W_k H(z_k^{-1})$ is uniformly bounded. From Lemmas 14.3 and 14.27, it follows that $w_k E(k)$ and $|w_k E(k)|^2$ are mixing of order 2. Hence, the partial sums $S(F)/F$ in (14-193) converge in mean square sense at the rate $O_{\text{m.s.}}(F^{-1/2})$ to their expected value (see Section 14.9, version 3 of the law of large numbers). Note that the noise $E(k)$ in (14-193) depends on the number of time domain samples N and, hence, also on the number of frequencies $F = O(N)$. Therefore, it should be denoted more precisely as $E_N(k)$, and to prove the strong convergence of $S(F)/F$, we must also verify that

$$\text{var}(\textstyle\sum_{k=1}^{s} x_r(k) - x_s(k)) = O(r-s) \text{ with } r \geq s \tag{14-194}$$

is satisfied for $x_r(k) = w_k E_r(k)$ and $x_r(k) = |w_k E_r(k)|^2$ (see Section 14.9, version 3 of the law of large numbers). To verify this condition we rearrange the order of the frequencies in (14-193) such that the new added frequencies appear at the end of the sum. In (14-194) we compare terms of the sums (14-193), at the same physical frequencies kf_s/s and NOT at the same DFT line numbers k; otherwise, the comparison makes no sense. This imposes a condition on the number of time domain samples $r \geq s$: r must be chosen such that the physical frequencies kf_s/s, $k = 0, 1, \ldots, s-1$, form a subset of the physical frequencies kf_s/r, $k = 0, 1, \ldots, r-1$. This condition is satisfied for the choice $r = ms$ with $m = 1, 2, 3, \ldots$. It means that we compare time domain experiments where the number of samples N is increased linearly as mN, $m = 1, 2, 3, \ldots$. Note that in a classical time domain analysis the number of samples is increased linearly as $N + m$, $m = 1, 2, 3, \ldots$.

The first partial sum in (14-193) converges strongly to its expected value if

$$\text{var}(\sum_{k \in \mathbf{F}_s} w_k(E_{ms}(mk) - E_s(k))) = O((m-1)s) \tag{14-195}$$

with $m \in \mathbb{N}_0$, $\mathbf{F}_s$ a set of $F_s = O(s)$ DFT frequencies, and

$$\begin{aligned} E_{ms}(mk) &= \frac{1}{\sqrt{ms}} \textstyle\sum_{t=0}^{ms-1} e(t) e^{-2\pi jkt/s} \\ E_s(k) &= \frac{1}{\sqrt{s}} \textstyle\sum_{t=0}^{s-1} e(t) e^{-2\pi jkt/s} \end{aligned} \tag{14-196}$$

Because $\sum_{t=0}^{s-1} e^{-2\pi jkt/s} = 0$ for $k \neq 0$, we can replace $e(t)$ by $e(t) - \mu$ in (14-196) without changing $E_{ms}(mk)$ and $E_s(k)$ for $k \neq 0$. Hence, we may assume in the sequel of the analysis that $e(t)$ has zero mean. If DC ($k = 0$) belongs to the set $\mathbf{F}_s$, then $\mu = \mathcal{E}\{ e(t) \}$ should be zero, otherwise the expected values of $E_{ms}(0)$ and $E_s(0)$ are not zero. Elaborating the variance expression in (14-195) gives

$$\begin{aligned} &\text{var}(\sum_{k \in \mathbf{F}_s} w_k(E_{ms}(mk) - E_s(k))) = \\ &\qquad \sum_{k_1, k_2 \in \mathbf{F}_s} w_{k_1} \bar{w}_{k_2} \mathcal{E}\{ (E_{ms}(mk_1) - E_s(k_1)(\bar{E}_{ms}(mk_2) - \bar{E}_s(k_2))) \end{aligned} \tag{14-197}$$

Because $e(t)$ is an independent random variable with zero mean and variance σ^2, we have

$$\begin{aligned} \mathcal{E}\{\, E_{ms}(mk_1)\bar{E}_{ms}(mk_2)\,\} &= \mathcal{E}\{\, E_s(k_1)\bar{E}_s(k_2)\,\} = \sigma^2\delta(k_1-k_2) \\ \mathcal{E}\{\, E_{ms}(mk_1)\bar{E}_s(k_2)\,\} &= \mathcal{E}\{\, E_s(k_1)\bar{E}_{ms}(mk_2)\,\} = \frac{\sigma^2}{\sqrt{m}}\delta(k_1-k_2) \end{aligned} \tag{14-198}$$

with $\delta(k)$ the Kronecker delta. Using (14-197) and (14-198), we find

$$\operatorname{var}\Big(\sum_{k\in\mathbf{F}} w_k(E_{ms}(mk)-E_s(k))\Big) = 2\sigma^2\frac{(m-1)s}{\sqrt{m}(\sqrt{m}+1)}\frac{1}{s}\sum_{k\in\mathbf{F}_s}|w_k|^2 \le O((m-1)s) \tag{14-199}$$

where the last inequality is due to $|w_k| \le c < \infty$ for any k.

The variance expression for the second partial sum in (14-193) equals

$$\begin{aligned} &\operatorname{var}\Big(\sum_{k\in\mathbf{F}}|w_k|^2(|E_{ms}(mk)|^2-|E_s(k)|^2)\Big) = \\ &\quad\sum_{k_1,k_2\in\mathbf{F}_s}|w_{k_1}|^2|w_{k_2}|^2\mathcal{E}\{\,(|E_{ms}(mk_1)|^2-|E_s(k_1)|^2(|E_{ms}(mk_2)|^2-|E_s(k_2)|^2))\end{aligned} \tag{14-200}$$

Because $e(t)$ is an independent random variable with zero mean, variance σ^2, and uniformly bounded fourth-order moment $\mu_4(t)$, we have

$$\begin{aligned} \mathcal{E}\{\,|E_{ms}(mk_1)|^2|E_{ms}(mk_2)|^2\} &= \sigma^4+\sigma^4\delta(k_1-k_2)+\frac{\kappa_4(ms)}{ms} \\ \mathcal{E}\{\,|E_s(k_1)|^2|E_s(k_2)|^2\} &= \sigma^4+\sigma^4\delta(k_1-k_2)+\frac{\kappa_4(s)}{s} \\ \mathcal{E}\{\,|E_{ms}(mk_1)|^2|E_s(k_2)|^2\} &= \mathcal{E}\{\,|E_{ms}(mk_1)|^2|E_s(k_2)|^2\} = \sigma^4+\frac{\sigma^4}{m}\delta(k_1-k_2)+\frac{\kappa_4(s)}{ms} \end{aligned} \tag{14-201}$$

where $\kappa_4(r) = \sum_{t=0}^{r-1}\mu_4(t)/r - 3\sigma^4$ is an $O(s^0)$. Using (14-200) and (14-201), we find

$$\begin{aligned} \operatorname{var}\Big(\sum_{k\in\mathbf{F}}|w_k|^2(|E_{ms}(mk)|^2-|E_s(k)|^2)\Big) &= \frac{(m-1)s}{m}\frac{2\sigma^2}{s}\sum\nolimits_{k\in\mathbf{F}_s}|w_k|^4 \\ &+\left(\frac{(m-1)s}{m}\kappa_4(s)+\frac{1}{m}\sum\nolimits_{t=s}^{ms-1}\mu_4(t)\right)\frac{1}{s^2}\sum\nolimits_{k_1,k_2\in\mathbf{F}_s}|w_{k_1}|^2|w_{k_2}|^2 \end{aligned} \tag{14-202}$$

Because $\kappa_4(s) = O(s^0)$, $|w_k| \le c < \infty$ is uniformly bounded, and

$$\left|\sum\nolimits_{t=s}^{ms-1}\mu_4(t)\right| \le (\max_t|\mu_4(t)|)((m-1)s)$$

(14-202) is bounded above by

$$\operatorname{var}\Big(\sum_{k\in\mathbf{F}}|w_k|^2(|E_{ms}(mk)|^2-|E_s(k)|^2)\Big) \le O((m-1)s)$$

which concludes the proof for the second partial sum. □

Appendix 14.P: Proof of Theorem 14.29

To prove the theorem it is sufficient to replace $V(k)$ by $H(z_k^{-1})E(k)$ (Lemma 14.23). The rest of the proof follows the lines of Appendix 14.L. Using (14-182) and property 2 of the cumulants (see Section 14.1), the Jth order cumulant of $N^{-1/2}S(N)$ becomes

$$C_J(N) = \frac{C_J}{N^{J-1}} \sum_{k_1, k_2, \ldots, k_{J-1} = 0}^{N-1} Y(k_1)\bar{Y}(k_2)\ldots Y(k_{J-1})Y(N - \textstyle\sum_{i=1}^{J-1} k_i) \tag{14-203}$$

with $Y(k) = \bar{H}(z_k^{-1})X(k)$. The right-hand side of (14-203) can be written as $J-1$ consecutive circular DFT convolutions of $Y(k)$ with itself

$$C_J(N) = \frac{C_J}{N^{(J-1)/2}} Y_J(N) \tag{14-204}$$

with $Y_J(k) = Y(k)*(Y(k)*(\ldots*Y(k)))$ and $Y(k)*Z(k) = N^{-1/2}\sum_{r=0}^{N-1} Y(r)Z(k-r)$. Using the property that the inverse discrete Fourier transform (IDFT) of a circular convolution of DFT spectra equals the product of the corresponding time signals, we can write $Y_J(k)$ as

$$Y_J(k) = \text{DFT}(\text{IDFT}(Y_J(k))) = \text{DFT}(y^J(t)) = \frac{1}{\sqrt{N}}\sum_{t=0}^{N-1} y^J(t) e^{-\frac{2\pi}{N}jkt} \tag{14-205}$$

Hence, $Y_J(N) = N^{-1/2}\sum_{t=0}^{N-1} y^J(t)$, and

$$C_J(N) = \frac{C_J}{N^{J/2-1}} \frac{1}{N} \sum_{t=0}^{N-1} y^J(t) \tag{14-206}$$

Applying the bounded-input, bounded-output property of stable linear systems (Kailath, 1980) to $\bar{Y}(k) = H(z_k^{-1})\bar{X}(k)$, with $\text{IDFT}(\bar{Z}(k)) = z(-t)$ $(Z = X, Y,\ z = x, y)$, shows that $\max_t |y(-t)| \le c_1 < \infty$ if $\max_t |x(-t)| \le c < \infty$ where c and c_1 are independent of N $(\max_t |z(-t)| = \max_t |z(t)|,\ z = x, y)$. Similarly, $N^{-1}\sum_{t=0}^{N-1} y^2(t) = O(N^0)$ since $N^{-1}\sum_{t=0}^{N-1} x^2(t) = O(N^0)$. Therefore, (14-206) becomes for $J = 2$

$$C_2(N) = \text{var}(N^{-1/2}S(N)) = \frac{C_2}{N^0} O(N^0) = O(N^0) \tag{14-207}$$

while for $J > 2$ (14-206) can be bounded above by $(N^{-1}\left|\sum_{t=0}^{N-1} y^J(t)\right| \le \max_t |y(t)|^J)$

$$|C_J(N)| \le \frac{C_J}{N^{J/2-1}} c_1^J \le O(N^{-J/2+1}) \tag{14-208}$$

It follows that the cumulants of order $J = 3, 4, \ldots$ are asymptotically zero, while $C_2(N) = O(N^0)$. According to the Fréchet-Shohat Lemma 14.11, $N^{-1/2}S(N)$ is asymptotically normally distributed. The convergence rate to the normal distribution function is established as in Appendix 14.H. □

Appendix 14.Q: Proof of Corollary 14.30

The proof follows the same lines as for Theorem 14.29. The only difference lies in the upper bound (14-208). Because (14-207) remains valid (we consider only signals with finite power), the signal can reach its peak value $O(\sqrt{\ln N})$ at most $O(N/\ln N)$ times, while the remaining $N - O(N/\ln N)$ samples have the value $O(N^0)$. Therefore, (14-206) with $J > 2$, is bounded above by

$$|C_J(N)| \le \frac{C_J}{N^{J/2-1}} O((\ln N)^{J/2-1}) \le O\left(\left(\frac{\ln N}{N}\right)^{J/2-1}\right) \tag{14-209}$$

which concludes the proof. □

Appendix 14.R: Proof of Lemma 14.31

First we prove the theorem for $X(k) = E_2(k)$ and $H_1(z_k^{-1}) = 1$. The generalization to the colored case follows directly from the uniformly boundedness of $H_1(z_k^{-1})$ and $H_2(z_k^{-1})$. Applying Lemma 14.35 with $\nu_{i1} = E_2(k)$ and $\nu_{i2} = \bar{E}_1(k)$ to the Jth order joint cumulant of $E_2(k)\bar{E}_1(k)$ gives

$$\begin{aligned}&\operatorname{cum}(E_2(k_1)\bar{E}_1(k_1), E_2(k_2)\bar{E}_1(k_2), \ldots, E_2(k_J)\bar{E}_1(k_J)) = \\ &\qquad \sum_{\mathbb{P}} \operatorname{cum}(E(k_{ij}) \in \mathbb{P}_1)\operatorname{cum}(E(k_{ij}) \in \mathbb{P}_2)\cdots\operatorname{cum}(E(k_{ij}) \in \mathbb{P}_K)\end{aligned} \tag{14-210}$$

where $E(k)$ equals $E_2(k)$ and/or $\bar{E}_1(k)$ and where the summation extends over all indecomposable partitions $\mathbb{P} = \mathbb{P}_1 \cup \mathbb{P}_2 \cup \cdots \cup \mathbb{P}_K$ of the table (14-137) with $J_1 = J_2 = \cdots = J_J = 2$. We study the mixing condition

$$\max_{k_J} \sum_{k_1, \ldots, k_{J-1} = 0}^{N-1} |\operatorname{cum}(E(k_{ij}) \in \mathbb{P}_1)\operatorname{cum}(E(k_{ij}) \in \mathbb{P}_2)\cdots\operatorname{cum}(E(k_{ij}) \in \mathbb{P}_K)| \tag{14-211}$$

for each term in the summation (14-210). As $\bar{E}_1(k)$ and $E_2(k)$ are mutually independent random variables, the partitions in (14-210) are limited to those where all $E(k_{ij}) \in \mathbb{P}_r$ are equal to $E_2(k)$ or to $\bar{E}_1(k)$. Therefore, we can apply formula (14-182) to each $\operatorname{cum}(E(k_{ij}) \in \mathbb{P}_r)$ in (14-211), which gives

$$\max_{k_J} \sum_{k_1, \ldots, k_{J-1} = 0}^{N-1} \left| \prod_{r=1}^{K} \frac{C_{J_r}}{N^{J_r/2-1}} \right| \tag{14-212}$$

with $\sum_{r=1}^{K} J_r = 2J$ (each partition contains all elements of the set) and where K constraints of the form

$$\left(\sum k_{ij}\right) \bmod N = 0, \; E(k_{ij}) \in \mathbb{P}_r, \; r = 1, 2, \ldots, K \tag{14-213}$$

should be satisfied. Because the partition $\mathbb{P} = \mathbb{P}_1 \cup \mathbb{P}_2 \cup \cdots \cup \mathbb{P}_K$ is indecomposable, there are exactly $K-1$ independent constraints in (14-213) (see Lemma 2.3.1, p. 20, Brillinger, 1981), so that (14-212) can be bounded above by

$$N^{(J-1)-(K-1)} \frac{\prod_{r=1}^{K} |C_{J_r}|}{N^{J-K}} = O(N^0) \tag{14-214}$$

This concludes the proof for the (colored) white noise case.

The result (14-214) is also valid when $X(k)$ is the DFT spectrum of an integer number of periods of normalized periodic noise or a normalized random multisine. To prove this statement, it is sufficient to note that

$$\mathrm{cum}(X(k_1), X(k_2), \ldots, X(k_J)) = C_J \delta(k_2 - k_1) \delta(k_3 - k_1) \cdots \delta(k_J - k_1)$$

by construction of these periodic signals (see Definitions 3.2 and 3.4). □

Appendix 14.S: Proof of Theorem 14.32

The proof follows the lines of Appendix 14.O.

14.S.1 First Partial Sum of (14-135). We distinguish two cases: (i) $X(k)$ is the DFT spectrum of filtered white noise, $x(t) = H_1(q)e_1(t)$, where $H_1(z^{-1})$ is stable and $e_1(t)$ is independently distributed noise with mean $\mu_1 < \infty$ and variance $\sigma_1^2 < \infty$, and (ii) $X(k)$ is the DFT spectrum of an integer number of periods of a normalized random multisine or normalized periodic noise.

Applying Lemma 14.23 to the first partial sum of (14-135) gives, for case (i),

$$\frac{1}{F} \sum_{k \in \mathbf{F}} W_k X(k) \bar{V}(k) \to \frac{1}{F} \sum_{k \in \mathbb{F}} w_k E_1(k) \bar{E}(k) \qquad \text{w.p. 1}$$

at the rate $O_p(F^{-1/2})$, where $w_k = W_k H_1(z_k^{-1}) \bar{H}(z_k^{-1})$ is uniformly bounded and where, at the same physical frequencies kf_s/N, w_k is independent of the number of time domain samples N. Hence, it is sufficient to study

$$S(F) = \sum_{k \in \mathbb{F}} w_k E_1(k) \bar{E}(k)$$

Using formulas (14-198) and the fact that $E_1(k)$ and $E(k)$ are independent, we find

$$\begin{aligned} \mathrm{var}\Big(\sum_{k \in \mathbb{F}} w_k (E_{1ms}(mk) \bar{E}_{ms}(mk) - E_{1s}(k) \bar{E}_s(k))\Big) &= 2\sigma_1^2 \sigma^2 \frac{(m-1)s}{m} \frac{1}{s} \sum_{k \in \mathbb{F}_s} |w_k|^2 \\ &\leq O((m-1)s) \end{aligned}$$

where the last inequality is due to $|w_k| \leq c < \infty$ for any k.

Applying Lemma 14.23 to the first partial sum of (14-135) gives, for case (ii),

$$\frac{1}{F}\sum_{k\in\mathbb{F}} W_k X(k)\bar{V}(k) \to \frac{1}{F}\sum_{k\in\mathbb{F}} w_k X(k)\bar{E}(k) \qquad \text{w.p. 1}$$

at the rate $O_{\mathrm{p}}(F^{-1/2})$, where $w_k = W_k\bar{H}(z_k^{-1})$ is uniformly bounded and where, at the same physical frequencies kf_s/N, w_k and $X(k)$ are independent of the number of time domain samples N. Hence, we must study

$$\text{var}\Big(\sum_{k\in\mathbb{F}_s} w_k X(k)(\bar{E}_{ms}(mk) - \bar{E}_s(k))\Big) \tag{14-215}$$

Because $X(k)$ is independent of $E(k)$, formula (14-199) of Appendix 14.O remains valid for (14-215), if $|w_k|^2$ is replaced by $|w_k|^2\mathcal{E}\{|X(k)|^2\}$.

14.S.2 Second Partial Sum of (14-135). As the case where $X(k)$ is the DFT spectrum of filtered white noise is already covered by Theorem 14.28, it is sufficient to study the case where $X(k)$ is the DFT spectrum of an integer number of periods of a periodic signal. For normalized random multisines, $|X(k)|^2$ is a uniformly bounded nonrandom number, while for normalized periodic noise, $|X(k)|^2$ is a random variable with uniformly bounded fourth-order moments. In both cases, at the same physical frequencies kf_s/N, $X(k)$ is independent of the number of time domain samples N. Hence, $S(F)/F$ obeys the strong law of large numbers (see Section 14.9, version 3 of the law of large numbers). □

Appendix 14.T: Proof of Theorem 14.33

Applying Lemma 14.23 to the first two partial sums of (14-136) gives

$$\begin{aligned} S(F)/F &\to \frac{1}{F}\sum_{k\in\mathbb{F}} W_k Y(k)\overline{H(z_k^{-1})E(k)} && \text{w.p. 1} \\ S(F)/F &\to \frac{1}{F}\sum_{k\in\mathbb{F}} |W_k H(z_k^{-1})E(k)|^2 && \text{w.p. 1} \end{aligned} \tag{14-216}$$

at the rate $O_{\mathrm{p}}(F^{-1/2})$. $Y(k) = H_1(z_k^{-1})E_1(k)$ for filtered iid noise, and $Y(k) = X(k)$ for the periodic signals. Because W_k, $H_1(z_k^{-1})$, and $H(z_k^{-1})$ are uniformly bounded, $W_kY(k)H(z_k^{-1})E(k)$ and $|W_kH(z_k^{-1})E(k)|^2$ are mixing of order infinity (proof: apply Lemmas 14.3, 14.26, and 14.31). Hence, according to version 4 of the central limit theorem (see Section 14.10), the sums in (14-216) are asymptotically normally distributed (convergence in law at the rate $O(F^{-1/2})$).

For the third partial sum of (14-136), we need only handle the case where $X(k)$ is the DFT spectrum of the periodic signals (the filtered iid case is already covered by the second partial sum of (14-136)). Because $|X(k)|^2$ is, by construction, independent over k (see Definitions 3.2 and 3.4), and W_k, $|X(k)|^2$ have, by assumption, existing moments of any order, $W_k|X(k)|^2$ is mixing of order infinity. Therefore, $W_k|X(k)|^2$ is asymptotically normally distributed at the rate $O(F^{-1/2})$ (see Section 14.10, version 4 of the central limit theorem). For a normalized random multisine, $|X(k)|^2$ is a nonrandom number and the normal distribution is degenerate. □

15

Properties of Least Squares Estimators with Deterministic Weighting

Abstract: This chapter studies the asymptotic stochastic properties (strong convergence, strong consistency, convergence rate, asymptotic bias, and asymptotic normality) of nonlinear least squares estimators with a deterministic weighting. The presented theory is applicable to a large class of estimators such as the quadratic prediction error methods, the Gaussian maximum likelihood estimators, and the total least squares–based methods. Readers who are unfamiliar with the analysis of the stochastic properties of estimators should first read Sections 14.11 to 14.13.

15.1 INTRODUCTION

In this chapter we consider the identification of a parametric plant and/or noise model $M(\theta, z_0, n_z)$ through the minimization of a weighted nonlinear least squares cost function

$$V_N(\theta, z) = \frac{1}{N} z^T W_N(\theta) z \tag{15-1}$$

with $W_N(\theta) \in \mathbb{R}^{N \times N}$ a deterministic positive semidefinite weighting matrix, $z \in \mathbb{R}^N$ the noisy measurements, $z = z_0 + n_z$, and $\theta \in \mathbb{R}^{n_\theta}$ the plant and/or noise model parameters with n_θ independent of N. Because the nonsymmetric part of $W_N(\theta)$ does not contribute to the quadratic form (15-1) (see Exercise 13.7), we can assume without any loss of generality that the weighting matrix $W_N(\theta)$ is symmetric. Note that all the elements of $W_N(\theta)$ may change as N increases. Because the true (unknown) observations z_0 can be a random variable, the expected values are taken everywhere w.r.t. the disturbing noise n_z and the true observations z_0.

The analysis of the stochastic properties of the minimizer(s) of (15-1) requires a closed and bounded (= compact) set of parameters, where the cost function (15-1) and/or its higher order derivatives exist and are finite. Such a regular compact set is constructed as follows. Let $\Theta \subset \mathbb{R}^{n_\theta}$, with $\dim(\Theta) = n_\theta$, be a compact parameter set. Define $\Theta_s \subset \Theta$ as the singular set of parameter values for which the cost function (15-1) does not exist or is infinite. Usually,

System Identification: A Frequency Domain Approach. Rik Pintelon and Johan Schoukens
Copyright © 2001 Institute of Electrical and Electronics Engineers. ISBN: 0-780-36000-1

the topological dimension of this singular set is smaller than n_θ. The regular set Θ_r are the parameters in Θ that are not within an ε-distance of the singular set Θ_s

$$\Theta_r = \Theta \backslash \{\theta \in \Theta | \|\theta - \theta_s\| < \varepsilon, \theta_s \in \Theta_s\} \tag{15-2}$$

Θ_r is compact (closed and bounded) by construction. Using the same reasoning, a regular compact set is constructed where the (higher order) derivatives of the cost function exist and are finite. Note that for the maximum likelihood estimation of ARMAX models the compactness assumption of the parameter space can be avoided (Hannan and Deistler, 1988).

Following the same lines as in Section 14.13, the asymptotic ($N \to \infty$) properties (strong convergence, strong consistency, convergence rate, asymptotic bias, and asymptotic normality) of the (set of) minimizer(s)

$$\hat{\theta}(z) = \arg \min_{\theta \in \Theta_r} V_N(\theta, z) \tag{15-3}$$

will be analyzed. Replacing $z \in \mathbb{C}^N$ by $z_{\mathrm{re}} \in \mathbb{R}^{2N}$ in (15-1) and/or $\theta \in \mathbb{C}^{n_\theta}$ by $\theta_{\mathrm{re}} \in \mathbb{R}^{2n_\theta}$ (see Section 13.8 for the definition of $(\ )_{\mathrm{re}}$), it follows directly that the results also apply to complex measurements and/or complex parameters $\theta \in \mathbb{C}^{n_\theta}$. The chapter ends with an overview of the asymptotic properties of $\hat{\theta}(z)$.

15.2 STRONG CONVERGENCE

The first step in the analysis consists of detecting the stochastic sum(s) w in the cost function (15-1) that averages the noise. It can easily be seen that there is only one such sum, namely the cost function itself. Following the notations of Section 14.13, we have

$$w(\theta, z_0, n_z) = V_N(\theta, z) \text{ and } \mu_w(\theta, z_0) = \mathcal{E}\{V_N(\theta, z)\} \tag{15-4}$$

so that

$$\tilde{\theta}(z_0) = \arg \min_{\theta \in \Theta_r} V_N(\theta) \tag{15-5}$$

with $V_N(\theta) = \mathcal{E}\{V_N(\theta, z)\}$.

In the second step (see Section 15.2.1), the uniform convergence (w.r.t. θ) of the stochastic sum w toward its expected value μ_w is established

$$\underset{N \to \infty}{\text{a.s.lim}}(V_N(\theta, z) - V_N(\theta)) = 0 \text{ or } \underset{N \to \infty}{\text{a.s.lim}} V_N(\theta, z) = \lim_{N \to \infty} V_N(\theta) = V_*(\theta) \tag{15-6}$$

This requires some assumptions concerning the true observations z_0, the disturbing noise n_z, the weighting matrix $W_N(\theta)$, and the strategy of adding the measurements. The convergence should be uniform w.r.t. the model parameters $\theta \in \Theta_r$ to ensure that the convergence of the cost functions implies the convergence of the minimizers. Figure 15-1 shows a counterexample where the cost function $V_N(\theta)$ converges nonuniformly to its limit value $V_*(\theta)$. It can be seen that the global minimum of the sequence $V_N(\theta)$ does not converge to the global minimum of $V_*(\theta)$.

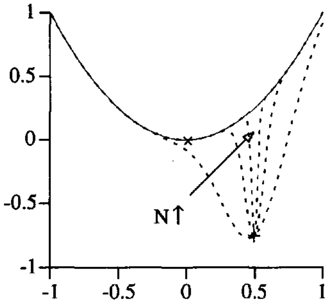


Figure 15-1. Although $V_N(\theta, z)$ (dashed lines) converges nonuniformly to $V_*(\theta)$ (solid line), their global minimizers (X) and (+) differ.

In a third step (see Section 15.2.2) the strong convergence of the minimizer(s) is established from the strong uniform convergence of the cost function

$$\underset{N\to\infty}{\text{a.s.lim}}(\hat{\theta}(z) - \tilde{\theta}(z_0)) = 0 \text{ or } \underset{N\to\infty}{\text{a.s.lim}}\,\hat{\theta}(z) = \lim_{N\to\infty}\tilde{\theta}(z_0) = \theta_* \tag{15-7}$$

It requires that adding measurements to z $(N \to \infty)$ increases the knowledge about the model parameters θ such that θ is uniquely identifiable. If this is the case, then the data are said to be *persistently exciting*. The weakest assumption that satisfies this condition is that the asymptotic cost function $V_*(\theta)$ has a unique global minimum θ_*. If this assumption is not fulfilled, then the uniform convergence of the cost functions does not imply the convergence of their minimizer(s). The following counterexample shows this. Consider, for example, the cost functions $V_{1N}(\theta) = 1 - N^{-1}\sin\theta$ and $V_{2N}(\theta) = 1 - N^{-1}\cos(\theta)$ with respective minimizers $\hat{\theta}_1(N) = \pi/2 + 2k\pi$ and $\hat{\theta}_2(N) = 2k\pi$, $k \in \mathbb{Z}$. Although $V_{1N}(\theta)$ converges uniformly in θ to $V_{2N}(\theta)$, $\hat{\theta}_1(N)$ does not converge to $\hat{\theta}_2(N)$: $\hat{\theta}_1(\infty) \neq \hat{\theta}_2(\infty)$. The problem with this counterexample is that all θ-values minimize the limit cost function $V_*(\theta) = 1$.

15.2.1 Strong Convergence of the Cost Function

Assumption 15.1 (Mixing Condition of Order P): The true observations $z_0 \in \mathbb{R}^N$ are disturbed by zero mean additive noise $z = z_0 + n_z$. The noise n_z is stochastically independent of z_0. Both n_z and z are mixing of order P.

Assumption 15.2 (Constraints on the Cost Function): (a) The weighting matrix $W_N(\theta) \in \mathbb{R}^{N \times N}$ in (15-1) is a symmetric positive semidefinite matrix, satisfying $\|W_N(\theta)\|_1 \leq c < \infty$, with c an N-independent constant, for all N, ∞ included, and all $\theta \in \Theta_r$. $W_N(\theta)$ is a continuous matrix function of θ in the compact set Θ_r. (b) There is an N_0 such that for any $r \geq s \geq N_0$, $\|W_{r[1:s,\,1:s]}(\theta) - W_s(\theta)\|_1^2 = O((r-s)/r)$ in Θ_r.

Note that Assumption 15.1 makes it possible to handle, simultaneously, the cases z_0 random and/or z_0 deterministic. Condition (b) in Assumption 15.2 limits the variation of the elements of $W_N(\theta)$ as N increases to infinity. This is necessary to ensure the strong convergence of the cost function (see proof of Lemma 15.3). If (b) is not satisfied, then only mean square convergence of the cost function can be shown. All lemmas and theorems of this chapter remain valid except that the strong convergence (w.p. 1) must be replaced by weak convergence (in prob.).

Lemma 15.3 (Strong Convergence of the Cost Function): Under Assumptions 15.1 $(P = 4)$ and 15.2 the cost function $V_N(\theta, z)$ converges uniformly w.p. 1 to its expected value

$V_N(\theta)$ in the compact set Θ_r. The uniform mean square convergence rate in Θ_r is $O_{\text{m.s.}}(N^{-1/2})$: $V_N(\theta, z) = V_N(\theta) + O_{\text{m.s.}}(N^{-1/2})$.

Proof. See Appendix 15.A. □

Lemma 15.3 does not guarantee that the limit cost function $V_*(\theta) = \lim_{N \to \infty} V_N(\theta)$ exists. Because $V_N(\theta)$ depends on $W_N(\theta)$ and z, the existence of $V_*(\theta)$ imposes some conditions on $W_N(\theta)$ and z that should be verified for each particular choice of the weighting $W_N(\theta)$ and for every experiment (strategy of adding measurements to z). Therefore, we make the following assumption.

Assumption 15.4 (Constraint on the Experiment): The expected value of the cost function $V_N(\theta)$ converges uniformly to the limit cost function $V_*(\theta)$ in the compact set Θ_r.

15.2.2 Strong Convergence of the Minimizer

Assumption 15.5 (Persistence of Excitation): There exists an N_0 such that for any $N \geq N_0$, ∞ included, the expected value of the cost function $V_N(\theta)$ has a unique global minimum $\tilde{\theta}(z_0)$, which is an interior point of Θ_r.

Theorem 15.6 (Strong Convergence of the Minimizer): Under Assumptions 15.1 $(P = 4)$, 15.2, and 15.5, the minimizer(s) $\hat{\theta}(z)$ converge(s) strongly to $\tilde{\theta}(z_0)$: $\text{a.s.}\lim_{N \to \infty}(\hat{\theta}(z) - \tilde{\theta}(z_0)) = 0$.

Proof. See Appendix 15.B. □

Theorem 15.6 does not guarantee that $\tilde{\theta}(z_0)$ converges to some limit value θ_*. Assumptions 15.4 and 15.7 ensure the existence of this limit value.

Assumption 15.7 (Persistence of Excitation): The asymptotic cost function $V_*(\theta)$ has a unique global minimum θ_*, which is an interior point of Θ_r.

If $V_N(\theta)$ and/or $V_*(\theta)$ are not convex, then in the presence of model errors it may happen that $V_N(\theta)$ and/or $V_*(\theta)$ have more than one global minimum. An example of this is given in Kabaila (1983) for the identification of particular parametric noise models (MA processes). To handle these cases we restrict the compact set Θ_r in Assumptions 15.5 and 15.7 such that $V_N(\theta)$ and/or $V_*(\theta)$ contain a unique global minimum in Θ_r.

Theorem 15.8 (Strong Convergence of the Minimizer): Under Assumptions 15.1 $(P = 4)$, 15.2, 15.4, and 15.7, $\tilde{\theta}(z_0)$ converges to θ_* and $\hat{\theta}(z)$ converges strongly to θ_*: $\lim_{N \to \infty} \tilde{\theta}(z_0) = \theta_*$ and $\text{a.s.}\lim_{N \to \infty} \hat{\theta}(z) = \theta_*$.

Proof. Note that $V_*(\theta)$ is a continuous function in Θ_r because it is the deterministic limit of a uniformly convergent sequence of continuous functions $V_N(\theta)$ in Θ_r (Theorem 2.1 of Henrici, 1974). The proof of the two limits then follows the same lines as for Theorem 15.6. □

Note that Theorems 15.6 and 15.8 do not require the existence of the derivative(s) of the cost function and are valid in the presence of model errors (the true model cannot be represented by $M(\theta, z_0, n_z)$).

15.3 STRONG CONSISTENCY

Consistency can be proved only if the true model belongs to the considered model set. Therefore, the following assumption is made.

Assumption 15.9 (True Model Belongs to Model Set): There is a $\theta_0 \in \Theta_r$ such that $M(\theta_0, z_0, n_z)$ represents the true model.

Using the results of Section 15.2, it follows directly that the estimates $\hat{\theta}(z)$ are strongly consistent if either $\theta_* = \theta_0$ (weakest assumption) or $\tilde{\theta}(z_0) = \theta_0$ for any $N \geq N_0$ (stronger assumption). This imposes some conditions on the expected value of the cost function, which can be written as

$$V_N(\theta) = \frac{1}{N}\mathcal{E}\{z_0^T W_N(\theta) z_0\} + \frac{1}{N}\text{trace}(W_N(\theta) C_{n_z}) \tag{15-8}$$

with C_{n_z} the covariance matrix of the disturbing noise n_z (see Exercise 15.2). The following theorems are in order of reduced conditions on $V_N(\theta)$.

Assumption 15.10 (Consistency Condition on the Cost Function): There exists an N_0 such that for any $N \geq N_0$, ∞ included, $\mathcal{E}\{z_0^T W_N(\theta) z_0\}$ is minimal in the true parameter values $\theta_0 \in \Theta_r$ and $\text{trace}(W_N(\theta) C_{n_z})$ is a θ-independent constant for any $\theta \in \Theta_r$.

Theorem 15.11 (Strong Consistency): Under the assumptions of Theorem 15.6 and Assumptions 15.9 and 15.10, the estimate $\hat{\theta}(z)$ is strongly consistent: $\text{a.s.}\lim_{N \to \infty} \hat{\theta}(z) = \theta_0$.

Proof. It follows directly from Theorem 15.6 and Assumptions 15.9 and 15.10. □

Assumption 15.12 (Consistency Condition on the Cost Function): There exists an N_0 such that for any $N \geq N_0$, ∞ included, the expected value of the cost function $V_N(\theta)$ is minimal in the true parameter values $\theta_0 \in \Theta_r$.

Theorem 15.13 (Strong Consistency): Under the assumptions of Theorem 15.6 and Assumptions 15.9 and 15.12, the estimate $\hat{\theta}(z)$ is strongly consistent: $\text{a.s.}\lim_{N \to \infty} \hat{\theta}(z) = \theta_0$.

Proof. It follows directly from Theorem 15.6 and Assumptions 15.9 and 15.12. □

Assumption 15.14 (Consistency Condition on the Cost Function): The asymptotic cost function $V_*(\theta)$ is minimal in the true parameter values $\theta_0 \in \Theta_r$.

Theorem 15.15 (Strong Consistency): Under the assumptions of Theorem 15.8 and Assumptions 15.9 and 15.14, the estimate $\hat{\theta}(z)$ is strongly consistent: $\text{a.s.}\lim_{N \to \infty} \hat{\theta}(z) = \theta_0$.

Proof. It follows directly from Theorem 15.8 and Assumptions 15.9 and 15.14. □

Although Assumptions 15.10 and 15.12 are stronger than Assumption 15.14, they are satisfied very often in practice. Assumption 15.10 often applies when a nonparametric noise model is identified (see, for example, Exercise 15.3 and Chapter 7).

15.4 CONVERGENCE RATE

Sections 15.2 and 15.3 study the conditions under which the estimate $\hat{\theta}(z)$ converges. This section studies how fast the estimate $\hat{\theta}(z)$ converges toward its limit value.

In a first step, the convergence rate of $\hat{\theta}(z)$ to $\tilde{\theta}(z_0)$ is analyzed. This is already sufficient for the strongly consistent estimators ($\tilde{\theta}(z_0) = \theta_0$) of Theorems 15.11 and 15.13 and the strongly converging estimator ($\tilde{\theta}(z_0) \neq \theta_0$) of Theorem 15.6. The key idea of the analysis consists of applying the mean value theorem (Kaplan, 1993) to the derivative of the cost function $V_N'(\theta, z)$ at the points $\hat{\theta}(z)$ and $\tilde{\theta}(z_0)$

$$V_N'(\hat{\theta}(z), z) = V_N'(\tilde{\theta}(z_0), z) + (\hat{\theta}(z) - \tilde{\theta}(z_0))^T V_N''(\overset{\frown}{\theta}, z) \tag{15-9}$$

where $\overset{\frown}{\theta}$ is a point on the straight line connecting $\hat{\theta}(z)$ to $\tilde{\theta}(z_0)$

$$\overset{\frown}{\theta} = t\hat{\theta}(z) + (1-t)\tilde{\theta}(z_0) \text{ with } t \in [0, 1] \tag{15-10}$$

Taking into account that $V_N'(\hat{\theta}(z), z) = 0$ ($\hat{\theta}(z)$ is the minimizer of $V'_N(\theta, z)$), an expression for $\hat{\theta}(z) - \tilde{\theta}(z_0)$ is found ($V_N''(\overset{\frown}{\theta}, z)$ is symmetric)

$$\hat{\theta}(z) - \tilde{\theta}(z_0) = -V_N''^{-1}(\overset{\frown}{\theta}, z) V_N'^{T}(\tilde{\theta}(z_0), z) \tag{15-11}$$

From (15-11) it follows that the convergence rate of $\hat{\theta}(z)$ to $\tilde{\theta}(z_0)$ is determined by the convergence rates of the first- and second-order derivatives of the cost function. Therefore, in Section 15.4.1 we will first analyze under which conditions these derivatives converge to their expected values. Next, in Section 15.4.2 the convergence rate of the minimizer will be established from the convergence rate of the derivatives of the cost function.

In a second step, the convergence of $\tilde{\theta}(z_0)$ to θ_* is analyzed (see Section 15.4.3). This second step is necessary for the strongly consistent estimator of Theorem 15.15 ($\theta_* = \theta_0$) and the strongly converging estimator of Theorem 15.8 ($\theta_* \neq \theta_0$). Following the same lines as in the first step, we find

$$\tilde{\theta}(z_0) - \theta_* = -V_N''^{-1}(\overset{\frown}{\theta_*}) V_N'^{T}(\theta_*) \tag{15-12}$$

with $\overset{\frown}{\theta_*}$ a point on the straight line connecting $\tilde{\theta}(z_0)$ to θ_*

$$\overset{\frown}{\theta_*} = t\tilde{\theta}(z_0) + (1-t)\theta_* \text{ with } t \in [0, 1] \tag{15-13}$$

From (15-12), it follows that the convergence rate of $\tilde{\theta}(z_0)$ to θ_* is determined by the deterministic convergence rates of the first- and second-order derivatives of the expected value of the cost function.

15.4.1 Convergence of the Derivatives of the Cost Function

The proof of the mean square converge of the derivatives of the cost function follows the same lines as in Section 15.2.1. Therefore, suitable assumptions concerning the derivatives of $W_N(\theta)$ w.r.t. θ should be made.

Assumption 15.16 (Constraints on First- and Second-Order Derivatives Cost Function): The weighting $W_N(\theta)$ has continuous first- and second-order derivatives w.r.t. θ with bounded 1-norm

$$\left\|\frac{\partial W_N(\theta)}{\partial \theta_{[i]}}\right\|_1 \leq c_1 < \infty, \quad \left\|\frac{\partial^2 W_N(\theta)}{\partial \theta_{[i]} \partial \theta_{[j]}}\right\|_1 \leq c_2 < \infty, \quad i, j = 1, 2, \ldots, n_\theta$$

for $N = 1, 2, \ldots, \infty$ and for any $\theta \in \Theta_{\mathrm{r}}$. c_1, c_2 are N-independent constants.

Lemma 15.17 (Convergence of the Derivatives of the Cost Function): Under Assumptions 15.1 ($P = 4$) and 15.16, the derivatives of the cost function $V_N'(\theta, z)$ and $V_N''(\theta, z)$ converge uniformly in mean square to their expected values $V_N'(\theta)$ and $V_N''(\theta)$ in the compact set Θ_{r}. The uniform mean square convergence rate in Θ_{r} is $O_{\mathrm{m.s.}}(N^{-1/2})$: $V_N'(\theta, z) = V_N'(\theta) + O_{\mathrm{m.s.}}(N^{-1/2})$ and $V_N''(\theta, z) = V_N''(\theta) + O_{\mathrm{m.s.}}(N^{-1/2})$.

Proof. Similar to Lemma 15.3. □

Lemma 15.17 does not guarantee that the Hessian (second-order derivative) of the expected value of the cost function is regular. This is, however, necessary to ensure the existence of the matrix inverse in (15-11) and (15-12). From (15-10) and the convergence of $\hat{\theta}(z)$ to $\tilde{\theta}(z_0)$, it follows that $\widehat{\theta}$ converges to $\tilde{\theta}(z_0)$. Hence, it is sufficient to assume that the Hessian of the expected value of the cost function is regular at $\tilde{\theta}(z_0)$. This assumption imposes some conditions on the data set z, and, therefore, it is also a persistence-of-excitation condition that is stronger than Assumption 15.5.

Assumption 15.18 (Persistence of Excitation): There exists an N_0 such that for any $N \geq N_0$, ∞ included, the Hessian of the expected value of the cost function is regular at the unique global minimizer $\tilde{\theta}(z_0)$, which is an interior point of Θ_{r}: $c_1 I_{n_\theta} \leq V_N''(\tilde{\theta}(z_0)) \leq c_2 I_{n_\theta}$, where $0 < c_1 \leq c_2 < \infty$ and c_1, c_2 are N-independent constants.

15.4.2 Convergence Rate of $\hat{\theta}(z)$ to $\tilde{\theta}(z_0)$

Theorem 15.19 (Convergence Rate of $\hat{\theta}(z)$ to $\tilde{\theta}(z_0)$): Under Assumptions 15.1 ($P = 4$), 15.2(a), 15.16, and 15.18 the convergence rate in probability of $\hat{\theta}(z)$ equals $O_{\mathrm{p}}(N^{-1/2})$: $\hat{\theta}(z) - \tilde{\theta}(z_0) = O_{\mathrm{p}}(N^{-1/2})$.

Proof. See Appendix 15.D. □

Note that Assumption 15.18 is essential for the convergence rate $O_{\mathrm{p}}(N^{-1/2})$. If the Hessian $V_N''(\tilde{\theta}(z_0))$ is not of full rank, then the convergence rate will decrease (see Exercise 15.4). Using the convergence rate of the minimizer $\hat{\theta}(z)$ and Assumption 15.20, we can strengthen Theorem 15.19.

Assumption 15.20 (Constraint on Third-Order Derivative Cost Function): The weighting $W_N(\theta)$ has continuous third-order derivatives w.r.t. θ with bounded 2-norm

$$\left\| \frac{\partial^3 W_N(\theta)}{\partial\theta_{[i]}\partial\theta_{[j]}\partial\theta_{[k]}} \right\|_2 \leq c_3 < \infty,\ i, j, k = 1, 2, \ldots, n_\theta$$

for $N = 1, 2, \ldots, \infty$ and for any $\theta \in \Theta_\mathrm{r}$.

Theorem 15.21 (Improved Convergence Rate of $\hat{\theta}(z)$ to $\tilde{\theta}(z_0)$): Under Assumptions 15.1 $(P = 4)$, 15.2 (a), 15.16, 15.18, and 15.20 the minimizer $\hat{\theta}(z)$ can be written as

$$\begin{aligned} \hat{\theta}(z) &= \tilde{\theta}(z_0) + \delta_\theta(z) + b_\theta(z) \\ \delta_\theta(z) &= -V_N''^{-1}(\tilde{\theta}(z_0))V_N'^T(\tilde{\theta}(z_0), z) \end{aligned} \tag{15-14}$$

where $\mathcal{E}\{\delta_\theta(z)\} = 0$, $\delta_\theta(z) = O_\mathrm{p}(N^{-1/2})$, and $b_\theta(z) = O_\mathrm{p}(N^{-1})$.

Proof. See Appendix 15.E. □

15.4.3 Convergence Rate of $\tilde{\theta}(z_0)$ to θ_*

Under Assumption 15.18, it follows from (15-12) that the convergence rate of $\tilde{\theta}(z_0)$ to θ_* is entirely determined by the deterministic convergence rate of $V_N'(\theta)$ to $V_*'(\theta)$. The latter depends on the way new data are added to the cost function and should be calculated for each particular weighting $W_N(\theta)$ and for every strategy of adding measurements to z. This is summarized in the following assumptions.

Assumption 15.22 (Constraint on the Experiment): The first- and second-order derivatives of the expected value of the cost function, $V_N'(\theta)$ and $V_N''(\theta)$, converge uniformly to their limit values, $V_*'(\theta)$ and $V_*''(\theta)$, in the compact set Θ_r.

Assumption 15.23 (Convergence Rate $V_N'(\theta)$): The convergence rate of the derivative of the expected value of the cost function is an $O(N^{-K})$ in Θ_r: $V_N'(\theta) = V_*'(\theta) + O(N^{-K})$.

Theorem 15.24 (Convergence Rate $\tilde{\theta}(z_0)$ to θ_*): Under Assumptions 15.2(a), 15.4, 15.18, 15.22, and 15.23 the deterministic convergence rate of $\tilde{\theta}(z_0)$ equals $O(N^{-K})$: $\tilde{\theta}(z_0) - \theta_* = O(N^{-K})$.

Proof. Similar to Theorem 15.19. □

15.5 ASYMPTOTIC BIAS

It makes sense to speak about the bias on the estimates $\hat{\theta}(z)$ only if a true model exists and if the true model belongs to the considered model set. Under Assumptions 15.9 and 15.10 or 15.12 or 15.14, Theorem 15.21, eventually combined with Theorem 15.24, gives information about the systematic errors on the estimates. This leads to the following two corollaries.

Corollary 15.25 (Improved Convergence Rate of $\hat{\theta}(z)$ to θ_0): Under the assumptions of Theorem 15.21 and Assumptions 15.9 and 15.10 or 15.12, the minimizer $\hat{\theta}(z)$ can be written as

$$\begin{aligned} \hat{\theta}(z) &= \theta_0 + \delta_\theta(z) + b_\theta(z) \\ \delta_\theta(z) &= -V_N''^{-1}(\theta_0) V_N'^{T}(\theta_0, z) \end{aligned} \tag{15-15}$$

where $\mathcal{E}\{\delta_\theta(z)\} = 0$, $\delta_\theta(z) = O_p(N^{-1/2})$, and $b_\theta(z) = O_p(N^{-1})$.

Proof. It follows directly from Theorem 15.21 and Assumptions 15.9 and 15.10 or 15.12. □

Corollary 15.26 (Improved Convergence Rate of $\hat{\theta}(z)$ to θ_0): Under the assumptions of Theorems 15.21 and 15.24 and Assumptions 15.9 and 15.14, the minimizer $\hat{\theta}(z)$ can be written as

$$\begin{aligned} \hat{\theta}(z) &= \theta_0 + \delta_\theta(z) + b_\theta(z) \\ \delta_\theta(z) &= -V_N''^{-1}(\tilde{\theta}(z_0)) V_N'^{T}(\tilde{\theta}(z_0), z) \end{aligned} \tag{15-16}$$

where $\mathcal{E}\{\delta_\theta(z)\} = 0$, $\delta_\theta(z) = O_p(N^{-1/2})$, and $b_\theta(z) = O_p(N^{-1}) + O(N^{-K})$.

Proof. It follows directly from Theorems 15.21 and 15.24 and Assumptions 15.9 and 15.14. □

Comparing Corollary 15.25 with Corollary 15.26 shows that the deterministic convergence rate of $\tilde{\theta}(z_0)$ to θ_0 influences only $b_\theta(z)$. Because $\delta_\theta(z)$ is a zero mean random variable, it can be concluded from Corollaries 15.25 and 15.26 that the $b_\theta(z)$ term is responsible for the systematic error on the estimate $\hat{\theta}(z)$ and that, in general (Corollary 15.25, Corollary 15.26 with $K > 1/2$), $b_\theta(z)$ tends faster to zero than $\delta_\theta(z)$ as N increases to infinity. Hence, in probability, no systematic errors should be expected when N is sufficiently large.

Although the previous analysis of the systematic errors is already sufficient for our purposes, we will also, briefly, discuss the bias error on $\hat{\theta}(z)$. It is very tempting to conclude from both corollaries that the bias $b_\theta = \mathcal{E}\{b_\theta(z)\}$ decreases to zero as an $O(N^{-1})$ $(O(N^{-1}) + O(N^{-K}))$. However, the expected value of $\hat{\theta}(z)$, and, hence, of $b_\theta(z)$, may, in general, not exist. This is due to the fact that convergence in probability does not exclude realizations of z for which $\hat{\theta}(z)$ tends to infinity. Additional assumptions on the measurements z are required to ensure the existence of $\mathcal{E}\{\hat{\theta}(z)\}$. For example, for quadratic prediction error methods, the eighth-order moments of the disturbing errors should be bounded (see Appendix 9.B of Ljung, 1999). The following pragmatic approach also ensures the existence of $\mathcal{E}\{\hat{\theta}(z)\}$. Define the truncated estimator $\underline{\hat{\theta}}(z)$ as

$$\underline{\hat{\theta}}(z) = \begin{cases} \hat{\theta}(z) & \|\hat{\theta}(z) - \tilde{\theta}(z_0)\|_2 \le L \\ 0 & \|\hat{\theta}(z) - \tilde{\theta}(z_0)\|_2 > L \end{cases} \tag{15-17}$$

where L is a(n) (arbitrarily) large number ($0 < L < \infty$) independent of N. Note that this is exactly what we do in practice: if the estimate is unacceptably large, then we reject it.

Lemma 15.27 (Equivalence between $\underline{\hat{\theta}}(z)$ and $\hat{\theta}(z)$): Under the assumptions of Theorem 15.6 or Theorem 15.8, there exists an N_0 such that for any $N \geq N_0$, $\underline{\hat{\theta}}(z) = \hat{\theta}(z)$ w.p. 1. Moreover, the results of Theorem 15.21 are still valid.

Proof. See Appendix 15.F. □

The estimate $\underline{\hat{\theta}}(z)$ is uniformly bounded and, hence, its expected value exists. This leads to the following theorem.

Theorem 15.28 (Asymptotic Bias on $\underline{\hat{\theta}}(z)$): Under the assumptions of Corollary 15.25 (Corollary 15.26) the asymptotic bias $b_\theta = \mathcal{E}\{b_\theta(z)\}$ of $\underline{\hat{\theta}}(z)$, and its derivative w.r.t. θ_0, $\partial b_\theta / \partial \theta_0$, are an $O(N^{-1})$ $(O(N^{-1}) + O(N^{-K}))$ for all $\theta_0 \in \Theta_{\mathrm{r}}$.

Proof. See Appendix 15.G. □

15.6 ASYMPTOTIC NORMALITY

It makes no sense to estimate parameters if no quality stamp on the result can be given. Otherwise, any random guess is a valuable estimate. For example, the quality stamp can be a region around the estimated value where the limit (true) value lies within some confidence level. Therefore, we would like to know the distribution function of $\hat{\theta}(z)$ for finite N. In most cases it is impossible to calculate, and we can only make statements about the asymptotic distribution function of $\hat{\theta}(z)$.

Theorem 15.29 (Asymptotic Normality of $\sqrt{N}(\hat{\theta}(z) - \tilde{\theta}(z_0))$): Under the assumptions of Theorem 15.21 and Assumption 15.1 ($P = \infty$), $\sqrt{N}(\hat{\theta}(z) - \tilde{\theta}(z_0))$ converges in law at the rate $O(N^{-1/2})$ to a Gaussian random variable with zero mean and covariance matrix $\mathrm{Cov}(\sqrt{N}\delta_\theta(z))$

$$
\begin{aligned}
\mathrm{Cov}(\sqrt{N}\delta_\theta(z)) &= V_N''^{-1}(\tilde{\theta}(z_0))\, Q_N(\tilde{\theta}(z_0))\, V_N''^{-1}(\tilde{\theta}(z_0)) \\
Q_N(\tilde{\theta}(z_0)) &= N\mathcal{E}\{V_N'^{T}(\tilde{\theta}(z_0), z)\, V_N'(\tilde{\theta}(z_0), z)\}
\end{aligned}
\tag{15-18}
$$

Proof. See Appendix 15.I. □

It follows that the uncertainty on the estimated parameters is small if the eigenvalues of the Hessian matrix $V_N''(\tilde{\theta}(z_0))$ are large. Although Theorem 15.29 guarantees neither the convergence of $\mathrm{Cov}(\sqrt{N}\hat{\theta}(z))$ to $\mathrm{Cov}(\sqrt{N}\delta_\theta(z))$ nor the existence of $\mathrm{Cov}(\sqrt{N}\hat{\theta}(z))$, it makes it possible to construct uncertainty regions on $\hat{\theta}(z)$ with a given confidence level. This is sufficient for our purposes.

Additional assumptions on the measurements z are required to ensure the existence and the convergence of $\mathrm{Cov}(\sqrt{N}\hat{\theta}(z))$ (for example, for quadratic prediction error methods the eighth-order moments of the disturbing errors should be bounded: see Appendix 9.B of Ljung, 1999). Another way to ensure its existence is to truncate the estimate $\hat{\theta}(z)$, see (15-17). It makes it possible to strengthen Theorem 15.29 as follows.

Theorem 15.30 (Asymptotic Covariance Matrix of $\underline{\hat{\theta}}(z)$): Under the assumptions of Theorem 15.21 and Assumption 15.1 ($P = \infty$), the covariance matrix $\mathrm{Cov}(\sqrt{N}\underline{\hat{\theta}}(z))$ exists and converges to $\mathrm{Cov}(\sqrt{N}\delta_\theta(z))$ at the rate $O(N^{-1/2})$.

Proof. See Appendix 15.J. □

Note that Theorems 15.29 and 15.30 are also valid for the strongly consistent estimators of Section 15.3 if in addition Assumption 15.9 and Assumption 15.10 or 15.12 or 15.14 (consistency conditions) are satisfied.

15.7 ASYMPTOTIC EFFICIENCY

Analyzing the (asymptotic) efficiency is possible only if a true model exists and if the probability density function of the disturbing noise is known and satisfies some regularity conditions (see Theorem 14.18). In the ideal case, the covariance matrix of the estimate $\hat{\theta}(z)$ should be compared with the generalized Cramér-Rao lower bound (14-84). Because an explicit expression of the bias and its derivative w.r.t. the model parameters is mostly not available, the analysis is simplified to comparing the covariance matrix of $\delta_\theta(z)$ to the unbiased Cramér-Rao lower bound (14-87). Although the classical definition of (asymptotic) efficiency applies only to estimators with finite second-order moments, the concept of efficiency is often extended to (weakly or strongly) consistent estimators. The (weakly or strongly) consistent estimate $\hat{\theta}(z)$ is then said to be asymptotically efficient if

$$\lim_{N \to \infty} (\mathrm{Cov}(\sqrt{N}\delta_\theta(z)) - NFi^{-1}(\theta_0)) = 0 \tag{15-19}$$

where $\mathrm{Cov}(\sqrt{N}\delta_\theta(z))$ is given by (15-18) with $\tilde{\theta}(z_0) = \theta_0$, and with $Fi(\theta_0)$ the Fisher information matrix of the model parameters. If (15-19) is valid for finite N and $b_\theta(z) = 0$, then the estimate is efficient. Note that (15-19) can be valid while $\mathrm{Cov}(\hat{\theta}(z))$ may not exist.

This classical definition of asymptotic efficiency can be applied to the truncated estimator $\hat{\underline{\theta}}(z)$ (15-17). Because the bias of $\hat{\underline{\theta}}(z)$ and its derivative w.r.t. θ_0 are asymptotically zero (Theorem 15.28), we can compare the covariance matrix of $\hat{\underline{\theta}}(z)$ to the unbiased Cramér-Rao lower bound (14-87). The asymptotically unbiased estimate $\hat{\underline{\theta}}(z)$ is asymptotically efficient if

$$\lim_{N \to \infty} (\mathrm{Cov}(\sqrt{N}\hat{\underline{\theta}}(z)) - NFi^{-1}(\theta_0)) = 0 \tag{15-20}$$

Note that Theorem 15.30 can be used to verify (15-20).

15.8 OVERVIEW OF THE ASYMPTOTIC PROPERTIES

In this section we give an overview of the asymptotic properties of the minimizer $\hat{\theta}(z)$ (15-3) of a cost function $V_N(\theta, z)$ (15-1) that is quadratic-in-the-measurements. In the analysis of the stochastic properties of $\hat{\theta}(z)$, it turned out that the expected value of the cost function $V_N(\theta)$, its limit value $V_*(\theta)$, and the corresponding minimizers $\tilde{\theta}(z_0)$ and θ_*, play an important role. Therefore, we summarize the notations in Table 15-1.

TABLE 15-1 Overview of the Notations Used

Cost function	$V_N(\theta, z)$	$V_N(\theta) = \mathcal{E}\{V_N(\theta, z)\}$	$V_*(\theta) = \lim\limits_{N \to \infty} V_N(\theta)$
Minimizer	$\hat{\theta}(z)$	$\tilde{\theta}(z_0)$	θ_*

The minimizer $\hat{\theta}(z)$ (15-3) of the cost function $V_N(\theta, z)$ (15-1) has the following asymptotic ($N \to \infty$) properties:

1. *Stochastic convergence:* $\hat{\theta}(z)$ converges strongly to $\tilde{\theta}(z_0)$ (Theorem 15.6).
2. *Stochastic convergence rate:* $\hat{\theta}(z)$ converges in probability at the rate $O_P(N^{-1/2})$ to $\tilde{\theta}(z_0)$ (Theorem 15.19).
3. *Systematic and stochastic errors:* $\hat{\theta}(z)$ converges in probability to $\tilde{\theta}(z_0)$ with

$$\begin{aligned}\hat{\theta}(z) &= \tilde{\theta}(z_0) + \delta_\theta(z) + b_\theta(z)\\ \delta_\theta(z) &= -V_N''^{-1}(\tilde{\theta}(z_0))V_N'^T(\tilde{\theta}(z_0), z)\end{aligned} \qquad (15\text{-}21)$$

   where $\delta_\theta(z) = O_p(N^{-1/2})$, with $\mathcal{E}\{\delta_\theta(z)\} = 0$, is the dominating stochastic error and where $b_\theta(z) = O_P(N^{-1})$ contains the contribution of the systematic errors (Theorem 15.21).
4. *Asymptotic normality:* $\sqrt{N}(\hat{\theta}(z) - \tilde{\theta}(z_0))$ converges in law at the rate $O(N^{-1/2})$ to a Gaussian random variable with zero mean and covariance matrix $\mathrm{Cov}(\sqrt{N}\delta_\theta(z))$

$$\begin{aligned}\mathrm{Cov}(\sqrt{N}\delta_\theta(z)) &= V_N''^{-1}(\tilde{\theta}(z_0))Q_N(\tilde{\theta}(z_0))V_N''^{-1}(\tilde{\theta}(z_0))\\ Q_N(\tilde{\theta}(z_0)) &= N\mathcal{E}\{V_N'^T(\tilde{\theta}(z_0), z)V_N'(\tilde{\theta}(z_0), z)\}\end{aligned} \qquad (15\text{-}22)$$

   (Theorem 15.29).
5. *Deterministic convergence:* $\tilde{\theta}(z_0)$ converges to θ_* (Theorem 15.8) at the rate $O(N^{-K})$ (Theorem 15.24).

If in addition $V_N(\theta, z)$ satisfies the consistency conditions then,

6. *Consistency:* $\hat{\theta}(z)$ is strongly consistent; in properties 1 to 4 replace $\tilde{\theta}(z_0)$ or $\lim\limits_{F \to \infty} \tilde{\theta}(Z_0) = \theta_*$ by θ_0 (Theorems 15.11, 15.13, and 15.15).
7. *Asymptotic bias:* the asymptotic bias $b_\theta = \mathcal{E}\{b_\theta(z)\}$ and its derivative w.r.t. θ_0, $\partial b_\theta/\partial\theta_0$, of $\underline{\hat{\theta}}(z)$ are an $O(N^{-1})$ or an $O(N^{-1}) + O(N^{-K})$ (Theorem 15.28).

Properties 1 to 7 make it possible to predict the stochastic behavior (uncertainty, bias, ...) of the estimate $\hat{\theta}(z)$ if, for example, nine times more data are collected. Property 1 ensures that $\hat{\theta}(z)$ will be closer to the minimizer $\tilde{\theta}(z_0)$ of the expected value of the cost function. Property 2 tells us that $\hat{\theta}(z)$ will be (in probability) three times closer to $\tilde{\theta}(z_0)$. From property 3 it follows that the systematic and stochastic errors in the residual $\hat{\theta}(z) - \tilde{\theta}(z_0)$ decrease with a factor of 9 and 3, respectively. Finally, property 4 ensures that the distribution function of $\hat{\theta}(z)$ is three times closer to a normal distribution. Similar results are obtained when no model errors are present $\tilde{\theta}(z_0) = \theta_0$.

15.9 EXERCISES

15.1. Prove the strong convergence of the nonlinear least squares cost function $\sum_{k=1}^{N}(y(k) - f(\theta, u_0(k)))^2/(N\sigma_k^2)$ with $y(k) = y_0(k) + n_y(k)$, $n_y(k)$ mixing of order 4, and $\sigma_k^2 = \mathrm{var}(n_y(k))$. Under which condition(s) is the convergence uniform w.r.t. θ (hint: follow the lines of Lemma 15.3)?

15.2. Show that the expected value of the cost function is given by (15-8) (hint: use $x^TAx = \mathrm{trace}(Axx^T)$ for any $x \in \mathbb{R}^N$ and $A \in \mathbb{R}^{N \times N}$).

15.3. Show that the term $\operatorname{trace}(W_N(\theta)C_{n_z})$ in the nonlinear least squares cost function of Exercise 15.1 is θ independent.

15.4. Let $\hat{\theta}(z) \in \mathbb{R}$ be the minimizer of $V_N(\theta, z)$ and θ_0 the unique global minimizer of $V_N(\theta)$ for any N, ∞ included. Assume that the cost function has continuous third-order derivative w.r.t. θ for all $\theta \in \Theta_r$ with $\|\partial^3 W_N(\theta)/\partial\theta^3\|_1 \leq c < \infty$. Assume, furthermore, that $V_N''(\theta_0) = 0$, $V_N'''(\theta_0) \neq 0$, and that Assumptions 15.1 ($P = 4$), 15.2, and 15.16 are satisfied. Show that $\hat{\theta}(z) = \theta_0 + O_p(N^{-1/4})$ (hint: follow the lines of the proof of Theorem 15.19 using $V_N'(\hat{\theta}(z), z) = V_N'(\theta_0, z) + V_N''(\theta_0, z)(\hat{\theta}(z) - \theta_0) + \frac{1}{2}V_N'''(\widehat{\theta}, z)(\hat{\theta}(z) - \theta_0)^2$).

15.5. Prove Theorem 15.24 (hint: follow the lines of the proof of Theorem 15.19).

15.6. Consider the nonlinear least square estimator of Exercise 15.1 and assume that $n_y(k)$ is independent over k. Assume, furthermore, that the true model is included in the model set ($y_0(k) = f(\theta_0, u_0(k))$). Show that the covariance matrix of $\delta_\theta(z)$ is given by $(\sum_{k=1}^{N}\sigma_k^{-2} f_{0k}'^T f_{0k}')^{-1}$, with $f_{0k}' = \partial f(\theta, u_0(k))/\partial\theta|_{\theta=\theta_0}$ (hint: use (15-18)).

15.7. Show the weak convergence and the weak consistency of the estimates $\hat{\theta}(z)$ (convergence in prob.) when Assumption 15.2(b) is not fulfilled (hint: use the mean square convergence of the cost function and interrelation 5 of Section 14.7).

15.10 APPENDIXES

Appendix 15.A: Proof of the Strong Convergence of the Cost Function (Lemma 15.3)

The cost function (15-1) can be written as

$$V_N(\theta, z) = \frac{1}{N}z^T y_N = \frac{1}{N}\sum_{t=1}^{N} x_N(t) \tag{15-23}$$

with $y_N = W_N(\theta)z$ and $x_N(t) = z_{[t]}y_{N[t]}$. Because $W_N(\theta)$ has a bounded 1-norm for $N = 1, 2, \ldots, \infty$ (Assumption 15.2) and $z_{[t]}$ is mixing over t of order 4, $N = 1, 2, \ldots, \infty$ (Assumption 15.1), the conditions of Corollary 14.7 are satisfied with $P = 4$. Hence, $y_{N[t]}$ is mixing over t of order 4, $N = 1, 2, \ldots, \infty$, so that $x_N(t) = z_{[t]}y_{N[t]}$ is mixing over t of order 2, $N = 1, 2, \ldots, \infty$ (Lemma 14.9). This proves that the law of large numbers for mixing sequences and the corresponding convergence rate is valid for (15-23) (see Section 14.9, version 3 of (14-69) and (14-72))

$$V_N(\theta, z) = V_N(\theta) + O_{\text{m.s.}}(N^{-1/2}) \tag{15-24}$$

If in addition it can be shown that $\operatorname{var}(\sum_{t=1}^{s} x_r(t) - x_s(t)) = O(r - s)$, $r \geq s$, then also the strong law of large numbers for mixing sequences applies to (15-23) (see Section 14.9, version 3 of (14-69)). Because $\sum_{t=1}^{s} x_r(t) - x_s(t) = \Delta_1 + \Delta_2$ with

$$\Delta_1 = \sum_{t=1}^{s} z_{[t]}\sum_{l=s+1}^{r} W_{r[t,l]}(\theta)z_{[l]}$$

$$\Delta_2 = \sum_{t=1}^{s} z_{[t]}\sum_{l=1}^{s} (W_{r[t,l]}(\theta) - W_{s[t,l]}(\theta))z_{[l]}$$

and $\operatorname{var}(\Delta_1 + \Delta_2) \leq 2\operatorname{var}(\Delta_1) + 2\operatorname{var}(\Delta_2)$ (see Exercise 14.1), it follows that it is sufficient to show that the variance of Δ_1 and Δ_2 are both an $O(r - s)$.

1. Study of Δ_1
   Rewriting Δ_1 as $\Delta_1 = \sum_{l=s+1}^{r} z_{[l]} y_s(l)$, where $y_s(l) = \sum_{t=1}^{s} z_{[t]} W_{r[t,l]}(\theta)$ is mixing over l of order 4 (Corollary 14.7), and $z_{[l]} y_s(l)$ is mixing over l of order 2 (Lemma 14.9), gives

$$\operatorname{var}(\Delta_1) = \operatorname{cum}(\Delta_1, \Delta_1) = \sum_{l_1, l_2 = s+1}^{r} \operatorname{cum}(z_{[l_1]} y_s(l_1), z_{[l_2]} y_s(l_2)) \le O(r-s) \quad (15\text{-}25)$$

   The last inequality in (15-25) is due to property (14-36) of the mixing condition.

2. Study of Δ_2
   If $\|W_{r[1:s,\,1:s]}(\theta) - W_s(\theta)\|_1^2 = O((r-s)/r)$, then there exists a matrix $T(\theta)$ such that

$$W_{r[1:s,\,1:s]}(\theta) - W_s(\theta) = T(\theta)\sqrt{(r-s)/r}$$

   with $\|T(\theta)\|_1 \le c < \infty$ for any $r \ge s$, infinity included. It facilitates rewriting Δ_2 as

$$\Delta_2 = \sqrt{(r-s)/r} \sum_{t=1}^{s} z_{[t]} y_s(t)$$

   where $y_s(t) = \sum_{l=1}^{s} T_{[t,l]}(\theta) z_{[l]}$ is mixing of order 4 (Corollary 14.7), and $z_{[t]} y_s(t)$ is mixing of order 2 (Lemma 14.9). Hence, we find

$$\operatorname{var}(\Delta_2) = \operatorname{cum}(\Delta_2, \Delta_2) = \frac{r-s}{r} \sum_{t_1, t_2 = 1}^{s} \operatorname{cum}(z_{[t_1]} y_s(t_1), z_{[t_2]} y_s(t_2)) \le O(r-s)$$

   where the last inequality is due to property (14-36) of the mixing condition and $r \ge s$.

We conclude that the cost function (15-23) converges w.p. 1 to its expected value, making (15-6) valid. It remains to be proved that the mean square convergence (15-24) and the almost sure convergence (15-6) are uniform in Θ_{r}. Because $V_N(\theta, z)$ is continuous in the compact set Θ_{r} (Assumption 15.2), it is uniformly bounded in Θ_{r}. Hence, the maximum over θ can be taken, where necessary, in the inequalities above and in those of the proof of the strong law of large numbers for mixing sequences (see Appendix 14.G). □

Appendix 15.B: Proof of the Strong Convergence of the Minimizer (Theorem 15.6)

The proof follows the lines of the proof of Theorem 2 in Söderström (1974). Because the assumptions of Lemma 15.3 are satisfied, we can consider only those realizations for which $V_N(\theta, z)$ converges to $V_N(\theta)$. These realizations have probability measure 1. Choose an arbitrary $\varepsilon > 0$ and construct the set $\Theta_\varepsilon = \{\theta \mid \|\theta - \tilde{\theta}(z_0)\|_2 < \varepsilon\} \subset \Theta_{\mathrm{r}}$. We will show that the global minimizer(s) of $V_N(\theta, z)$ is (are) located in Θ_ε for N sufficiently large. This proves the theorem because ε can be made arbitrarily small.

Because $V_N(\theta)$ is a continuous function in Θ_{r} (Assumption 15.2), we can choose a $\delta > 0$ such that

$$\min_{\theta \in \Theta_r \backslash \Theta_\varepsilon} V_N(\theta) \geq V_N(\tilde{\theta}(z_0)) + \delta \tag{15-26}$$

As $V_N(\theta, z)$ converges uniformly to $V_N(\theta)$ in Θ_r, there exists an N_0 such that for any $N \geq N_0$ and any $\theta \in \Theta_r$

$$-\delta/3 \leq V_N(\theta, z) - V_N(\theta) \leq \delta/3 \tag{15-27}$$

Using the upper inequality in (15-27), evaluated in $\tilde{\theta}(z_0)$, we get

$$\min_{\theta \in \Theta_r} V_N(\theta, z) \leq V_N(\tilde{\theta}(z_0), z) \leq V_N(\tilde{\theta}(z_0)) + \delta/3 \tag{15-28}$$

Using the lower inequality in (15-27) and result (15-26), we find

$$\min_{\theta \in \Theta_r \backslash \Theta_\varepsilon} V_N(\theta, z) \geq \min_{\theta \in \Theta_r \backslash \Theta_\varepsilon} V_N(\theta) - \delta/3 \geq V_N(\tilde{\theta}(z_0)) + 2\delta/3 \tag{15-29}$$

From (15-28) and (15-29) it follows that

$$\min_{\theta \in \Theta_r} V_N(\theta, z) < \min_{\theta \in \Theta_r \backslash \Theta_\varepsilon} V_N(\theta, z)$$

which shows that the minimizer(s) of $V_N(\theta, z)$ is (are) located in Θ_ε. □

Appendix 15.C: Lemmas

In this appendix we study the asymptotic ($N \to \infty$) properties of the function $f_N(\hat{\theta}, z)$ where $\hat{\theta} \in \mathbb{R}^{n_\theta}$ is a stochastic vector of finite dimension (n_θ is an N-independent integer) and $z \in \mathbb{R}^N$ are the noisy observations. The convergence, the convergence rate, the asymptotic bias, and the asymptotic distribution function are analyzed. For the bias analysis, the concept of the truncated estimate of Section 15.5 is used. Although all the theorems are proved assuming convergence w.p. 1, they are also valid for convergence in probability (see Section 14.7, interrelation 5).

Lemma 15.31 (Strong or Weak Convergence): Let $f_N(\theta, z) \in \mathbb{R}$ be a continuous function of θ in Θ_r, a compact subset of $\mathbb{R}^{n_\theta}$, and $z \in \mathbb{R}^N$ a stochastic variable. If

1. $f_N(\theta, z)$ converges uniformly w.p. 1 (in prob.) to $f(\theta)$ in Θ_r,
2. $\hat{\theta}$ converges w.p. 1 (in prob.) to θ_*, an interior point of Θ_r,
   then $f_N(\hat{\theta}, z)$ converges w.p. 1 (in prob.) to $f(\theta_*)$.

Proof (Strong Convergence). Consider the stochastic realizations of z for which $f_N(\theta, z)$ converges uniformly to $f(\theta)$ in Θ_r and $\hat{\theta}$ converges to θ_*. Due to the almost sure convergence, these realizations have probability measure one. Choose an arbitrary $\varepsilon > 0$ and construct the set $\Theta_\varepsilon = \{\theta | \|\theta - \theta_*\|_2 < \varepsilon\} \subset \Theta_r$. Because for any of the considered realizations z, $f_N(\theta, z)$ converges uniformly to $f(\theta)$ in Θ_{r} and $\hat{\theta}$ converges to θ_*, there exists, for any $\delta > 0$, an N_0 independent of θ such that for any $N \geq N_0$

$$\hat{\theta} \in \Theta_\varepsilon \text{ and } |f_N(\theta, z) - f(\theta)| \leq \frac{\delta}{2} \text{ for any } \theta \in \Theta_{\mathrm{r}} \tag{15-30}$$

The function $f(\theta)$ is continuous in Θ_{r} because it is the limit of a uniformly convergent sequence of continuous functions (see Kaplan, 1993, Theorem 31, Remark 2). Hence, there exists an ε such that

$$|f(\theta) - f(\theta_*)| \leq \frac{\delta}{2} \text{ for any } \theta \in \Theta_\varepsilon \tag{15-31}$$

Combining (15-30) and (15-31) shows that for any $\delta > 0$ there exist an ε and an N_0 such that for any $N \geq N_0$

$$|f_N(\hat{\theta}, z) - f(\theta_*)| \leq |f_N(\hat{\theta}, z) - f(\hat{\theta})| + |f(\hat{\theta}) - f(\theta_*)| \leq \delta \tag{15-32}$$

Making δ arbitrarily small and noting that the considered realizations z occur w.p. 1 reveals directly that $\underset{N \to \infty}{\text{a.s.lim}}(f_N(\hat{\theta}, z) - f(\theta_*)) = 0$ or $f_N(\hat{\theta}, z) = f(\theta_*) + o_{\text{a.s.}}(N^0)$. □

Corollary 15.32 (Strong or Weak Convergence): Let $z \in \mathbb{R}^N$ be a stochastic variable. Let $f_N(\theta, \psi, \eta, z) \in \mathbb{R}$ be a jointly continuous function of $\theta \in \Theta_{\mathrm{r}}$, $\psi \in \Psi_{\mathrm{r}}$, and $\eta \in \boldsymbol{\eta}_{\mathrm{r}}$. Θ_{r}, Ψ_{r}, and $\boldsymbol{\eta}_{\mathrm{r}}$ are compact subsets of, respectively, $\mathbb{R}^{n_\theta}$, $\mathbb{R}^{n_\psi}$, and $\mathbb{R}^{n_\eta}$. If

1. $f_N(\theta, \psi, \eta, z)$ converges uniformly w.p. 1 (in prob.) to $f(\theta, \psi, \eta)$ in Θ_{r}, Ψ_{r}, and $\boldsymbol{\eta}_{\mathrm{r}}$.
2. $\hat{\theta}$, $\hat{\psi}$ converge w.p. 1 (in prob.) to θ_*, ψ_*, interior points of Θ_{r}, Ψ_{r},
   then $f_N(\hat{\theta}, \hat{\psi}, \eta, z)$ converges uniformly w.p. 1 (in prob.) to $f(\theta_*, \psi_*, \eta)$ in $\boldsymbol{\eta}_{\mathrm{r}}$.

Lemma 15.33 (Strong or Weak Convergence): Let $f_N(\theta, z) \in \mathbb{R}$ be a continuous function of θ in Θ_{r}, a compact subset of $\mathbb{R}^{n_\theta}$, and $z \in \mathbb{R}^N$ a stochastic variable. If

1. $f_N(\theta, z) = O_{\text{a.s.}}(N^k)$ $(O_{\mathrm{p}}(N^k))$ uniformly in Θ_{r},
2. $\hat{\theta}$ converges w.p. 1 (in prob.) to θ_*, an interior point of Θ_{r},
   then $f_N(\hat{\theta}, z) = O_{\text{a.s.}}(N^k)$ $(O_{\mathrm{p}}(N^k))$.

Proof. Similar to that of Lemma 15.31. □

Lemma 15.34 (Convergence Rate): Let $z \in \mathbb{R}^N$ be a stochastic variable. Let $f_N(\theta, z) \in \mathbb{R}$ and $f_N'(\theta, z)$, its derivative w.r.t. θ, be continuous functions of θ in Θ_{r}, a compact subset of $\mathbb{R}^{n_\theta}$. If

1. $f_N(\theta, z)$ converges uniformly w.p. 1 (in prob.) to $f(\theta)$ in Θ_{r} at the rate $O_{\mathrm{p}}(N^{-1/2})$,
2. $\|f_N'(\theta, z)\|_2 \leq O_{\text{a.s.}}(N^0)$ $(O_{\mathrm{p}}(N^0))$ uniformly in Θ_{r},

3. $\hat{\theta}$ converges w.p. 1 (in prob.) to θ_* at the rate $O_{\rm p}(N^{-1/2})$, with θ_* an interior point of $\boldsymbol{\Theta}_{\rm r}$,

then $f_N(\hat{\theta}, z)$ converges w.p. 1 (in prob.) to $f(\theta_*)$ at the rate $O_{\rm p}(N^{-1/2})$.

Proof (w.p. 1). Note that the conditions of Lemma 15.31 are satisfied so that only the convergence rate must be proved. Applying the mean value theorem (Kaplan, 1993) to $f_N(\theta, z)$ at the points $\hat{\theta}$, θ_* gives

$$f_N(\hat{\theta}, z) = f_N(\theta_*, z) + f_N'(\overset{\frown}{\theta}, z)(\hat{\theta} - \theta_*) \tag{15-33}$$

with $\overset{\frown}{\theta}$ a point on the straight line connecting $\hat{\theta}$ to θ_* ($\overset{\frown}{\theta} = t\hat{\theta} + (1-t)\theta_*$ with $t \in [0, 1]$). $\overset{\frown}{\theta}$ converges w.p. 1 to θ_* because

$$\operatorname*{a.s.lim}_{N\to\infty}(\overset{\frown}{\theta} - \theta_*) = (\lim_{N\to\infty} t)\operatorname*{a.s.lim}_{N\to\infty}(\hat{\theta} - \theta_*) = 0 \tag{15-34}$$

Consider the realizations z for which $\overset{\frown}{\theta}$ converges to θ_* and $\|f_N'(\theta, z)\|_2 \le O(N^0)$ uniformly in $\boldsymbol{\Theta}_{\rm r}$. For these realizations, there is an N_0 such that for any $N \ge N_0$, $\overset{\frown}{\theta} \in \boldsymbol{\Theta}_{\rm r}$ and, hence, $\|f_N'(\overset{\frown}{\theta}, z)\|_2 \le O(N^0)$. Because these realizations occur w.p. 1, we have $\|f_N'(\overset{\frown}{\theta}, z)\|_2 \le O_{\rm a.s.}(N^0)$ and, hence, $\|f_N'(\overset{\frown}{\theta}, z)\|_2 \le O_{\rm p}(N^0)$ (see Section 14.7). Putting this result in (15-33), taking into account that $f_N(\theta_*, z) = f(\theta_*) + O_{\rm p}(N^{-1/2})$ (condition 1), $\hat{\theta} = \theta_* + O_{\rm p}(N^{-1/2})$ (condition 3), and that n_θ is an N-independent integer proves the lemma. □

Corollary 15.35 (Convergence Rate): Let $z \in \mathbb{R}^N$ be a stochastic variable. Let $f_N(\theta, \psi, \eta, z) \in \mathbb{R}$ and its derivatives w.r.t. θ and ψ be jointly continuous functions of $\theta \in \boldsymbol{\Theta}_{\rm r}$, $\psi \in \boldsymbol{\psi}_{\rm r}$, and $\eta \in \boldsymbol{\eta}_r$, $\boldsymbol{\Theta}_{\rm r}$, $\boldsymbol{\psi}_{\rm r}$, and $\boldsymbol{\eta}_{\rm r}$ are compact subsets of respectively $\mathbb{R}^{n_\theta}$, $\mathbb{R}^{n_\psi}$, and $\mathbb{R}^{n_\eta}$. If

1. $f_N(\theta, \psi, \eta, z)$ converges uniformly w.p. 1 (in prob.) to $f(\theta, \psi, \eta)$ in $\boldsymbol{\Theta}_{\rm r}$, $\boldsymbol{\psi}_{\rm r}$, and $\boldsymbol{\eta}_{\rm r}$ at the rate $O_{\rm p}(N^{-1/2})$,
2. $\|\partial f_N(\theta, \psi, \eta, z)/\partial\theta\|_2 \le O_{\rm a.s.}(N^0)$ $(O_{\rm p}(N^0))$, $\|\partial f_N(\theta, \psi, \eta, z)/\partial\psi\|_2 \le O_{\rm a.s.}(N^0)$ $(O_{\rm p}(N^0))$ uniformly in $\boldsymbol{\Theta}_{\rm r}$, $\boldsymbol{\psi}_{\rm r}$, and $\boldsymbol{\eta}_{\rm r}$,
3. $\hat{\theta}$, $\hat{\psi}$ converge w.p. 1 (in prob.) to θ_*, ψ_* at the rate $O_{\rm p}(N^{-1/2})$, where θ_*, ψ_* are interior points of $\boldsymbol{\Theta}_{\rm r}$, $\boldsymbol{\psi}_{\rm r}$,

then $f_N(\hat{\theta}, \hat{\psi}, \eta, z)$ converges uniformly w.p. 1 (in prob.) to $f(\theta_*, \psi_*, \eta)$ in $\boldsymbol{\eta}_{\rm r}$ at the rate $O_{\rm p}(N^{-1/2})$.

Lemma 15.36 (Asymptotic Bias): Let $z \in \mathbb{R}^N$ be a stochastic variable. Let $f_N(\theta, z) \in \mathbb{R}$ and $f_N^{(k)}(\theta, z)$, $k = 1, 2$, its derivatives w.r.t. θ, be continuous functions of θ in $\boldsymbol{\Theta}_{\rm r}$, a compact subset of $\mathbb{R}^{n_\theta}$. If

1. $f_N(\theta, z)$, $f_N'(\theta, z)$ converge uniformly w.p. 1 (in prob.) to $\mathscr{E}\{f_N(\theta, z)\} = f_N(\theta)$, $\mathscr{E}\{f_N'(\theta, z)\} = f_N'(\theta)$ in $\boldsymbol{\Theta}_{\rm r}$ at the rate $O_{\rm p}(N^{-1/2})$,
2. $\|f_N''(\theta, z)\|_2 \le O_{\rm a.s.}(N^0)$ $(O_{\rm p}(N^0))$ uniformly in $\boldsymbol{\Theta}_{\rm r}$,

3. $\hat{\theta}$ converges w.p. 1 (in prob.) to θ_0 at the rate $O_{\text{p}}(N^{-1/2})$, with θ_0 an interior point of $\mathbb{\Theta}_{\text{r}}$,
4. The bias of the truncated estimate $\underline{\hat{\theta}}$ is an $O(N^{-1})$: $\mathcal{E}\{\underline{\hat{\theta}}\} = \theta_0 + O(N^{-1})$,

then $f_N(\hat{\theta}, z)$ converges w.p. 1 (in prob.) to $f(\theta_0)$ at the rate $O_{\text{p}}(N^{-1/2})$ and the bias of the truncated estimate $\underline{f_N}(\hat{\theta}, z)$ is an $O(N^{-1})$: $\mathcal{E}\{\underline{f_N}(\hat{\theta}, z)\} = f_N(\theta_0) + O(N^{-1})$.

Proof (w.p. 1). The conditions of Lemma 15.34 are fulfilled for $f_N(\hat{\theta}, z)$ so that only the claim about the asymptotic bias must be proved. Applying the mean value theorem to $f_N(\theta, z)$ at the points $\hat{\theta}$, θ_0 gives

$$f_N(\hat{\theta}, z) = f_N(\theta_0, z) + f_N'(\overset{\frown}{\theta}, z)(\hat{\theta} - \theta_0) \tag{15-35}$$

where $\overset{\frown}{\theta} = t\hat{\theta} + (1-t)\theta_0$ with $t \in [0, 1]$. Note that $f_N'(\overset{\frown}{\theta}, z)$ satisfies the conditions of Lemma 15.34 because $\overset{\frown}{\theta}$ converges w.p. 1 to θ_0 at the rate $O_{\text{p}}(N^{-1/2})$ (proof: similar to (15-34)). Referring to the equivalence between the truncated and the original estimate (Lemma 15.27), there is an N_0 such that for any $N \geq N_0$ w.p. 1, $\underline{\hat{\theta}} = \hat{\theta}$, $\underline{f_N}(\hat{\theta}, z) = f_N(\hat{\theta}, z)$, and $\underline{f_N}'(\overset{\frown}{\theta}, z) = f_N'(\overset{\frown}{\theta}, z)$. Assume now that $N \geq N_0$ and define $\mathbb{Z}_L$ as the set of realizations z for which $\hat{\theta}$ converges to θ_0, $f_N(\theta_0, z)$ converges to $f_N(\theta_0)$, $f_N'(\overset{\frown}{\theta}, z)$ converges to $f_N'(\theta_0)$, $\underline{\hat{\theta}} = \hat{\theta}$, $\underline{f_N}(\hat{\theta}, z) = f_N(\hat{\theta}, z)$, and $\underline{f_N}'(\overset{\frown}{\theta}, z) = f_N'(\overset{\frown}{\theta}, z)$. The set $\mathbb{Z}_L$ has probability measure one. For $z \in \mathbb{Z}_L$, (15-35) can be written as

$$\underline{f_N}(\hat{\theta}, z) = f_N(\theta_0, z) + \underline{f_N}'(\overset{\frown}{\theta}, z)(\underline{\hat{\theta}} - \theta_0) \tag{15-36}$$

Using the convergence rates of $\underline{\hat{\theta}}$ (condition 3) and $\underline{f_N}'(\overset{\frown}{\theta}, z)$ (Lemma 15.34), (15-36) becomes

$$\underline{f_N}(\hat{\theta}, z) = f_N(\theta_0, z) + f_N'(\theta_0)(\underline{\hat{\theta}} - \theta_0) + O_{\text{p}}(N^{-1}) \tag{15-37}$$

where $O_{\text{p}}(N^{-1})$ is a uniformly bounded random variable. Calculating the expected value of (15-37) over all realizations $z \in \mathbb{Z}_L$, taking into account that $\mathcal{E}\{O_{\text{p}}(N^{-1}) | z \in \mathbb{Z}_L\} = O(N^{-1})$ (see Section 14.8, (14-63)), and that by definition of the truncated estimate, $\mathcal{E}\{\underline{f_N}(\hat{\theta}, z) | z \in \mathbb{Z}_L\} = \mathcal{E}\{\underline{f_N}(\hat{\theta}, z)\}$, gives

$$\mathcal{E}\{\underline{f_N}(\hat{\theta}, z)\} = \mathcal{E}\{f_N(\theta_0, z) | z \in \mathbb{Z}_L\} + O(N^{-1})$$

Because $\text{Prob}(z \notin \mathbb{Z}_L) = 0$ for $N \geq N_0$ and $\mathcal{E}\{f_N(\theta_0, z)\}$ exists and is finite, we have

$$\mathcal{E}\{f_N(\theta_0, z) | z \in \mathbb{Z}_L\} = \mathcal{E}\{f_N(\theta_0, z)\} = f_N(\theta_0)$$

which concludes the proof. □

Corollary 15.37 (Asymptotic Bias): Let $z \in \mathbb{R}^N$ be a stochastic variable. Let $f_N(\theta, \psi, \eta, z) \in \mathbb{R}$, and its first- and second-order derivatives w.r.t. θ and ψ, be jointly continuous functions of $\theta \in \mathbb{\Theta}_{\text{r}}$, $\psi \in \mathbb{\Psi}_{\text{r}}$, and $\eta \in \mathbb{\eta}_{\text{r}}$. $\mathbb{\Theta}_{\text{r}}$, $\mathbb{\Psi}_{\text{r}}$, and $\mathbb{\eta}_{\text{r}}$ are compact subsets of, respectively, $\mathbb{R}^{n_\theta}$, $\mathbb{R}^{n_\psi}$, and $\mathbb{R}^{n_\eta}$. If

1. $f_N(\theta, \psi, \eta, z)$, $\partial f_N(\theta, \psi, \eta, z)/\partial\theta$, $\partial f_N(\theta, \psi, \eta, z)/\partial\psi$ converge uniformly w.p. 1 (in prob.) to their expected values $f_N(\theta, \psi, \eta)$, $\partial f_N(\theta, \psi, \eta)/\partial\theta$, $\partial f_N(\theta, \psi, \eta)/\partial\psi$ in $\mathbb{\Theta}_\mathrm{r}$, $\mathbb{\Psi}_\mathrm{r}$, and $\mathbb{\eta}_\mathrm{r}$ at the rate $O_\mathrm{p}(N^{-1/2})$,
2. $\left\|\dfrac{\partial^2 f_N(\theta, \psi, \eta, z)}{\partial\theta^2}\right\|_2 \le O_{\mathrm{a.s.}}(N^0)\ (O_\mathrm{p}(N^0))$, $\left\|\dfrac{\partial^2 f_N(\theta, \psi, \eta, z)}{\partial\psi^2}\right\|_2 \le O_{\mathrm{a.s.}}(N^0)\ (O_\mathrm{p}(N^0))$ uniformly in $\mathbb{\Theta}_\mathrm{r}$, $\mathbb{\Psi}_\mathrm{r}$, and $\mathbb{\eta}_\mathrm{r}$,
3. $\hat\theta$, $\hat\psi$ converge w.p. 1 (in prob.) to θ_0, ψ_0 at the rate $O_\mathrm{p}(N^{-1/2})$, where θ_0, ψ_0 are interior points of $\mathbb{\Theta}_\mathrm{r}$, $\mathbb{\Psi}_\mathrm{r}$,
4. The bias of the truncated estimates $\underline{\hat\theta}$, $\underline{\hat\psi}$ is an $O(N^{-1})$,

then $f_N(\hat\theta, \hat\psi, \eta, z)$ converges uniformly w.p. 1 (in prob.) to $f_N(\theta_0, \psi_0, \eta)$ in $\mathbb{\eta}_\mathrm{r}$ at the rate $O_\mathrm{p}(N^{-1/2})$, and the bias of the truncated estimate $\underline{f_N}(\hat\theta, \hat\psi, \eta, z)$ is an $O(N^{-1})$.

Lemma 15.38 (Asymptotic Distribution Function): Let $z \in \mathbb{R}^N$ be a stochastic variable. Let $f_N(\theta, z) \in \mathbb{R}$, and $f_N^{(k)}(\theta, z)$, $k = 1, 2$, its derivatives w.r.t. θ, be continuous functions of θ in $\mathbb{\Theta}_\mathrm{r}$, a compact subset of $\mathbb{R}^{n_\theta}$. If

1. $f_N(\theta, z)$ converges uniformly w.p. 1 (in prob.) to $f(\theta)$ in $\mathbb{\Theta}_\mathrm{r}$ at the rate $O_\mathrm{p}(N^{-1/2})$ and is asymptotically normally distributed at the rate $O(N^{-1/2})$,
2. $f_N'(\theta, z)$ converges uniformly w.p. 1 (in prob.) to $f'(\theta)$ in $\mathbb{\Theta}_\mathrm{r}$ at the rate $O_\mathrm{p}(N^{-1/2})$,
3. $\|f_N''(\theta, z)\|_2 \le O_{\mathrm{a.s.}}(N^0)\ (O_\mathrm{p}(N^0))$ uniformly in $\mathbb{\Theta}_\mathrm{r}$,
4. $\hat\theta$ converges w.p. 1 (in prob.) to θ_* at the rate $O_\mathrm{p}(N^{-1/2})$, with θ_* an interior point of $\mathbb{\Theta}_\mathrm{r}$, and is asymptotically normally distributed at the rate $O(N^{-1/2})$,

then $f_N(\hat\theta, z)$ converges w.p. 1 (in prob.) to $f(\theta_*)$ at the rate $O_\mathrm{p}(N^{-1/2})$ and is asymptotically normally distributed at the rate $O(N^{-1/2})$. Moreover, we have

$$f_N(\hat\theta, z) = f_N(\theta_*, z) + f'(\theta_*)(\hat\theta - \theta_*) + O_\mathrm{p}(N^{-1})$$

Proof (w.p. 1). Condition 2 implies that $\|f_N'(\theta, z)\|_2 \le O_{\mathrm{a.s.}}(N^0)$ and, hence, all the assumptions of Lemma 15.34 are fulfilled. Therefore, only the asymptotic normality must be proved. Applying the mean value theorem to $f_N(\theta, z)$ at the points $\hat\theta$, θ_* gives

$$f_N(\hat\theta, z) = f_N(\theta_*, z) + f_N'(\overset{\frown}{\theta}, z)(\hat\theta - \theta_*) \tag{15-38}$$

where $\overset{\frown}{\theta} = t\hat\theta + (1-t)\theta_*$ with $t \in [0, 1]$. Because $\overset{\frown}{\theta}$ converges w.p. 1 to θ_* at the same rate as $\hat\theta$ (proof: see (15-34)) it follows from conditions 2 and 3 that $f_N'(\overset{\frown}{\theta}, z) = f'(\theta_*) + O_\mathrm{p}(N^{-1/2})$ (Lemma 15.34). Putting this result in (15-38) taking into account conditions 1 ($f_N(\theta_*, z) - f(\theta_*) = O_\mathrm{p}(N^{-1/2})$) and 4 ($\hat\theta - \theta_* = O_\mathrm{p}(N^{-1/2})$) gives

$$\begin{aligned} f_N(\hat\theta, z) - f(\theta_*) &= \delta_N(z) + O_\mathrm{p}(N^{-1}) \\ \delta_N(z) &= f_N(\theta_*, z) - f(\theta_*) + f'(\theta_*)(\hat\theta - \theta_*) \end{aligned} \tag{15-39}$$

where $\delta_N(z) = O_\mathrm{p}(N^{-1/2})$ is asymptotically normally distributed at the rate $O(N^{-1/2})$ (a finite linear combination of asymptotically normally distributed random variables is asymptotically normally distributed and the convergence rate is preserved). Multiplying (15-39) by $\sqrt{N}$ and taking the limit gives

$$\operatorname*{plim}_{N\to\infty} \sqrt{N}(f_N(\hat{\theta}, z) - f(\theta_*) - \delta_N(z)) = 0 \tag{15-40}$$

Because convergence in probability implies convergence in law (see Section 14.7, interrelation 3), it follows from (15-40) that $\sqrt{N}(f_N(\hat{\theta}, z) - f(\theta_*))$ is asymptotically normally distributed at the rate $O(N^{-1/2})$. □

Corollary 15.39 (Asymptotic Distribution Function): Let $z \in \mathbb{R}^N$ be a stochastic variable. Let $f_N(\theta, \psi, \eta, z) \in \mathbb{R}$, and its first- and second-order derivatives w.r.t. θ and ψ, be jointly continuous functions of $\theta \in \boldsymbol{\Theta}_\mathrm{r}$, $\psi \in \boldsymbol{\psi}_\mathrm{r}$, and $\eta \in \boldsymbol{\eta}_\mathrm{r}$. $\boldsymbol{\Theta}_\mathrm{r}$, $\boldsymbol{\psi}_\mathrm{r}$, and $\boldsymbol{\eta}_\mathrm{r}$ are compact subsets of, respectively, $\mathbb{R}^{n_\theta}$, $\mathbb{R}^{n_\psi}$, and $\mathbb{R}^{n_\eta}$. If

1. $f_N(\theta, \psi, \eta, z)$ converges uniformly w.p. 1 (in prob.) to $f(\theta, \psi, \eta)$ in $\boldsymbol{\Theta}_\mathrm{r}$, $\boldsymbol{\psi}_\mathrm{r}$, and $\boldsymbol{\eta}_\mathrm{r}$ at the rate $O_\mathrm{p}(N^{-1/2})$ and is asymptotically normally distributed at the rate $O(N^{-1/2})$,
2. $\partial f_N(\theta, \psi, \eta, z)/\partial\theta$, $\partial f_N(\theta, \psi, \eta, z)/\partial\psi$ converge uniformly w.p. 1 (in prob.) to $\partial f(\theta, \psi, \eta)/\partial\theta$, $\partial f(\theta, \psi, \eta)/\partial\psi$ in $\boldsymbol{\Theta}_\mathrm{r}$, $\boldsymbol{\psi}_\mathrm{r}$, and $\boldsymbol{\eta}_\mathrm{r}$ at the rate $O_\mathrm{p}(N^{-1/2})$,
3. $\left\|\dfrac{\partial^2 f_N(\theta, \psi, \eta, z)}{\partial\theta^2}\right\|_2 \le O_\mathrm{a.s.}(N^0)$ $(O_\mathrm{p}(N^0))$, $\left\|\dfrac{\partial^2 f_N(\theta, \psi, \eta, z)}{\partial\psi^2}\right\|_2 \le O_\mathrm{a.s.}(N^0)$ $(O_\mathrm{p}(N^0))$ uniformly in $\boldsymbol{\Theta}_\mathrm{r}$, $\boldsymbol{\psi}_\mathrm{r}$, and $\boldsymbol{\eta}_\mathrm{r}$,
4. $\hat{\theta}$, $\hat{\psi}$ converge w.p. 1 (in prob.) to θ_*, ψ_* at the rate $O_\mathrm{p}(N^{-1/2})$, where θ_*, ψ_* are interior points of $\boldsymbol{\Theta}_\mathrm{r}$, $\boldsymbol{\psi}_\mathrm{r}$, and are asymptotically normally distributed at the rate $O(N^{-1/2})$,

then $f_N(\hat{\theta}, \hat{\psi}, \eta, z)$ converges uniformly w.p. 1 (in prob.) to $f(\theta_*, \psi_*, \eta)$ in $\boldsymbol{\eta}_\mathrm{r}$ at the rate $O_\mathrm{p}(N^{-1/2})$ and is asymptotically normally distributed at the rate $O(N^{-1/2})$. Moreover, we have

$$\begin{aligned} f_N(\hat{\theta}, \hat{\psi}, \eta, z) &= f_N(\theta_*, \psi_*, \eta, z) + \frac{\partial f_N(\theta, \psi_*, \eta)}{\partial\theta_*}(\hat{\theta} - \theta_*) \\ &\quad + \frac{\partial f_N(\theta_*, \psi, \eta)}{\partial\psi_*}(\hat{\psi} - \psi_*) + O_\mathrm{p}(N^{-1}) \end{aligned}$$

Appendix 15.D: Proof of the Convergence Rate of the Minimizer (Theorem 15.19)

The proof consists of three steps: first, the convergence rate of $V_N'(\tilde{\theta}(z_0), z)$ is studied; next, the convergence rate of $V_N''(\hat{\theta}, z)$; and finally, both results are combined to establish the convergence rate of $\hat{\theta}(z)$ to $\tilde{\theta}(z_0)$.

1. Convergence rate of $V_N'(\tilde{\theta}(z_0), z)$

   Under Assumptions 15.1 ($P = 4$) and 15.16, $V_N'(\theta, z)$ and $V_N''(\theta, z)$ converge uniformly in mean square at the rate $O_{\text{m.s.}}(N^{-1/2})$ to their expected values $V_N'(\theta)$ and $V_N''(\theta)$, respectively (Lemma 15.17)

   $$V_N'(\theta, z) = V_N'(\theta) + O_{\text{m.s.}}(N^{-1/2}) \tag{15-41}$$

   Because $\tilde{\theta}(z_0)$ is deterministic we may evaluate (15-41) at $\theta = \tilde{\theta}(z_0)$. Taking into account that $\tilde{\theta}(z_0)$ minimizes $V_N(\theta)$, this gives

   $$V_N'(\tilde{\theta}(z_0), z) = V_N'(\tilde{\theta}(z_0)) + O_{\text{m.s.}}(N^{-1/2}) = O_{\text{m.s.}}(N^{-1/2}) \tag{15-42}$$

2. Convergence rate of $V_N''(\overset{\frown}{\theta}, z)$

   Here, $\overset{\frown}{\theta}$ is a random variable (see (15-10)) so that the reasoning applied to $V_N'(\tilde{\theta}(z_0), z)$ does not hold for $V_N''(\overset{\frown}{\theta}, z)$. Indeed, the convergence rate of $V_N''(\overset{\frown}{\theta}, z)$ depends on the stochastic properties of $\overset{\frown}{\theta}$. Under Assumptions 15.1 ($P = 4$), 15.2(a), and 15.18, $\hat{\theta}(z)$ converges in prob. to $\tilde{\theta}(z_0)$ (Theorem 15.6 without Assumption 15.2(b) shows convergence in prob., see Exercise 15.7). From (15-10) it follows also that $\overset{\frown}{\theta}$ converges in probability to $\tilde{\theta}(z_0)$

   $$\underset{N\to\infty}{\text{a.s.lim}}(\overset{\frown}{\theta} - \tilde{\theta}(z_0)) = (\underset{N\to\infty}{\text{a.s.lim}}\, t)(\underset{N\to\infty}{\text{a.s.lim}}(\hat{\theta}(z) - \tilde{\theta}(z_0))) = 0 \tag{15-43}$$

   Hence, $V_N''(\overset{\frown}{\theta}, z)$ converges in probability to $V_N''(\tilde{\theta}(z_0))$ (see Appendix 15.C, Lemma 15.31)

   $$V_N''(\overset{\frown}{\theta}, z) = V_N''(\tilde{\theta}(z_0)) + o_p(N^0) \tag{15-44}$$

3. Convergence rate of $\hat{\theta}(z)$ to $\tilde{\theta}(z_0)$

   Because convergence in mean square and almost sure convergence imply convergence in probability (see Section 14.7, interrelations 1 and 2) and a continuous function and the limit in probability may be interchanged (see Section 14.8, property 3), it follows from (15-11), (15-42), and (15-44) that

   $$\begin{aligned}\hat{\theta}(z) - \tilde{\theta}(z_0) &= (V_N''(\tilde{\theta}(z_0)) + o_p(N^0))^{-1} O_p(N^{-1/2}) \\ &= (V_N''^{-1}(\tilde{\theta}(z_0)) + o_p(N^0)) O_p(N^{-1/2})\end{aligned}$$

   Under Assumption 15.18, $V_N''^{-1}(\tilde{\theta}(z_0))$ is an $O(N^0)$, so that $\hat{\theta}(z) - \tilde{\theta}(z_0) = O_p(N^{-1/2})$. □

Appendix 15.E: Proof of the Improved Convergence Rate of the Minimizer (Theorem 15.21)

Using Assumption 15.20 and the result of Theorem 15.19, the convergence rate of $V_N''(\overset{\frown}{\theta}, z)$ will be established. Combined with the convergence rate of $V_N'(\tilde{\theta}(z_0), z)$ (see Appendix 15.D), this will lead to a refined expression for $\hat{\theta}(z) - \tilde{\theta}(z_0)$.

To establish the convergence rate of $V_N''(\widehat{\theta}, z)$, we verify that all the conditions of Lemma 15.34 (see Appendix 15.C) are satisfied. Consistent with the assumptions of Lemma 15.17, $V_N''(\theta, z)$ converges uniformly in mean square to $V_N''(\theta)$ in Θ_r at the rate $O_{\text{m.s.}}(N^{-1/2})$, which implies $O_p(N^{-1/2})$ (condition 1, Lemma 15.34). From (15-10) it follows that $\widehat{\theta} - \tilde{\theta}(z_0) = t(\hat{\theta}(z) - \tilde{\theta}(z_0))$ with $t \in [0, 1]$. Hence, $\widehat{\theta} - \tilde{\theta}(z_0)$ converges w.p. 1 to zero at the rate of $\hat{\theta}(z) - \tilde{\theta}(z_0)$, which is given by Theorem 15.19, and, hence, $\widehat{\theta} - \tilde{\theta}(z_0) = O_p(N^{-1/2})$ (condition 3, Lemma 15.34). We will now show under Assumptions 15.1 ($p = 4$), 15.18, and 15.20 that $V_N'''(\theta, z)$ is an $O_{\text{a.s.}}(N^0)$ uniformly in Θ_r (condition 2, Lemma 15.34). The absolute value of the third-order derivative of the cost function $V_N(\theta, z)$ is bounded by

$$\left|\frac{\partial^3 V_N(\theta, z)}{\partial\theta_{[i]}\partial\theta_{[j]}\partial\theta_{[k]}}\right| \leq \frac{\|z\|_2^2}{N}\left\|\frac{\partial^3 W_N(\theta)}{\partial\theta_{[i]}\partial\theta_{[j]}\partial\theta_{[k]}}\right\|_2 \quad \text{for } i, j, k = 1, 2, \ldots, n_\theta \tag{15-45}$$

Assumption 15.20 guarantees that the second factor in the right-hand side of (15-45) is an $O(N^0)$ uniformly in Θ_r. Under Assumption 15.1 ($P = 4$), $\|z\|_2^2/N$ obeys the strong law of large numbers for mixing sequences (see Section 14.9, version 3 of (14-69)). Under Assumptions 15.1 ($P = 2$) the expected value of $\|z\|_2^2/N$ is an $O(N^0)$ so that $\|z\|_2^2/N = O_{\text{a.s.}}(N^0)$. This shows that the right-hand side of (15-45) is an $O_{\text{a.s.}}(N^0)$. The three conditions of Lemma 15.34 (see Appendix 15.C) are satisfied and, hence,

$$V_N''(\widehat{\theta}, z) = V_N''(\tilde{\theta}(z_0)) + O_p(N^{-1/2}) \tag{15-46}$$

Following the same lines as in the third step in the proof of Theorem 15.19 (see Appendix 15.D) we conclude from (15-11), (15-42), and (15-46) that

$$\begin{aligned}\hat{\theta}(z) - \tilde{\theta}(z_0) &= -(V_N''^{-1}(\tilde{\theta}(z_0)) + O_p(N^{-1/2}))V_N'^T(\tilde{\theta}(z_0), z)\\ &= \delta_\theta(z) + O_p(N^{-1})\end{aligned} \tag{15-47}$$

with $\delta_\theta(z) = -V_N''^{-1}(\tilde{\theta}(z_0))V_N'^T(\tilde{\theta}(z_0), z) = O_p(N^{-1/2})$. As $\mathcal{E}\{V_N'(\theta, z)\} = V_N'(\theta)$ in Θ_r and $V_N'(\tilde{\theta}(z_0)) = 0$, we have that $\mathcal{E}\{\delta_\theta(z)\} = 0$, which concludes the proof. □

Appendix 15.F: Equivalence between the Truncated and the Original Minimizer (Lemma 15.27)

Define $\mathbb{Z}_L$ as the set of realizations z for which the estimate $\hat{\theta}(z)$ lies within the hyperball with center $\tilde{\theta}(z_0)$ and radius L

$$\mathbb{Z}_L = \{z \mid \|\hat{\theta}(z) - \tilde{\theta}(z_0)\|_2 \leq L\}$$

Under the assumptions of Theorem 15.6 or Theorem 15.8, $\hat{\theta}(z)$ converges strongly to $\tilde{\theta}(z_0)$: there exists an N_0 such that for any $N \geq N_0$, $\|\hat{\theta}(z) - \tilde{\theta}(z_0)\|_2 \leq L$ w.p. 1. Hence, for $N \geq N_0$, the realizations $z \in \mathbb{Z}_L$ happen with probability measure 1. From the definition of $\underline{\hat{\theta}}(z)$ (15-17) we conclude that $\underline{\hat{\theta}}(z) = \hat{\theta}(z)$ w.p. 1 for any $N \geq N_0$.

To prove that the results of Corollaries 15.25 and 15.26 are valid, we still need to show that the expected value of $\delta_\theta(z)$ over the set $\mathbb{Z}_L$ is zero. Using $\mathcal{E}\{\delta_\theta(z)\} = 0$, we get

$$\mathcal{E}\{\delta_\theta(z) | z \in \mathbb{Z}_L\} = -\mathcal{E}\{\delta_\theta(z) | z \notin \mathbb{Z}_L\} = 0$$

where the last equality is due to the fact that $\text{Prob}(z \notin \mathbb{Z}_L) = 0$ for $N \geq N_0$ and that the second-order moments of z are uniformly bounded. □

Appendix 15.G: Proof of the Asymptotic Bias on the Truncated Minimizer (Theorem 15.28)

Define $\mathbb{Z}_L$ as in the proof of Lemma 15.27 with $\tilde{\theta}(z_0) = \theta_0$ (see Appendix 15.F). Under the assumptions of Corollary 15.25 (Corollary 15.26) and Eq. (15-17), it follows for any $z \in \mathbb{Z}_L$, that $\hat{\underline{\theta}}(z) = \hat{\theta}(z) = \theta_0 + \delta_\theta(z) + b_\theta(z)$ and that $b_\theta(z)$ is uniformly bounded. The interchangeability property of the expected value and the limit in probability for uniformly bounded random variables (see Section 14.8, (14-63)) guarantees that the expected value of $b_\theta(z)$ taken over $\mathbb{Z}_L$, $\mathcal{E}\{b_\theta(z) | z \in \mathbb{Z}_L\}$, is an $O(N^{-1})$ $(O(N^{-1}) + O(N^{-K}))$. From Lemma 15.27 it follows that there exists an N_0 such that for any $N \geq N_0$, $\mathcal{E}\{\delta_\theta(z) | z \in \mathbb{Z}_L\} = 0$. This shows that $\mathcal{E}\{b_\theta(z) | z \in \mathbb{Z}_L\}$ is the bias error of $\hat{\underline{\theta}}(z)$ for $N \geq N_0$.

If we can show that the derivative of the bias w.r.t. θ_0 is continuous for all $\theta_0 \in \Theta_r$, then it is also uniformly bounded in the compact set Θ_r, and, hence, it behaves as an $O(N^{-1})$ $(O(N^{-1}) + O(N^{-K}))$. Therefore, it is sufficient to show that the derivative of $\hat{\underline{\theta}}(z)$ w.r.t. θ_0 is continuous in Θ_r. Consider Eq. (15-9) with $\tilde{\theta}(z_0)$ replaced by θ_0, giving

$$V_N'(\hat{\theta}(z), z) = V_N'(\theta_0, z) + (\hat{\theta}(z) - \theta_0)^T V_N''(\overset{\frown}{\theta}, z) \tag{15-48}$$

where $\overset{\frown}{\theta} = t\hat{\theta}(z) + (1-t)\theta_0$ with $t \in [0, 1]$. Because $V_N(\theta, z)$ has continuous first-, second-, and third-order derivatives for all $\theta \in \Theta_r$ (Assumptions 15.16 and 15.20), it follows from the implicit function theorem (Kaplan, 1993) that $\overset{\frown}{\theta}$ is a continuous function of $\hat{\theta}(z)$ and θ_0 with continuous partial derivatives. Putting $\overset{\frown}{\theta} = g(\hat{\theta}(z), \theta_0)$ in (15-48), taking into account that $V_N'(\hat{\theta}(z), z) = 0$, gives

$$0 = V_N'(\theta_0, z) + (\hat{\theta}(z) - \theta_0)^T V_N''(g(\hat{\theta}(z), \theta_0), z) \tag{15-49}$$

(15-49) defines $\hat{\theta}(z)$ implicitly as a function of θ_0. Applying, again, the implicit function theorem shows that $\partial\hat{\theta}(z)/\partial\theta$ is continuous in Θ_r. By definition of $\hat{\underline{\theta}}(z)$ (15-17), this is also true for $\partial\hat{\underline{\theta}}(z)/\partial\theta$. □

Appendix 15.H: Cumulants of the Partial Sum of a Mixing Sequence

Let $S(N) = \sum_{t=1}^{N} x_N(t)$ with $x_N(t)$ mixing over t of order P for $N = 1, 2, \ldots, \infty$. The kth order cumulant of $S(N)$ is an $O(N)$, $k = 1, 2, \ldots, P$.

Proof. The kth order cumulant C_k of $S(N)$ is given by

$$C_k = \sum_{t_1, t_2, \ldots, t_k = 1}^{N} \operatorname{cum}(x_N(t_1), x_N(t_2), \ldots, x_N(t_k)) \tag{15-50}$$

Applying (14-36) to (15-50) gives $|C_k| = O(N)$. □

Appendix 15.I: Proof of the Asymptotic Distribution of the Minimizer (Theorem 15.29)

From Theorem 15.21, it follows that $\sqrt{N}(\hat{\theta}(z) - \tilde{\theta}(z_0))$ converges in probability and, hence, also in law (see Section 14.7, interrelation 3) to $\sqrt{N}\delta_\theta(z)$. According to (15-14) the stochastic part of $\delta_\theta(z)$ is given by $V_N'^T(\tilde{\theta}(z_0), z)$. Since the matrix dimensions of the Hessian of the cost function are independent of N, we can study the stochastic behavior of $V_N'^T(\tilde{\theta}(z_0), z)$ separately from $V_N''^{-1}(\tilde{\theta}(z_0))$. We will show that the cumulants of $\sqrt{N}V_N'^T(\tilde{\theta}(z_0), z)$ tend to those of a Gaussian random variable. Because a normal distribution is uniquely determined by its moments, it follows from the Fréchet-Shohat Lemma 14.11 that $\sqrt{N}V_N'^T(\tilde{\theta}(z_0), z)$ converges in law to a Gaussian random variable. A linear combination of a finite number (independent of N) of Gaussian random variables is also a Gaussian random variable, which shows that $\sqrt{N}\delta_\theta(z)$ is asymptotically normally distributed. Expression (15-18) for the covariance matrix follows directly from the definition of $\delta_\theta(z)$ (15-14) and the fact that $\mathcal{E}\{\delta_\theta(z)\} = 0$.

Under Assumptions 15.1 ($P = 2K$), 15.2, and 15.16, the derivative of the cost function can be written for any $\theta \in \Theta_r$ as

$$\frac{\partial V_N(\theta, z)}{\partial \theta_{[i]}} = \frac{1}{N}\sum_{t=1}^{N} x_N(t) \tag{15-51}$$

$i = 1, 2, \ldots, n_\theta$, with $x_N(t)$ mixing over t of order K, $N = 1, 2, \ldots, \infty$ (proof: similar to that of $V_N(\theta, z)$ in Appendix 15.A). Hence, the kth order joint cumulant of $\sqrt{N}V_N'^T(\tilde{\theta}(z_0), z)$ is an $O(N^{1-k/2})$, $k = 1, 2, \ldots, K$ (see Appendix 15.H). Under Assumption 15.1 ($P = \infty$) this is valid for $K = 1, 2, \ldots, \infty$. It shows that the covariance matrix (second-order cumulant) is an $O(N^0)$ and that all the joint cumulants of order $k = 3, 4, \ldots, \infty$ are asymptotically ($N \to \infty$) zero. This concludes the proof because the joint cumulants, of order 3 and larger, of a Gaussian random variable are zero (see Example 14.2).

The proof of the convergence rate follows the lines of Appendix 14.H. □

Appendix 15.J: Proof of the Existence and the Convergence of the Covariance Matrix of the Truncated Minimizer (Theorem 15.30)

Lemma 15.27 is valid under the assumptions of Theorem 15.21. It states that there exists an N_0 such that for any $N \geq N_0$, $\underline{\hat{\theta}}(z) = \tilde{\theta}(z_0) + \delta_\theta(z) + b_\theta(z)$ w.p. 1. Because $\underline{\hat{\theta}}(z)$, defined by (15-17), is uniformly bounded, its expected value and covariance matrix exist. The same is true for $b_\theta(z)$ for all realizations $z \in \mathbb{Z}_L$, where $\mathbb{Z}_L$ is defined in Appendix 15.F. Taking into account that for all $N \geq N_0$, $\mathcal{E}\{\delta_\theta(z) | z \in \mathbb{Z}_L\} = 0$, and $\mathcal{E}\{b_\theta(z) | z \in \mathbb{Z}_L\} = O(N^{-1})$ (proof: similar to Appendix 15.G), we find

$$\sqrt{N}\hat{\underline{\theta}}(z) - \mathcal{E}\{\sqrt{N}\hat{\underline{\theta}}(z) \,|\, z \in \mathbb{Z}_L\} = \sqrt{N}\delta_\theta(z) + O_\mathrm{p}(N^{-1/2}) \tag{15-52}$$

where $O_\mathrm{p}(N^{-1/2})$ is a uniformly bounded random variable. Calculating the covariance matrix of (15-52), taking into account that by definition of $\hat{\underline{\theta}}(z)$

$$\mathrm{Cov}(\sqrt{N}\hat{\underline{\theta}}(z)) = \mathrm{Cov}(\sqrt{N}\hat{\underline{\theta}}(z) \,|\, z \in \mathbb{Z}_L)$$

gives

$$\mathrm{Cov}(\sqrt{N}\hat{\underline{\theta}}(z)) = \mathrm{Cov}(\sqrt{N}\delta_\theta(z) \,|\, z \in \mathbb{Z}_L) + O(N^{-1/2}) \tag{15-53}$$

Because $\mathrm{Prob}(z \notin \mathbb{Z}_L) = 0$ for $N \geq N_0$ (see Appendix 15.F) and the fourth-order moments of z are uniformly bounded, we have

$$\mathrm{Cov}(\sqrt{N}\delta_\theta(z) \,|\, z \in \mathbb{Z}_L) = \mathrm{Cov}(\sqrt{N}\delta_\theta(z)) \tag{15-54}$$

Combining (15-53) and (15-54) proves that $\mathrm{Cov}(\sqrt{N}\hat{\underline{\theta}}(z))$ converges to $\mathrm{Cov}(\sqrt{N}\delta_\theta(z))$ at the rate $O(N^{-1/2})$. □

16

Properties of Least Squares Estimators with Stochastic Weighting

Abstract: This chapter studies the asymptotic stochastic properties (strong convergence, strong consistency, convergence rate, asymptotic bias, and asymptotic normality) of nonlinear least squares estimators with a stochastic weighting. The presented theory is applicable to, for example, (total) least squares estimators using nonparametric noise models. Because this chapter relies strongly on the results of Chapter 15, it cannot be read independently of that chapter. Readers who are unfamiliar with the analysis of the stochastic properties of estimators should, in addition, first read Sections 14.11 to 14.13.

16.1 INTRODUCTION—NOTATIONAL CONVENTIONS

In this chapter we consider the identification of a parametric plant and/or noise model $M(\theta, z_0, n_z)$ through the minimization of a weighted nonlinear least square cost function

$$V_N(\theta, z) = \frac{1}{N} z^T W_N(\theta, \eta(z), w(\theta, \eta(z), z))z \tag{16-1}$$

with $z \in \mathbb{R}^N$ the noisy measurements, $z = z_0 + n_z$, $\theta \in \mathbb{R}^{n_\theta}$ the plant model parameters, and $\eta(z) \in \mathbb{R}^q$ a stochastic vector. $w(\theta, \eta(z), z) \in \mathbb{R}^p$ is the vector of the stochastic sums that average the noisy measurements z. $W_N \in \mathbb{R}^{N \times N}$ is a stochastic positive semidefinite weighting matrix depending on θ, $\eta(z)$, and $w(\theta, \eta(z), z)$. n_θ, p and q are N-independent integers. Just as in Chapter 15, we can assume without any loss of generality that the weighting matrix W_N is symmetric. We will often rewrite (16-1) as

$$V_N(\theta, z) = f_N(\theta, \eta(z), w(\theta, \eta(z), z), z) \tag{16-2}$$

and denote the function by

$$f_N(\theta, \eta, w, z) = \frac{1}{N} z^T W_N(\theta, \eta, w)z \tag{16-3}$$

System Identification: A Frequency Domain Approach. Rik Pintelon and Johan Schoukens
Copyright © 2001 Institute of Electrical and Electronics Engineers. ISBN: 0-780-36000-1

where the stochastic vectors $\eta(z)$, $w(\theta, \eta(z), z)$ have been replaced by the deterministic vectors η, w. Note that (16-3) is a nonlinear least squares cost function with deterministic weighting. Similarly to Θ_r (see Section 15.1), we define $\mathbb{W} \subset \mathbb{R}^p$ as a compact (closed and bounded) set of w values for which the cost function (16-3) and/or its higher order derivatives w.r.t. w exist and are finite. From (16-2) and (16-3) it follows that it is not possible to put (16-1) within the framework (14-95) of Section 14.13, even without $\eta(z)$. The reason for this is that (16-1) is a stochastic sum which depends, itself, on other stochastic sums $w(\theta, \eta(z), z)$ and a stochastic vector $\eta(z)$.

Following the same lines as in Chapter 15, the asymptotic ($N \to \infty$) properties (strong convergence, strong consistency, convergence rate, asymptotic bias, and asymptotic normality) of the (set of) minimizer(s)

$$\hat{\theta}(z) = \arg\min_{\theta \in \Theta_r} V_N(\theta, z) \tag{16-4}$$

will be analyzed (Θ_r, see Section 15.1, is a compact set of parameters where the cost function and/or its higher order derivatives exist and are finite). It is clear that the asymptotic properties of the minimizer (16-4) strongly depend on the stochastic behavior of $\eta(z)$ and $w(\theta, \eta(z), z)$. Assumptions similar to those of Chapter 15 guarantee the stochastic properties of (16-3). The main difference from Chapter 15 is that additional assumptions concerning $\eta(z)$ and $w(\theta, \eta(z), z)$ have to be made to ensure that (16-4) has asymptotic properties similar to those of the nonlinear least squares estimator with deterministic weighting (15-3). The analysis of the nonlinear least squares estimator with stochastic weighting (16-4) relies heavily on the stochastic properties of a converging sequence of functions that also depends on some converging random vector(s). Therefore, it is recommended to read Appendix 15.C first, before going through the proofs of this chapter. The chapter ends with an overview of the asymptotic properties of $\hat{\theta}(z)$.

16.2 STRONG CONVERGENCE

The strong convergence of the minimizer is a direct consequence of the strong convergence of the cost function (see Section 15.2). The cost function (16-1) can converge only if $\eta(z)$ tends to some nonrandom number and if $w(\theta, \eta(z), z)$ tends to some deterministic function of θ. Therefore, it is natural to make the following assumptions (see Lemma 15.31, Appendix 15.C).

Assumption 16.1 (Strong Convergence $\eta(z)$): The stochastic vector $\eta(z)$ converges w.p. 1 to a nonrandom value η_*.

Define $\mathbb{H}_\varepsilon$ as a compact set of η values in the neighborhood of η_*

$$\mathbb{H}_\varepsilon = \{\eta \in \mathbb{R}^q \,|\, \|\eta - \eta_*\|_2 \leq \varepsilon\} \tag{16-5}$$

and let $w(\theta, \eta, z)$ denote the stochastic sums, where the stochastic vector $\eta(z)$ is replaced by the deterministic vector η. Although the strong convergence of the minimizer puts conditions only on the convergence of $w(\theta, \eta(z), z)$, the assumption will be stated more generally for the kth order derivative w.r.t. θ.

Assumption 16.2 (Strong or Weak Convergence $w^{(k)}(\theta, \eta, z)$): $w^{(k)}(\theta, \eta, z)$, the kth order derivative of $w(\theta, \eta, z)$ w.r.t. θ, converges uniformly w.p. 1 (in prob.) to $\mathcal{E}\{w^{(k)}(\theta, \eta, z)\} = w^{(k)}(\theta, \eta)$, in Θ_r, $\mathbb{H}_\varepsilon$. For any N, ∞ included, $w^{(k)}(\theta, \eta, z)$ is a jointly continuous function of θ, η in Θ_r, $\mathbb{H}_\varepsilon$.

16.2.1 Strong Convergence of the Cost Function

Assumption 16.3 (Constraints on the Cost Function): (a) The weighting matrix $W_N(\theta, \eta, w)$ is a symmetric positive semidefinite matrix, satisfying $\|W_N(\theta, \eta, w)\|_1 \le c < \infty$, with c an N-independent constant, for all N (∞ included) and all $\theta \in \mathbb{\Theta}_{\mathrm{r}}$, $\eta \in \mathbb{\eta}_\varepsilon$, $w \in \mathbb{W}$. $W_N(\theta, \eta, w)$ is a jointly continuous matrix function of θ, η, w in the compact sets $\mathbb{\Theta}_{\mathrm{r}}$, $\mathbb{\eta}_\varepsilon$, $\mathbb{W}$. (b) There exists an N_0 such that for any $r \ge s \ge N_0$, $\|W_{r[1:s,\,1:s]}(\theta, \eta, w) - W_s(\theta, \eta, w)\|_1^2 = O((r-s)/r)$ in $\mathbb{\Theta}_{\mathrm{r}}$, $\mathbb{\eta}_\varepsilon$, $\mathbb{W}$.

Condition (b) is necessary to ensure the strong convergence of the cost function. If it is not fulfilled, then all the lemmas and theorems of this chapter remain valid except that the strong convergence (convergence w.p. 1) must be replaced everywhere by weak convergence (convergence in prob.).

Lemma 16.4 (Strong Convergence of the Cost Function): Under Assumptions 15.1 $(P = 4)$, 16.1, 16.2 (w.p. 1, $k = 0$), and 16.3 the cost function $V_N(\theta, z)$ converges uniformly w.p. 1 to

$$V_N(\theta) = \mathcal{E}\{f_N(\theta, \eta_*, w(\theta, \eta_*), z)\} = f_N(\theta, \eta_*, w(\theta, \eta_*)) \tag{16-6}$$

in the compact set $\mathbb{\Theta}_{\mathrm{r}}$. $V_N(\theta, z)$ and $V_N(\theta)$ are continuous functions of θ in $\mathbb{\Theta}_{\mathrm{r}}$.

Proof. See Appendix 16.A. □

Note that $V_N(\theta) = f_N(\theta, \eta_*, w(\theta, \eta_*))$ is obtained as follows: first replace the stochastic weighting $W_N(\theta, \eta(z), w(\theta, \eta(z), z))$ in (16-1) by the deterministic weighting $W_N(\theta, \eta_*, w(\theta, \eta_*))$, and next, take the expected value. This shows that under Assumptions 16.1 and 16.2 (w. p. 1), the stochastic behavior of the cost function with stochastic weighting is similar to that with deterministic weighting.

16.2.2 Strong Convergence of the Minimizer

Using definition (16-6) of $V_N(\theta)$, the theorems of Section 15.2.2 (strong convergence of the minimizer of nonlinear least squares cost functions with deterministic weighting) remain valid under Assumptions 16.1, 16.2 (w.p. 1, $k = 0$).

Theorem 16.5 (Strong Convergence of the Minimizer): Under Assumptions 15.1 $(P = 4)$, 16.1, 16.2 (w.p. 1, $k = 0$), 16.3, and 15.5 the minimizer(s) $\hat{\theta}(z)$ converge(s) w.p. 1 to

$$\tilde{\theta}(z_0) = \arg\min_{\theta \in \mathbb{\Theta}_{\mathrm{r}}} f_N(\theta, \eta_*, w(\theta, \eta_*)) \tag{16-7}$$

Proof. Similar to Theorem 15.6 (see Appendix 15.B). □

Theorem 16.6 (Strong Convergence of the Minimizer): Under Assumptions 15.1 $(P = 4)$, 16.1, 16.2 (w.p. 1, $k = 0$), 16.3, 15.4, and 15.7, $\tilde{\theta}(z_0)$ converges to θ_* and $\hat{\theta}(z)$ converges strongly θ_*, with

$$\theta_* = \arg\min_{\theta \in \mathbb{\Theta}_{\mathrm{r}}} V_*(\theta) \text{ and } V_*(\theta) = \lim_{N \to \infty} f_N(\theta, \eta_*, w(\theta, \eta_*)) \tag{16-8}$$

Proof. Similar to Theorem 15.8. □

16.3 STRONG CONSISTENCY

The cost function (16-6) can be written as

$$V_N(\theta) = \frac{1}{N}\mathcal{E}\{z_0^T W_N(\theta, \eta_*, w(\theta, \eta_*))z_0\} + \frac{1}{N}\mathrm{trace}(W_N(\theta, \eta_*, w(\theta, \eta_*))C_{n_z}) \qquad (16\text{-}9)$$

and equals $V_N(\theta)$ in (15-8) where $W_N(\theta)$ is replaced by $W_N(\theta, \eta_*, w(\theta, \eta_*))$. Hence, replacing $W_N(\theta)$ by $W_N(\theta, \eta_*, w(\theta, \eta_*))$ in the assumptions of Section 15.3 shows that the theorems of Section 15.3 (strong consistency of the minimizer of nonlinear least squares cost functions with deterministic weighting) remain valid under Assumptions 16.1, 16.2 (w.p. 1, $k = 0$). We cite them in order of reduced conditions on $V_N(\theta)$.

Theorem 16.7 (Strong Consistency): Under the assumptions of Theorem 16.5 and Assumptions 15.9 and 15.10, the estimate $\hat{\theta}(z)$ converges w.p. 1 to θ_0.

Proof. It follows directly from Theorem 16.5 and Assumptions 15.9 and 15.10. □

Theorem 16.8 (Strong Consistency): Under the assumptions of Theorem 16.5 and Assumptions 15.9 and 15.12, the estimate $\hat{\theta}(z)$ converges w.p. 1 to θ_0.

Proof. It follows directly from Theorem 16.5 and Assumptions 15.9 and 15.12. □

Theorem 16.9 (Strong Consistency): Under the assumptions of Theorem 16.6 and Assumptions 15.9 and 15.14, the estimate $\hat{\theta}(z)$ converges w.p. 1 to θ_0.

Proof. It follows directly from Theorem 16.6 and Assumptions 15.9 and 15.14. □

16.4 CONVERGENCE RATE

The convergence rate of the minimizer is a direct consequence of the convergence rate of the first- and second-order derivatives of the cost function w.r.t. θ (see Section 15.4). These derivatives can be written as

$$\begin{aligned} V_N'^T(\theta, z) &= g_N(\theta, \eta(z), w(\theta, \eta(z), z), w'(\theta, \eta(z), z), z) \\ V_N''(\theta, z) &= h_N(\theta, \eta(z), w(\theta, \eta(z), z), w'(\theta, \eta(z), z), w''(\theta, \eta(z), z), z) \end{aligned} \qquad (16\text{-}10)$$

From (16-10) it follows that the convergence rate of the derivatives of the cost function is influenced by the convergence rates of $\eta(z)$ and $w^{(k)}(\theta, \eta(z), z)$, $k = 0, 1, 2$. Therefore, it is natural to make the following assumptions (see Lemma 15.34, Appendix 15.C).

Assumption 16.10 (Convergence Rate $\eta(z)$): The convergence rate in probability of $\eta(z)$ equals $O_p(N^{-1/2})$: $\eta(z) - \eta_* = O_p(N^{-1/2})$.

Assumption 16.11 (Convergence Rate $w^{(k)}(\theta, \eta, z)$): The convergence rate in probability of the kth order derivative $w^{(k)}(\theta, \eta, z)$ equals $O_p(N^{-1/2})$ uniformly in $\mathbb{\Theta}_r$, $\mathbb{\eta}_\varepsilon$: $w^{(k)}(\theta, \eta, z) - w^{(k)}(\theta, \eta) = O_p(N^{-1/2})$.

Assumption 16.12 (Constraint on the Derivative of $w^{(k)}(\theta, \eta, z)$): The derivative of $w^{(k)}(\theta, \eta, z)$ w.r.t. η satisfies

$$\left\|\frac{\partial}{\partial\eta}\frac{\partial^k w(\theta, \eta, z)}{\partial\theta_{[i_1]}\partial\theta_{[i_1]}\ldots\partial\theta_{[i_k]}}\right\|_2 \leq O_{\mathrm{p}}(N^0)$$

uniformly in $\mathbb{\Theta}_{\mathrm{r}}$, $\mathbb{\eta}_{\varepsilon}$, for $i_1, i_2, \ldots, i_k = 1, 2, \ldots, n_\theta$ and for any N (∞ included).

We discuss only the convergence rate of $\hat{\theta}(z)$ to $\tilde{\theta}(z_0)$. The reader is referred to Section 15.4.3 for a discussion of the convergence rate of $\tilde{\theta}(z_0)$ to θ_*. The first step in the analysis of the convergence rate of $\hat{\theta}(z)$ is the weak convergence of the derivatives (16-10) of the cost function.

16.4.1 Convergence of the Derivatives of the Cost Function

The derivatives of the cost function (16-10) can be written more explicitly as

$$\begin{aligned} g_{N[i]}(\theta, \eta(z), w(\theta, \eta(z), z), w'(\theta, \eta(z), z), z) &= \frac{1}{N}z^T\frac{dW_N(\theta, \eta(z), w(\theta, \eta(z), z))}{d\theta_{[i]}}z \\ h_{N[i,\,j]}(\theta, \eta(z), w(\theta, \eta(z), z), w'(\theta, \eta(z), z), w''(\theta, \eta(z), z), z) &= \frac{1}{N}z^T\frac{d^2W_N(\theta, \eta(z), w(\theta, \eta(z), z))}{d\theta_{[i]}d\theta_{[j]}}z \end{aligned} \quad (16\text{-}11)$$

for $i, j = 1, 2, \ldots, n_\theta$, with

$$\begin{aligned} \frac{dW_N(\theta, \eta(z), w(\theta, \eta(z), z))}{d\theta_{[i]}} &= W_{1i}(\theta, \eta(z), w(\theta, \eta(z), z), w'(\theta, \eta(z), z)) \\ \frac{d^2W_N(\theta, \eta(z), w(\theta, \eta(z), z))}{d\theta_{[i]}d\theta_{[j]}} &= W_{2ij}(\theta, \eta(z), w(\theta, \eta(z), z), w'(\theta, \eta(z), z), w''(\theta, \eta(z), z)) \end{aligned} \quad (16\text{-}12)$$

Similarly to (16-3), we will denote the functions (16-10) by

$$g_N(\theta, \eta, w, w_1, z) \text{ and } h_N(\theta, \eta, w, w_1, w_2, z) \quad (16\text{-}13)$$

where the random variables $\eta(z)$ and $w^{(k)}(\theta, \eta(z), z)$, $k = 0, 1, 2$, have been replaced by the deterministic variables η, w, w_1, w_2. Define $\mathbb{W}_1 \subset \mathbb{R}^{p \times n_\theta}$ and $\mathbb{W}_2 \subset \mathbb{R}^{p \times n_\theta \times n_\theta}$ as the compact sets of respectively w_1 and w_2 values for which the functions (16-13) and/or their higher order derivatives w.r.t. w_1 and w_2 exist and are finite.

The proof of the convergence in probability follows the same lines as for Lemma 16.4. Therefore, the following assumptions concerning W_{1i}, W_{2ij} must be made.

Assumption 16.13 (Constraints on the Derivatives of the Cost Function): The matrices $W_{1i}(\theta, \eta, w, w_1)$ and $W_{2ij}(\theta, \eta, w, w_1, w_2)$ have bounded 1-norm

$$\|W_{1i}(\theta, \eta, w, w_1)\|_1 \leq c_1 < \infty, \ \|W_{2ij}(\theta, \eta, w, w_1, w_2)\|_1 \leq c_2 < \infty$$

for $i, j = 1, 2, \ldots, n_\theta$, and are jointly continuous functions of θ, η, w, w_1, w_2 in the compact sets $\mathbb{\Theta}_{\mathrm{r}}$, $\mathbb{\eta}_\varepsilon$, $\mathbb{W}$, $\mathbb{W}_1$, $\mathbb{W}_2$ for any N (∞ included). c_1, c_2 are N-independent constants.

Lemma 16.14 (Convergence of the Derivatives of the Cost Function): Under Assumptions 15.1 ($P = 4$), 16.1, 16.2 (in prob., $k = 0, 1, 2$), and 16.13, the derivatives of the cost function $V_N'(\theta, z)$ and $V_N''(\theta, z)$ converge uniformly in probability, respectively, to

$$V_N'^T(\theta) = \mathcal{E}\{g_N(\theta, \eta_*, w(\theta, \eta_*), w'(\theta, \eta_*), z)\} = g_N(\theta, \eta_*, w(\theta, \eta_*), w'(\theta, \eta_*))$$

$$V_N''(\theta) = \mathcal{E}\{h_N(\theta, \eta_*, w(\theta, \eta_*), w'(\theta, \eta_*), w''(\theta, \eta_*), z)\} = h_N(\theta, \eta_*, w(\theta, \eta_*), w'(\theta, \eta_*), w''(\theta, \eta_*))$$

in the compact set $\mathbb{\Theta}_{\mathrm{r}}$. $V_N'(\theta, z)$ and $V_N''(\theta, z)$ are continuous functions of θ in $\mathbb{\Theta}_{\mathrm{r}}$.

Proof. Similar to Lemma 16.4. □

16.4.2 Convergence Rate of $\hat{\theta}(z)$ to $\tilde{\theta}(z_0)$

Assumption 16.15 (Constraint on Derivative $g_N(\theta, \eta, w, w_1)$ w.r.t. η, w, w_1): The weighting matrices in $g_N(\theta, \eta, w, w_1)$ have continuous first-order derivatives satisfying

$$\left\| \frac{\partial W_{1i}(\theta, \eta, w, w_1)}{\partial x_{[j]}} \right\|_2 \leq c < \infty \tag{16-14}$$

uniformly in $\mathbb{\Theta}_{\mathrm{r}}$, $\mathbb{\eta}_\varepsilon$, $\mathbb{W}$, and $\mathbb{W}_1$, for $i = 1, 2, \ldots, n_\theta$, $j = 1, 2, \ldots, \dim(x)$, and for any N (∞ included). x is a vector that contains all the elements of η, w, w_1 ($\dim(x) = q + p + pn_\theta$), and c is an N-independent constant.

Theorem 16.16 (Convergence Rate of $\hat{\theta}(z)$ to $\tilde{\theta}(z_0)$): Under Assumptions 15.1 ($P = 4$), 15.18, 16.1, 16.2 (w.p. 1, $k = 0$; in prob., $k = 1, 2$), 16.3(a), 16.10, 16.11 ($k = 0, 1$), 16.12 ($k = 0, 1$), 16.13, and 16.15 the convergence rate of $\hat{\theta}(z)$ to $\tilde{\theta}(z_0)$ equals $O_{\mathrm{p}}(N^{-1/2})$: $\hat{\theta}(z) - \tilde{\theta}(z_0) = O_{\mathrm{p}}(N^{-1/2})$.

Proof. See Appendix 16.B. □

An additional assumption on the third-order derivatives of the cost function allows the refinement of the expression for the convergence rate.

Assumption 16.17 (Constraint on Derivative $h_N(\theta, \eta, w, w_1, w_2)$ w.r.t. θ, η, w, w_1, w_2): The weighting matrices in $h_N(\theta, \eta, w, w_1, w_2)$ have jointly continuous first-order derivatives satisfying

$$\left\| \frac{\partial W_{2ij}(\theta, \eta, w, w_1, w_2)}{\partial x_{[k]}} \right\|_2 \leq c < \infty \tag{16-15}$$

uniformly in $\mathbb{\Theta}_{\mathrm{r}}$, $\mathbb{\eta}_\varepsilon$, $\mathbb{W}$, $\mathbb{W}_1$, and $\mathbb{W}_2$, for $i, j = 1, 2, \ldots, n_\theta$, $k = 1, 2, \ldots, \dim(x)$ and for any N (∞ included). x is a vector that contains all the elements of θ, η, w, w_1, w_2 ($\dim(x) = n_\theta + q + p + pn_\theta + pn_\theta^2$), and c is an N-independent constant.

Theorem 16.18 (Improved Convergence Rate of $\hat{\theta}(z)$ to $\tilde{\theta}(z_0)$): Under Assumptions 15.1 ($P = 4$), 15.18, 16.1, 16.2 (w.p. 1, $k = 0$; in prob., $k = 1, 2$), 16.3(a), 16.10, 16.11 ($k = 0, 1, 2$), 16.12 ($k = 0, 1, 2$), 16.13, 16.15, and 16.17, the minimizer $\hat{\theta}(z)$ can be written as

$$\begin{aligned} \hat{\theta}(z) &= \tilde{\theta}(z_0) + \delta_\theta(z) + b_\theta(z) \\ \delta_\theta(z) &= -V_N''^{-1}(\tilde{\theta}(z_0)) V_N'^T(\tilde{\theta}(z_0), z) \end{aligned} \tag{16-16}$$

with $\delta_\theta(z) = O_\mathrm{p}(N^{-1/2})$ and $b_\theta(z) = O_\mathrm{p}(N^{-1})$.

Proof. Follow the lines of the proof of Theorem 15.21, generalized as in Theorem 16.16 for the stochastic weighting. □

Note that compared with the deterministic weighting (Theorem 15.21), the expected value of $\delta_\theta(z)$ is, in general, not zero and may not even exist. Indeed, $\tilde{\theta}(z_0)$ is not the minimizer of $\mathcal{E}\{V_N(\tilde{\theta}(z_0), z)\}$ (see (16-2) with (16-6) and (16-7)), and the expected value of the weighting $\mathcal{E}\{W_N(\theta, \eta(z), w(\theta, \eta(z), z))\}$ may not exist.

16.5 ASYMPTOTIC BIAS

We assume in this section that a true model exists and that it belongs to the considered model set. Compared with the case with deterministic weighting (see Section 15.5), we need additional assumptions to guarantee that the asymptotic bias behaves as an $O(N^{-1})$. This is due to the fact that the expected value of $\delta_\theta(z)$ in (16-16) is, in general, not zero or equivalently $\mathcal{E}\{V_N'(\theta_0, z)\} \neq 0$. Therefore, using the concept of the truncated estimate (see Section 15.5), the asymptotic bias of $V_N'(\theta_0, z)$ will be analyzed in more details. This requires the following assumptions (see Lemma 15.36).

Assumption 16.19 (Asymptotic Bias $\underline{\eta}(z)$): The bias of the truncated estimate $\underline{\eta}(z)$ is an $O(N^{-1})$: $\mathcal{E}\{\underline{\eta}(z)\} = \eta_* + O(N^{-1})$.

Assumption 16.20 (Derivative $w^{(k)}(\theta, \eta, z)$ w.r.t. η): The derivative of $w^{(k)}(\theta, \eta, z)$ w.r.t. η, $\partial w^{(k)}(\theta, \eta, z)/\partial\eta$, converges uniformly in probability to its expected value $\partial w^{(k)}(\theta, \eta)/\partial\eta$ in Θ_r, η_ε at the rate $O_\mathrm{p}(N^{-1/2})$. The second-order derivative of $w^{(k)}(\theta, \eta, z)$ w.r.t. η is uniformly bounded in Θ_r, η_ε for any N (∞ included): $\left\|\frac{\partial^2 w^{(k)}(\theta, \eta)}{\partial\eta_{[i]}\partial\eta_{[j]}}\right\|_2 \leq O_\mathrm{p}(N^0)$, $i, j = 1, 2, \ldots, q$.

Assumption 16.21 (Constraint on Derivatives $g_N(\theta, \eta, w, w_1)$ w.r.t. η, w, w_1): The weighting matrices in $g_N(\theta, \eta, w, w_1)$ have jointly continuous first- and second-order derivatives satisfying

$$\left\|\frac{\partial W_{1i}(\theta, \eta, w, w_1)}{\partial x_{[j]}}\right\|_1 \leq c_1 < \infty, \quad \left\|\frac{\partial^2 W_{1i}(\theta, \eta, w, w_1)}{\partial x_{[j]}\partial x_{[k]}}\right\|_2 \leq c_2 < \infty$$

uniformly in $\mathbb{\Theta}_r$, $\mathbb{\eta}_\varepsilon$, $\mathbb{W}$, and $\mathbb{W}_1$, for $i = 1, 2, \ldots, n_\theta$, $j, k = 1, 2, \ldots, \dim(x)$ and for any N, ∞ included. x is a vector that contains all the elements of η, w, w_1 ($\dim(x) = q + p + pn_\theta$) and c_1, c_2 are N-independent constants.

Theorem 16.22 (Asymptotic Bias of $\hat{\underline{\theta}}(z)$): Under the assumptions of Theorem 16.18 and Assumptions 15.9, 15.10 or 15.12, 16.19, 16.20 ($k = 0, 1$), and 16.21, the bias $b_\theta = \mathcal{E}\{b_\theta(z)\}$ of the truncated estimate $\hat{\underline{\theta}}(z)$ is an $O(N^{-1})$. If, in addition, $w^{(3)}(\theta, \eta(z), z)$ is continuous in $\mathbb{\Theta}_r$ then the derivative of the bias w.r.t. θ_0, $\partial b_\theta / \partial \theta_0$, is an $O(N^{-1})$.

Proof. See Appendix 16.C. □

16.6 ASYMPTOTIC NORMALITY

The asymptotic distribution function of the minimizer is determined by the asymptotic distribution function of the first derivative of the cost function w.r.t. θ (see Section 15.6). The asymptotic distribution function of the derivative of the cost function (16-10) is influenced by the asymptotic distribution functions of $\eta(z)$ and $w^{(k)}(\theta, \eta(z), z)$, $k = 0, 1$. Therefore, it is natural to make the following assumptions (see Lemma 15.38, Appendix 15.C).

Assumption 16.23 (Asymptotic Distribution $\eta(z)$): $\eta(z)$ can be written as $\eta(z) = \tilde{\eta}(z) + O_p(N^{-1})$ with $\tilde{\eta}(z) = O_p(N^{-1/2})$, and where $\tilde{\eta}(z)$ has finite second-order moments and converges in law at the rate $O(N^{-1/2})$ to a Gaussian random variable with mean value η_*.

Assumption 16.24 (Asymptotic Distribution $w^{(k)}(\theta, \eta, z)$): $w^{(k)}(\theta, \eta, z)$, $k = 0, 1$, can be written as $w^{(k)}(\theta, \eta, z) = \tilde{w}^{(k)}(\theta, \eta, z) + O_p(N^{-1})$ with $\tilde{w}^{(k)}(\theta, \eta, z) = O_p(N^{-1/2})$, and where $\tilde{w}^{(k)}(\theta, \eta, z)$ has finite second-order moments and converges in law at the rate $O(N^{-1/2})$ to a Gaussian random variable with mean value $w^{(k)}(\theta, \eta)$. The convergence is uniform in $\mathbb{\Theta}_r$, $\mathbb{\eta}$.

Theorem 16.25 (Asymptotic Normality of $\sqrt{N}(\hat{\theta}(z) - \tilde{\theta}(z_0))$): Under the assumptions of Theorem 16.18 and Assumptions 15.1 ($P = \infty$), 16.20 ($k = 0, 1$), 16.21, 16.23, and 16.24, $\sqrt{N}(\hat{\theta}(z) - \tilde{\theta}(z_0))$ converges in law at the rate $O(N^{-1/2})$ to a Gaussian random variable. The expression for the covariance matrix (15-18) is still valid if $\delta_\theta(z)$ is replaced by

$$\begin{aligned}
d_\theta(z) &= -V_N{}''^{-1}(\tilde{\theta}(z_0))d_N(z) \\
d_N(z) &= g_N(\tilde{\theta}(z_0), \eta_*, w(\tilde{\theta}(z_0), \eta_*), w'(\tilde{\theta}(z_0), \eta_*), z) \\
&\qquad + \left.\frac{dg_N(\tilde{\theta}(z_0), \eta, w(\tilde{\theta}(z_0), \eta), w'(\tilde{\theta}(z_0), \eta))}{dx}\right|_{x = x_*} (\tilde{x}(z) - x_*)
\end{aligned} \tag{16-17}$$

with

$$\begin{aligned}
x^T &= [\eta^T \; w^T(\tilde{\theta}(z_0), \eta) \; \mathrm{vec}^T(w'(\tilde{\theta}(z_0), \eta))] \\
x_*^T &= [\eta_*^T \; w^T(\tilde{\theta}(z_0), \eta_*) \; \mathrm{vec}^T(w'(\tilde{\theta}(z_0), \eta_*))] \\
\tilde{x}^T(z) &= [\tilde{\eta}^T(z) \; \tilde{w}^T(\tilde{\theta}(z_0), \eta_*, z) \; \mathrm{vec}^T(\tilde{w}'(\tilde{\theta}(z_0), \eta_*, z))]
\end{aligned}$$

$d_\theta(z) = O_p(N^{-1/2})$, and $\mathcal{E}\{d_\theta(z)\} = 0$. The functions $g_N(\ )$ are defined in (16-10) and Lemma 16.14, and $\bar{\eta}(z)$, $\bar{w}^{(k)}(\tilde{\theta}(z_0), \eta, z)$ are defined in Assumptions 16.23 and 16.24. The derivative w.r.t. η in (16-17) is calculated using the chain rule

$$\frac{dg_N}{d\eta} = \frac{\partial g_N}{\partial \eta} + \frac{\partial g_N}{\partial w}\frac{\partial w}{\partial \eta} + \frac{\partial g_N}{\partial \mathrm{vec}(w')}\frac{\partial \mathrm{vec}(w')}{\partial \eta}$$

Proof. See Appendix 16.D.

16.7 OVERVIEW OF THE ASYMPTOTIC PROPERTIES

In this section we give an overview of the asymptotic properties of the minimizer $\hat{\theta}(z)$ (16-4) of a cost function $V_N(\theta, z)$ (16-1), which is quadratic-in-the-measurements when the stochastic vectors $\eta(z)$, $w(\theta, \eta(z), z)$ in the weighting matrix W_N are replaced by deterministic vectors η, w. In the analysis of the stochastic properties of $\hat{\theta}(z)$, the cost functions and minimizers of Table 16-1 play an important role.

TABLE 16-1 Overview of the Notations Used: $\eta(z)$ and $w(\theta, \eta(z), z)$ Are Stochastic Vectors, and η_*, $w(\theta, \eta_*) = \mathcal{E}\{w(\theta, \eta_*, z)\}$ Are the Corresponding Limit Values

Cost function	$V_N(\theta, z) =$ $f_N(\theta, \eta(z), w(\theta, \eta(z), z), z)$	$V_N(\theta) = \mathcal{E}\{f_N(\theta, \eta_*, w(\theta, \eta_*), z)\}$ $= f_N(\theta, \eta_*, w(\theta, \eta_*))$	$V_*(\theta) = \lim_{N \to \infty} V_N(\theta)$
Minimizer	$\hat{\theta}(z)$	$\tilde{\theta}(z_0)$	θ_*

The minimizer $\hat{\theta}(z)$ (16-4) of the cost function $V_N(\theta, z)$ (16-1) has the following asymptotic ($N \to \infty$) properties:

1. *Stochastic convergence:* $\hat{\theta}(z)$ converges strongly to $\tilde{\theta}(z_0)$ (Theorem 16.5).
2. *Stochastic convergence rate:* $\hat{\theta}(z)$ converges in probability at the rate $O_P(N^{-1/2})$ to $\tilde{\theta}(z_0)$ (Theorem 16.16).
3. *Systematic and stochastic errors:* $\hat{\theta}(z)$ converges in probability to $\tilde{\theta}(z_0)$ with

$$\begin{aligned} \hat{\theta}(z) &= \tilde{\theta}(z_0) + \delta_\theta(z) + b_\theta(z) \\ \delta_\theta(z) &= -V_N''^{-1}(\tilde{\theta}(z_0))V_N'^T(\tilde{\theta}(z_0), z) \end{aligned} \tag{16-18}$$

where $\delta_\theta(z) = O_p(N^{-1/2})$ is the dominating stochastic error and where $b_\theta(z) = O_P(N^{-1})$ contains the contribution of the systematic errors (Theorem 16.18).

4. *Asymptotic normality:* $\sqrt{N}(\hat{\theta}(z) - \tilde{\theta}(z_0))$ converges in law at the rate $O(N^{-1/2})$ to a Gaussian random variable with zero mean and covariance matrix $\mathrm{Cov}(\sqrt{N}d_\theta(z))$

$$\begin{aligned} \mathrm{Cov}(\sqrt{N}d_\theta(z)) &= V_N''^{-1}(\tilde{\theta}(z_0))Q_N(\tilde{\theta}(z_0))V_N''^{-1}(\tilde{\theta}(z_0)) \\ Q_N(\tilde{\theta}(z_0)) &= N\mathcal{E}\{d_N(z)d_N^T(z)\} \end{aligned} \tag{16-19}$$

where $d_N(z)$ is defined in (16-17) (Theorem 16.25).

5. *Deterministic convergence:* $\tilde{\theta}(z_0)$ converges to θ_* (Theorem 16.6).

If in addition $V_N(\theta, z)$ satisfies the consistency conditions, then

6. *Consistency:* $\hat{\theta}(z)$ is strongly consistent; replace in properties 1 to 4 $\tilde{\theta}(z_0)$ or $\lim_{F \to \infty} \tilde{\theta}(Z_0) = \theta_*$ by θ_0 (Theorems 16.7, 16.8, and 16.9).
7. *Asymptotic bias:* the asymptotic bias $b_\theta = \mathcal{E}\{b_\theta(z)\}$, and its derivative w.r.t. θ_0, $\partial b_\theta / \partial \theta_0$, of $\underline{\hat{\theta}}(z)$ are an $O(N^{-1})$ (Theorem 16.22).

Note the similarity to the properties of Section 15.8. Compared with the deterministic weighting, the uncertainty (16-19) is increased. This is due to the stochastic vectors $\eta(z)$ and $w(\theta, \eta(z), z)$ in the weighting W_N (compare $d_N(z)$ in (16-17) with $V_N'^T(\tilde{\theta}(z_0), z)$ in (15-18)).

16.8 EXERCISES

16.1. Consider the model equation $y_0(k) = f(\theta, u_0(k))$. Assume that $M \geq 2$ independent repeated experiments of $N \geq n_\theta$ measurements each are available: $y^{[r]}(k) = y_0(k) + n_y^{[r]}(k)$ for $r = 1, 2, \ldots, M$ and $k = 1, 2, \ldots, N$, where the disturbing noise $n_y^{[r]}(k)$ is independent and identically distributed (over r, k) with finite fourth-order moment and $\text{var}(n_y^{[r]}(k)) = \sigma_y^2$. Consider the nonlinear least squares cost function with stochastic weighting

$$\left[\sum_{k=1}^{N} \hat{y}(k) - f(\theta, u_0(k))^2\right] \Big/ \left[\sum_{k=1}^{N} \hat{\sigma}_y^2(k)\right]$$

where $\hat{y}(k)$ and $\hat{\sigma}_y^2(k)$ are, respectively, the sample mean and sample variance of the kth measurement over the M experiments. Show that $\hat{\theta}(z)$ is a strongly consistent estimate if $f(\theta, u_0(k))$ is a continuous function of θ. Under what conditions on $f(\theta, u_0(k))$ are the other results of this chapter valid? What additional assumption on $n_y(k)$ is required for the asymptotic normality property? (Hint: first write the cost function as $\frac{1}{N} z^T W_N(\theta, w(z)) z$ with $z^T = \left[\hat{y}(1)\ 1\ \hat{y}(2)\ 1\ \ldots\ \hat{y}(N)\ 1\right]$, $w(z) = \frac{1}{N}\sum_{k=1}^{N} \hat{\sigma}_y^2(k)$, $W_N(\theta, w(z)) = \text{diag}(C_1, C_2, \ldots, C_N)$, and $C_k = \frac{1}{w(z)}\begin{bmatrix} 1 & -f(\theta, u_0(k)) \\ -f(\theta, u_0(k)) & f^2(\theta, u_0(k)) \end{bmatrix}$.)

16.2. Consider the linear model $y_0(k) = a_0 u_0(k) + b_0$ with $\theta^T = [a\ b]$. Assume that $M \geq 2$ independent repeated experiments of $N \geq 2$ measurements each are available: $y^{[r]}(k) = y_0(k) + n_y^{[r]}(k)$, $u^{[r]}(k) = u_0(k) + n_u^{[r]}(k)$ for $r = 1, 2, \ldots, M$ and $k = 1, 2, \ldots, N$. $n_y^{[r]}(k)$, $n_u^{[r]}(k)$ are independent (over r, k) uniformly bounded random variables. Consider the nonlinear least squares cost function with stochastic weighting

$$\left[\sum_{k=1}^{N} (\hat{y}(k) - a\hat{u}(k) - b)^2\right] \Big/ \left[\sum_{k=1}^{N} (\hat{\sigma}_y^2(k) + a^2 \hat{\sigma}_u^2(k))\right]$$

where $\hat{y}(k)$, $\hat{u}(k)$ and $\hat{\sigma}_y^2(k)$, $\hat{\sigma}_u^2(k)$ are, respectively, the sample means and sample variances of the kth measurement over the M experiments. Show that $\hat{\theta}(z)$ is a strongly consistent estimate. Show that all the other results of this chapter are also valid (hint: follow the lines of Exercise 16.1).

16.3. Consider the linear model $y_0(k) = a_0u_0(k) + b_0$ with $\theta^T = [a\ b]$. Assume that N noisy observations of the input and output are available: $y(k) = y_0(k) + n_y(k)$ and $u(k) = u_0(k) + n_u(k)$, $k = 1, 2, \ldots, N$. $n_y(k)$, $n_u(k)$ are independent Gaussian random variables. The maximum likelihood solution of this problem minimizes

$$\sum_{k=1}^{N}(y(k) - au(k) - b)^2/(\sigma_y^2(k) + a^2\sigma_u^2(k))$$

w.r.t. θ. This requires a nonlinear minimization and, therefore, the following weighted linear least squares approximation, also called iterative quadratic maximum likelihood (IQML), is often calculated:

$$\hat{\theta}_{\mathrm{IQML}} = \arg\min_{\theta}\sum_{k=1}^{N}(y(k) - au(k) - b)^2/(\sigma_y^2(k) + \hat{a}_{\mathrm{LS}}^2\sigma_u^2(k))$$

where $\hat{\theta}_{\mathrm{LS}}$ minimizes the linear least squares cost function $\sum_{k=1}^{N}(y(k) - au(k) - b)^2$. Show that $\hat{\theta}_{\mathrm{IQML}}$ is an inconsistent estimate (hint: apply Theorem 16.5 with $\eta(z) = \hat{a}_{\mathrm{LS}}$ and $w(\theta, \eta(z), z) = 1$, and show that $\tilde{\theta}(z_0) \neq \theta_0$ for any N, ∞ included).

16.9 APPENDIXES

Appendix 16.A: Proof of the Strong Convergence of the Cost Function (Lemma 16.4)

We will show that $V_N(\theta, z) = f_N(\theta, \eta(z), w(\theta, \eta(z), z), z)$ satisfies the two conditions of Corollary 15.32 (see Appendix 15.C). Under Assumptions 15.1 ($P = 4$) and 16.3, $f_N(\theta, \eta, w, z)$ converges uniformly w.p. 1 to $\mathcal{E}\{f_N(\theta, \eta, w, z)\} = f_N(\theta, \eta, w)$ in $\boldsymbol{\Theta}_{\mathrm{r}}$, $\boldsymbol{\eta}_{\varepsilon}$, and $\mathbb{W}$ (Lemma 15.3), so that condition 1 of Corollary 15.32 is satisfied. Under Assumptions 16.1 and 16.2 (w.p. 1, $k = 0$), $w(\theta, \eta(z), z)$ converges uniformly w.p. 1 to $w(\theta, \eta_*)$, an interior point of $\mathbb{W}$, in $\boldsymbol{\Theta}_{\mathrm{r}}$ (Lemma 15.31, Appendix 15.C). Combining this result with Assumption 16.1 shows that condition 2 of Corollary 15.32 is satisfied. We conclude that $V_N(\theta, z)$ converges uniformly w.p. 1 to $f_N(\theta, \eta_*, w(\theta, \eta_*))$ in $\boldsymbol{\Theta}_{\mathrm{r}}$.

$V_N(\theta, z)$, $V_N(\theta)$ are continuous in $\boldsymbol{\Theta}_{\mathrm{r}}$ because $f_N(\theta, \eta, w, z)$, $f_N(\theta, \eta, w)$ are jointly continuous functions of θ, w in $\boldsymbol{\Theta}_{\mathrm{r}}$, $\mathbb{W}$ and $w(\theta, \eta, z)$, $w(\theta, \eta_*)$ are continuous functions of θ in $\boldsymbol{\Theta}_{\mathrm{r}}$. □

Appendix 16.B: Proof of the Convergence Rate of the Minimizer (Theorem 16.16)

The proof follows the same lines as for Theorem 15.19 (see Appendix 15.D). Applying the mean value theorem to the derivative of the cost function $V_N'(\theta, z)$ at the points $\hat{\theta}(z)$ and $\tilde{\theta}(z_0)$ gives

$$V_N'(\hat{\theta}(z), z) = V_N'(\tilde{\theta}(z_0), z) + (\hat{\theta}(z) - \tilde{\theta}(z_0))^T V_N''(\overset{\frown}{\theta}, z) \tag{16-20}$$

where $V_N'(\hat{\theta}(z), z) = 0$ by definition of $\hat{\theta}(z)$, and $\overset{\frown}{\theta} = t\hat{\theta}(z) + (1 - t)\tilde{\theta}(z_0)$ with $t \in [0, 1]$. The proof consists of three main steps. In a first step, the convergence rate of $V_N'(\tilde{\theta}(z_0), z)$ is shown using Corollary 15.35 of Appendix 15.C. Because $V_N'(\tilde{\theta}(z_0)) = 0$, we find

$$V_N'(\tilde{\theta}(z_0), z) = V_N'(\tilde{\theta}(z_0)) + O_{\mathrm{p}}(N^{-1/2}) = O_{\mathrm{p}}(N^{-1/2}) \tag{16-21}$$

uniformly in $\mathbb{\Theta}_r$. In a second step, the convergence of $V_N''(\widehat{\theta}, z)$ is shown using Corollary 15.32 of Appendix 15.C. Because $V_N''(\tilde{\theta}(z_0)) = O(N^0)$ (Assumption 15.18), we get

$$V_N''(\widehat{\theta}, z) = V_N''(\tilde{\theta}(z_0)) + o_{\text{a.s.}}(N^0) = O_{\text{a.s.}}(N^0) \tag{16-22}$$

uniformly in $\mathbb{\Theta}_r$. In the third and last step (16-20), (16-21), and (16-22) are combined, giving

$$\hat{\theta}(z) - \tilde{\theta}(z_0) = V_N''^{-1}(\widehat{\theta}, z) V_N'^T(\tilde{\theta}(z_0), z) = O_p(N^{-1/2}) \tag{16-23}$$

In the first step we verify that $V_N'(\tilde{\theta}(z_0), z)$ fulfills all the conditions of Corollary 15.35. Under Assumptions 15.1 ($P = 4$) and 16.13 $g_N(\theta, \eta, w, w_1, z)$ converges uniformly in prob. to $g_N(\theta, \eta, w, w_1)$ in $\mathbb{\Theta}_r$, $\mathbb{\eta}_\varepsilon$, $\mathbb{W}$, and $\mathbb{W}_1$ at the rate $O_p(N^{-1/2})$ (Lemma 15.17), so that condition 1 of Corollary 15.35 is satisfied. The conditions of Lemma 15.34 are satisfied so that $w^{(k)}(\theta, \eta(z), z)$, $k = 0, 1$, converges uniformly in prob. to $w^{(k)}(\theta, \eta_*)$ in $\mathbb{\Theta}_r$ at the rate $O_p(N^{-1/2})$. This, together with Assumption 16.10, guarantees that condition 3 of Corollary 15.35 is satisfied. Following the same lines as in Appendix 15.E, Eq. (15-45), we conclude from Assumptions 15.1 ($P = 4$) and 16.15 that

$$\left\| \frac{\partial g_{N[i]}(\theta, \eta, w, w_1, z)}{\partial x_{[j]}} \right\|_2 \le \frac{\|z\|_2^2}{N} \left\| \frac{\partial W_{1i}(\theta, \eta, w, w_1)}{\partial x_{[j]}} \right\|_2 = O_{\text{a.s.}}(N^0) \tag{16-24}$$

$i = 1, 2, \ldots, n_\theta$. Hence, condition 2 of Corollary 15.35 is satisfied, thus concluding the first step of the proof.

In the second step we verify that $V_N''(\widehat{\theta}, z)$ fulfills all the conditions of Corollary 15.32. The assumptions of Lemma 16.14 are satisfied and, hence, $V_N''(\theta, z)$ converges uniformly in prob. to $V_N''(\theta)$ in $\mathbb{\Theta}_r$ (condition 1 of Corollary 15.32). The assumptions of Theorem 16.5 (Assumption 15.18 is stronger than Assumption 15.5) are satisfied so that $\hat{\theta}(z)$, and, hence, also $\widehat{\theta}$, converges in prob. to $\tilde{\theta}(z_0)$ (Theorem 16.5 without Assumption 16.3(b) shows convergence in prob.). The conditions of Lemma 15.31 are satisfied so that $w^{(k)}(\theta, \eta(z), z)$, $k = 0, 1, 2$, converges uniformly in prob. to $w^{(k)}(\theta, \eta_*)$ in $\mathbb{\Theta}_r$. Together with Assumption 16.1, it shows that condition 2 of Corollary 15.32 is satisfied, which concludes the second step of the proof. □

Appendix 16.C: Proof of the Asymptotic Bias of the Truncated Minimizer (Theorem 16.22)

The results of Theorem 16.18 are valid so that only $V_N'(\theta, z)$ must be studied. We will show that $V_N'(\theta, z) = g_N(\theta, \eta(z), w(\theta, \eta(z), z), w'(\theta, \eta(z), z), z)$ satisfies the conditions of Corollary 15.37 (see Appendix 15.C), which proves the theorem.

Under Assumptions 16.1, 16.2 ($k = 0, 1$), 16.10, 16.11 ($k = 0, 1$), 16.19, and 16.20 ($k = 0, 1$), $w^{(k)}(\theta, \eta(z), z)$ satisfies the conditions of Lemma 15.36. Hence, it converges uniformly in prob. to $w^{(k)}(\theta, \eta_*)$ at the rate $O_p(N^{-1/2})$ with bias

$$\mathcal{E}\{\underline{w}^{(k)}(\theta, \eta(z), z)\} = w^{(k)}(\theta, \eta_*) + O(N^{-1})$$

It follows that conditions 3 and 4 of Corollary 15.37 are satisfied for $\eta(z)$ and $w^{(k)}(\theta, \eta(z), z)$.

Under Assumptions 15.1 $(P = 4)$ and 16.21, $\partial g_N(\theta, \eta, w, w_1, z)/\partial x_{[j]}$ converges uniformly in prob. to $\partial g_N(\theta, \eta, w, w_1)/\partial x_{[i]}$ at the rate $O_p(N^{-1/2})$ (proof: similar to Lemma 15.17). Under the same assumptions, we also have

$$\left\| \frac{\partial^2 g_{N[i]}(\theta, \eta, w, w_1, z)}{\partial x_{[j]} \partial x_{[k]}} \right\|_2 \le O_{\text{a.s.}}(N^0)$$

$i = 1, 2, \ldots, n_\theta$ (proof: similar to (16-24)). From the proof of Theorem 16.18, it follows that $g_N(\theta, \eta, w, w_1, z)$ satisfies condition 1 of Corollary 15.35. Hence, conditions 1 and 2 of Corollary 15.37 are satisfied, thus concluding the proof for the bias.

If $w^{(3)}(\theta, \eta(z), z)$ is continuous in Θ_r, then under Assumptions 16.2 $(k = 0, 1, 2)$ and 16.17, $V_N(\theta, z)$ has continuous first-, second-, and third-order derivatives w.r.t. θ in Θ_r. This is sufficient to show that $\partial b_\theta/\partial \theta_0$ is an $O(N^{-1})$ (see the proof of Theorem 15.28 in Appendix 15.G). □

Appendix 16.D: Proof of the Asymptotic Normality of the Minimizer (Theorem 16.25)

Multiplying (16-16) by $\sqrt{N}$ and taking the limit for $N \to \infty$ gives

$$\operatorname*{plim}_{N \to \infty} \sqrt{N}(\hat{\theta}(z) - \tilde{\theta}(z_0) - \delta_\theta(z)) = 0 \tag{16-25}$$

Because convergence in probability implies convergence in law (see Section 14.7 , interrelation 3), it follows directly from (16-25) that $\sqrt{N}(\hat{\theta}(z) - \tilde{\theta}(z_0))$ has the same asymptotic distribution function as $\delta_\theta(z)$. We will show that $\delta_\theta(z)$ converges in law at the rate $O(N^{-1/2})$ to a Gaussian random variable. To prove this it is sufficient to show that the stochastic part of $\delta_\theta(z)$, namely,

$$V_N'^T(\tilde{\theta}(z_0), z) = g_N(\tilde{\theta}(z_0), \eta(z), w(\tilde{\theta}(z_0), \eta(z), z), w'(\tilde{\theta}(z_0), \eta(z), z), z) \tag{16-26}$$

satisfies the conditions of Corollary 15.39.

Condition 1 of Corollary 15.39 is satisfied under Assumptions 15.1 $(P = \infty)$, 16.3(a), and 16.13 (proof: similar to Theorem 15.29). Conditions 2 and 3 of Corollary 15.39 are satisfied under Assumptions 15.1 $(P = 4)$, and 16.21 (proof: see Theorem 16.22, Appendix 16.C). Under Assumptions 16.1, 16.2 (in prob., $k = 0, 1$), 16.10, 16.11 $(k = 0, 1)$, 16.20 $(k = 0, 1)$, 16.23, and 16.24, $w^{(k)}(\theta, \eta(z), z)$ satisfies the conditions of Lemma 15.38. Hence, $w^{(k)}(\theta, \eta(z), z)$ converges uniformly in prob. to $w^{(k)}(\theta, \eta_*)$ at the rate $O_p(N^{-1/2})$, is asymptotically normally distributed at the rate $O(N^{-1/2})$, and can be written as

$$w^{(k)}(\theta, \eta(z), z) = w^{(k)}(\theta, \eta_*, z) + \frac{\partial w^{(k)}(\theta, \eta)}{\partial \eta_*}(\eta(z) - \eta_*) + O_p(N^{-1}) \tag{16-27}$$

Combined with Assumptions 16.1, 16.10, and 16.23, it shows that condition 4 of Corollary 15.39 is also fulfilled. We conclude from Corollary 15.39 and (16-27) that (16-26) can be written as

$$\begin{aligned} V_N'^T(\tilde{\theta}(z_0), z) &= g_N(\tilde{\theta}(z_0), \eta_*, w(\tilde{\theta}(z_0), \eta_*), w'(\tilde{\theta}(z_0), \eta_*), z) \\ &\quad + \left.\frac{dg_N(\tilde{\theta}(z_0), \eta, w(\tilde{\theta}(z_0), \eta), w'(\tilde{\theta}(z_0), \eta))}{dx}\right|_{x = x_*} (x(z) - x_*) + O(N^{-1}) \end{aligned} \tag{16-28}$$

with

$$x^T(z) = \tilde{x}^T(z) = [\eta^T(z)\ w^T(\tilde{\theta}(z_0), \eta_*, z)\ \text{vec}^T(w'(\tilde{\theta}(z_0), \eta_*, z))]$$

and where the derivative w.r.t. η is calculated using the chain rule

$$\frac{dg_N}{d\eta} = \frac{\partial g_N}{\partial \eta} + \frac{\partial g_N \partial w}{\partial w\, \partial \eta} + \frac{\partial g_N}{\partial \text{vec}(w')} \frac{\partial \text{vec}(w')}{\partial \eta}$$

The first two terms in the right-hand side of (16-28) are asymptotically normally distributed. Because $x(z) = \tilde{x}(z) + O_p(N^{-1})$, we can replace $x(z)$ by $\tilde{x}(z)$ in (16-28), which concludes the proof. □

17

Identification of Semilinear Models

Abstract: Many signal and system modeling problems lead to parametric models that are linear-in-the-measurements. This chapter treats the identification (parameter estimation and model selection) of such models using the Markov estimator. The asymptotic properties (consistency, convergence rate, asymptotic bias, asymptotic normality, and asymptotic efficiency) of the Markov estimates are analyzed. The different aspects of model selection, such as model validation, and detection of undermodeling and overmodeling are discussed. Explicit expressions for the Cramér-Rao lower bound are derived and conditions for the asymptotic efficiency of the Gaussian maximum likelihood estimator are given. The presented theory is applicable to general signal modeling and system identification problems. Readers who are unfamiliar with the analysis of the stochastic properties of estimators should first read Sections 14.11 to 14.13 and Chapter 15.

17.1 THE SEMILINEAR MODEL

Consider the following general model based on N observations:

$$M_0(\theta) + M_1(\theta)z = 0 \qquad (17\text{-}1)$$

which is linear-in-the-measurements $z \in \mathbb{R}^{sN}$ and (non)linear-in-the-model-parameters $\theta \in \mathbb{R}^{n_\theta}$. $M_0(\theta) \in \mathbb{R}^{rN}$, $M_1(\theta) \in \mathbb{R}^{rN \times sN}$ with $s \geq r$ has rank rN and n_θ, s and r fixed integers, independent of the number of observations N. Each time a new observation is added, the number of model equations and the number of measurements increase with, respectively, r and s. In frequency domain applications, (17-1) often has a block diagonal structure

$$M_{0k}(\theta) + M_{1k}(\theta)z_k = 0 \text{ for } k = 1, 2, \ldots, N \qquad (17\text{-}2)$$

System Identification: A Frequency Domain Approach. Rik Pintelon and Johan Schoukens
Copyright © 2001 Institute of Electrical and Electronics Engineers. ISBN: 0-780-36000-1

with $M_{0k}(\theta) \in \mathbb{R}^r$, $M_{1k}(\theta) \in \mathbb{R}^{r \times s}$, and $z_k \in \mathbb{R}^s$. The relationship with (17-1) is given by

$$\begin{aligned} M_0^T(\theta) &= [M_{01}^T(\theta)...M_{0N}^T(\theta)] \\ M_1(\theta) &= \text{diag}(M_{10}(\theta), ..., M_{1N}(\theta)) \\ z^T &= [z_1^T...z_N^T] \end{aligned}$$

Two special cases of model (17-1) are worth mentioning.

17.1.1 Signal Model

Putting $s = r$ and $M_1(\theta) = -I_{sN}$ in (17-1) gives the signal model

$$z = M_0(\theta) \tag{17-3}$$

This is typically the form encountered when estimating a linear combination of basis signals such as sine waves (Pintelon and Schoukens, 1996), cisoids (Cadzow, 1990), and exponential functions (Van den Bos and Swarte, 1993).

17.1.2 Transfer Function Model

Putting $M_0(\theta) = 0$, $M_1(\theta) = [A(\theta) \quad -B(\theta)]$, and $z^T = [y^T \quad u^T]$ gives the transfer function model

$$A(\theta)y = B(\theta)u \tag{17-4}$$

where $y \in \mathbb{R}^{rN}$ and $u \in \mathbb{R}^{qN}$ with $r + q = s$. $A(\theta) \in \mathbb{R}^{rN \times rN}$ is a regular matrix. The model equations of a linear time-invariant discrete-time multivariable system can be written in the time domain under this form (Exercise 17.1). u and y are, respectively, the stacked input and output signals of the system, while r and q are, respectively, the numbers of outputs n_y and inputs n_u. The left matrix fraction description (5-56) can be written under the block diagonal form (17-2) with

$$\begin{aligned} M_{0k}(\theta) &= 0 \\ M_{1k}(\theta) &= \left[A_{\text{Re}}(\Omega_k, \theta) \quad -B_{\text{Re}}(\Omega_k, \theta)\right] \\ z_k^T &= \left[Y_{\text{re}}^T(k) \quad U_{\text{re}}^T(k)\right] \end{aligned}$$

and where the operators $(\ )_{\text{Re}}$ and $(\ )_{\text{re}}$ are defined in Section 13.8 (proof: apply Lemma 13.4 to (5-56)). $Y(k)$ and $U(k)$ are, respectively, the n_y by 1 output and the n_u by 1 input DFT spectra at frequency k $(s = 2(n_y + n_u))$. If the initial conditions are included in the model, then $M_{0k}(\theta) = I_{\text{re}}(\Omega_k, \theta)$, with $I(\Omega_k, \theta)$ the n_y by 1 vector of the equivalent initial conditions (see Section 5.6), and (17-2) becomes (Exercise 17.3)

$$A_{\text{Re}}(\Omega_k, \theta)Y_{\text{re}}(k) = B_{\text{Re}}(\Omega_k, \theta)U_{\text{re}}(k) + I_{\text{re}}(\Omega_k, \theta) \tag{17-5}$$

17.2 THE MARKOV ESTIMATOR

First we construct the Markov estimates for real observations and real model parameters. Afterward, the results are generalized to complex observations and complex model parameters.

17.2.1 Real Case

An estimate $\hat{\theta}$ of the model parameters θ of the semilinear model (17-1) is calculated using noisy observations $z = z_0 + n_z$ of the true (unknown) values z_0. Because z_0 is unknown, it should also be estimated and is parameterized as z_p. Under Assumption 15.1(2), the Markov estimator minimizes the squared residuals $(z - z_p)$ weighted with the noise covariance matrix $C_{n_z} = \mathrm{Cov}(n_z)$, taking into account the model equations (17-1). This constrained minimization problem, with parameters θ and z_p, can be solved using Lagrange multipliers $\lambda \in \mathbb{R}^{rN}$ (Kaplan, 1993)

$$\frac{1}{2}(z - z_p)^T C_{n_z}^{+}(z - z_p) + \lambda^T(M_0(\theta) + M_1(\theta)z_p) \tag{17-6}$$

with + the Moore-Penrose pseudoinverse (see Section 13.5). Singular noise covariance matrices are allowed to cover the case where parts of the measurements may be known exactly. This is, for example, the case in transfer function modeling with known inputs. Because (17-6) is quadratic in z_p and linear in λ, z_p and λ can be explicitly eliminated. This gives the following expression for z_p (see Appendix 17.A):

$$C_{n_z}C_{n_z}^{+}z_p = C_{n_z}C_{n_z}^{+}z - C_{n_z}M_1^T(\theta)(M_1(\theta)C_{n_z}M_1^T(\theta))^{-1}(M_0(\theta) + M_1(\theta)z) \tag{17-7}$$

It makes it possible to eliminate the parameters z_p in (17-6), which results in a significant reduction of the size of the minimization problem. The following Markov cost function is found:

$$V_{\mathrm{Markov}}(\theta, z) = \frac{1}{2}e^T(\theta, z)C_e^{-1}(\theta)e(\theta, z) = \frac{1}{2}\varepsilon^T(\theta, z)\varepsilon(\theta, z) \tag{17-8}$$

with

$$e(\theta, z) = M_0(\theta) + M_1(\theta)z, \quad C_e(\theta) = M_1(\theta)C_{n_z}M_1^T(\theta), \quad \varepsilon(\theta, z) = \Lambda(\theta)e(\theta, z) \tag{17-9}$$

and where $\Lambda(\theta) \in \mathbb{R}^{rN \times rN}$ satisfies $\Lambda^T(\theta)\Lambda(\theta) = C_e^{-1}(\theta)$ (see Appendix 17.A). Note that $\mathrm{Cov}(e(\theta, n_z)) = C_e(\theta)$ and $\mathrm{Cov}(\varepsilon(\theta, n_z)) = I_{rN}$. Minimizing the cost function (17-8) w.r.t. θ gives the Markov estimates of the model parameters

$$\hat{\theta}(z) = \arg\min_{\theta \in \boldsymbol{\theta}_r} V_{\mathrm{Markov}}(\theta, z) \tag{17-10}$$

The Markov estimates $\hat{z}$ of the true observations z_0 are found by evaluating (17-7) at $\theta = \hat{\theta}(z)$

$$C_{n_z} C_{n_z}^{+} \hat{z} = C_{n_z} C_{n_z}^{+} z - C_{n_z} M_1^T(\hat{\theta}(z)) C_e^{-1}(\hat{\theta}(z)) e(\hat{\theta}(z), z) \tag{17-11}$$

Note that this formula estimates the observations lying in the regular space of C_{n_z}. Those lying in the null space of C_{n_z} are known exactly.

The Markov estimates require knowledge of the noise covariance matrix C_{n_z}. It can be estimated from independent repeated experiments (see Chapter 8 for transfer function modeling of SISO systems). This consumes a great deal of computer time and memory space if N is large. In frequency domain identification only a (block) diagonal version of C_{n_z} is required. Using the block diagonal structure (17-2) of the model equations and replacing C_{n_z} in (17-8) by $\text{diag}(C_{n_{z1}}, \ldots, C_{n_{zN}})$, with $C_{n_{zk}} = \text{Cov}(n_{zk})$, gives the simplified Markov cost function

$$V_{\text{Markov}}(\theta, z) = \frac{1}{2}\sum\nolimits_{k=1}^{N} e_k^T(\theta, z_k) C_{e_k}^{-1}(\theta) e_k(\theta, z_k) \tag{17-12}$$

with $e_k(\theta, z_k) = M_{0k}(\theta) + M_{1k}(\theta) z_k$, $C_{e_k}(\theta) = \text{Cov}(e_k(\theta, n_{zk})) = M_{1k}(\theta) C_{n_{zk}} M_{1k}^T(\theta)$, and $z_k = z_{0k} + n_{zk}$. Neglecting the nondiagonal terms of $C_{e_k}(\theta)$ in (17-12), the Markov cost function can even be simplified further to

$$V_{\text{Markov}}(\theta, z) = \frac{1}{2}\sum_{k=1}^{N}\sum_{i=1}^{r} \frac{e_{k[i]}^2(\theta, z_k)}{\text{var}(e_{k[i]}(\theta, n_{zk}))} \tag{17-13}$$

The stochastic properties of the minimizers of (17-8), (17-12), and (17-13) are analyzed in Section 17.4.

17.2.2 Complex Case

Expressions (17-8), (17-12), and (17-13) are still valid for *complex-valued observations* $z \in \mathbb{C}^{sN}$ and *complex-valued model parameters* $\theta \in \mathbb{C}^{n_\theta}$, if applied to z_{re} and θ_{re}. If in addition, *the errors* n_z *are circular complex distributed* (see (14-12) and (14-13)), then (17-8) can be written as (see Example 13.5)

$$\begin{aligned} V_{\text{Markov}}(\theta, z) &= e^H(\theta, z) C_e^{-1}(\theta) e(\theta, z) \\ &= \varepsilon^H(\theta, z)\varepsilon(\theta, z) \end{aligned} \tag{17-14}$$

with $\varepsilon(\theta, z) = \Lambda(\theta) e(\theta, z)$ and where $\Lambda(\theta) \in \mathbb{C}^{rN \times rN}$ satisfies $\Lambda^H(\theta)\Lambda(\theta) = C_e^{-1}(\theta)$.

17.3 CRAMÉR-RAO LOWER BOUND

We first derive the Carmér-Rao lower bound for real observations and real model parameters. Afterward, the results are generalized to complex observations and complex parameters.

17.3.1 Real Case

The concept of Cramér-Rao lower bound requires the existence of a true model

$$M_0(\theta_0) + M_1(\theta_0) z_0 = 0 \tag{17-15}$$

with z_0 the true observations and θ_0 the true model parameters, and knowledge of the probability density function of the measurements $z = z_0 + n_z$. The Cramér-Rao lower bound for unbiased estimators (14-87) is constructed under the following assumption.

Assumption 17.1 (Gaussian Errors): The observations z_0 are deterministic and the errors n_z are normally distributed with known covariance matrix C_{n_z}.

The log-likelihood function becomes

$$\ln f_z(z, z_\mathrm{p}, \theta) = -\frac{1}{2}(z - z_\mathrm{p})^T C_{n_z}^{+}(z - z_\mathrm{p}) + c \tag{17-16}$$

with c a constant independent of z_p and θ, and where the parameters z_p and θ satisfy the constraint (17-1)

$$M_0(\theta) + M_1(\theta) z_\mathrm{p} = 0 \tag{17-17}$$

Straightforward calculation of the Fisher information matrix $Fi(z_0, \theta_0)$ (see (14-85)) from (17-16) is impossible because z_p contains too many unknowns. Indeed, the parameters of z_p lying in the singular space of C_{n_z} are known exactly and should not appear in the CR bound. Next, (17-17) puts rN linear constraints on the entries of z_p. The resulting rN linear dependent variables should also not appear in the CR bound.

Hence, the first step in calculating the CR bound consists of reducing the sN parameters z_p to the $\mathrm{rank}(C_{n_z}) - rN$ parameters x: $z_\mathrm{p} = z_\mathrm{p}(x, \theta)$ (see Appendix 17.B). It facilitates writing the log-likelihood function as

$$\ln f_z(z, x, \theta) = -\frac{1}{2}(z - z_\mathrm{p}(x, \theta))^T C_{n_z}^{+}(z - z_\mathrm{p}(x, \theta)) + c \tag{17-18}$$

The corresponding Fisher information matrix $Fi(x_0, \theta_0)$ is

$$Fi(x_0, \theta_0) = \begin{bmatrix} F_{xx} & F_{x\theta} \\ F_{x\theta}^T & F_{\theta\theta} \end{bmatrix} \tag{17-19}$$

Applying the inverse of block matrices (13-8) to $CR(x_0, \theta_0) = Fi^{-1}(x_0, \theta_0)$ gives the Cramér-Rao lower bound on the model parameters

$$CR(\theta_0) = Fi^{-1}(\theta_0) = (F_{\theta\theta} - F_{x\theta}^T F_{xx}^{-1} F_{x\theta})^{-1} \tag{17-20}$$

Filling out the explicit expressions of $F_{\theta\theta}$, F_{xx}, and $F_{x\theta}$ in (17-20) gives, after some calculations (see Appendix 17.B),

$$\begin{aligned} Fi(\theta_0) &= V_{\mathrm{Markov}}''(\theta_0, z_0) \\ &= \left(\frac{\partial e(\theta, z_0)}{\partial \theta_0}\right)^T C_e^{-1}(\theta_0)\left(\frac{\partial e(\theta, z_0)}{\partial \theta_0}\right) \\ &= \left(\frac{\partial \varepsilon(\theta, z_0)}{\partial \theta_0}\right)^T \left(\frac{\partial \varepsilon(\theta, z_0)}{\partial \theta_0}\right) \end{aligned} \tag{17-21}$$

This shows that the Fisher information matrix of the model parameters $Fi(\theta_0)$ equals the Hessian of the Markov cost function (17-8), evaluated at the true observations and the true model parameters. Hence, the larger the eigenvalues of the Hessian matrix, the smaller the CR bound of the model parameters.

17.3.2 Complex Case

We first consider the case where *the observations* $z \in \mathbb{C}^{sN}$ *are complex, the errors* $n_z \in \mathbb{C}^{sN}$ *are circular complex distributed* (see (14-12) and (14-13)), and *the model parameters* θ *are real.* This is, for example, true in frequency domain system identification. Formula (17-21) of the real case still applies to z_{re} and $e_{\text{re}}(\theta_0, z_0)$. Using the isomorphism between complex and real matrices (see Section 13.8), (17-21) becomes

$$\begin{aligned} Fi(\theta_0) &= V_{\text{Markov}}''(\theta_0, z_0) \\ &= 2\text{Re}\left(\left(\frac{\partial e(\theta, z_0)}{\partial \theta_0}\right)^H C_e^{-1}(\theta_0)\left(\frac{\partial e(\theta, z_0)}{\partial \theta_0}\right)\right) \\ &= 2\text{Re}\left(\left(\frac{\partial \varepsilon(\theta, z_0)}{\partial \theta_0}\right)^H \left(\frac{\partial \varepsilon(\theta, z_0)}{\partial \theta_0}\right)\right) \end{aligned} \tag{17-22}$$

with $V_{\text{Markov}}(\theta, z)$ is given by (17-14) (see Exercise 17.6).

Next, we consider the case where, also, the *model parameters* θ *are complex.* Applications where θ is complex can be found in nuclear magnetic resonance spectroscopy (Kumaresan et al., 1990) and the diagnosis of asymmetry of rotating machinery (Lee and Joh, 1994; Peeters et al., 2000). Formula (17-22) is still valid for θ_{re}. If $e(\theta, z)$ is an analytic function of $\theta \in \mathbb{C}^{n_\theta}$, then

$$\frac{\partial e(\theta, z)}{\partial \theta_{\text{re}}} = \left[\frac{\partial e(\theta, z)}{\partial \text{Re}(\theta)} \quad \frac{\partial e(\theta, z)}{\partial \text{Im}(\theta)}\right] = \left[\frac{\partial e(\theta, z)}{\partial \theta} \quad j\frac{\partial e(\theta, z)}{\partial \theta}\right]$$

and (17-22) can be rewritten as

$$\begin{aligned} Fi(\theta_0) &= 2\left(\frac{\partial e(\theta, z_0)}{\partial \theta_0}\right)^H C_e^{-1}(\theta_0)\left(\frac{\partial e(\theta, z_0)}{\partial \theta_0}\right) \\ &= 2\left(\frac{\partial \varepsilon(\theta, z_0)}{\partial \theta_0}\right)^H \left(\frac{\partial \varepsilon(\theta, z_0)}{\partial \theta_0}\right) \end{aligned} \tag{17-23}$$

(see Exercise 17.7).

17.4 PROPERTIES MARKOV ESTIMATOR

The properties of the Markov estimator will be studied for real observations and real model parameters. Following the lines of Sections 17.2.2 and 17.3.2, it is easy to see that the results are also valid for complex observations and complex model parameters. Note that the Markov cost function (17-8) fits within the framework of Chapter 15. Indeed, (17-8) can be written as

$$V_N(\theta, z) = \frac{1}{N}V_{\text{Markov}}(\theta, z) = \frac{1}{N}\begin{bmatrix}1\\z\end{bmatrix}^T W_N(\theta)\begin{bmatrix}1\\z\end{bmatrix} \qquad (17\text{-}24)$$

with $W_N(\theta)$ an $(sN+1)$ by $(sN+1)$ weighting matrix

$$W_N(\theta) = \frac{1}{2}[M_0(\theta) \;\; M_1(\theta)]^T C_e^{-1}(\theta)[M_0(\theta) \;\; M_1(\theta)] \qquad (17\text{-}25)$$

Cost function (17-24) has exactly the same form as (15-1) and, hence, all the results of Chapter 15 apply to the Markov estimator (17-10). The same is true for the Markov estimates based on the simplified cost functions (17-12) and (17-13). Only the assumptions and the properties that can be worked out more specifically for the Markov estimator are discussed here.

Approximate expressions for "large" signal-to-noise ratios ($\|z_0\|_2/\|n_z\|_2 \gg 1$) and "small" model errors are obtained by replacing n_z by υn_z (and, hence, C_{n_z} by $\upsilon^2 C_{n_z}$) with $\upsilon \to 0$ and $e(\tilde{\theta}(z), z_0)$ by $\mu e(\tilde{\theta}(z), z_0)$ with $\mu \to 0$.

17.4.1 Consistency

Besides the model parameters, (a part of) the measurements are also estimated. First, the consistency of the estimates $\hat{\theta}(z)$ (17-10) of the model parameters is analyzed. Next, the consistency of the estimates $\hat{z}$ (17-11) of the measurements is discussed.

17.4.1.1 Model Parameters. Assumption 15.2 requires that $C_e(\theta)$ is positive definite in the compact set Θ_r. The condition $\|W_N(\theta)\|_1 \le c < \infty$ imposes restrictions on the one and infinity norm of $M_1(\theta)$ and C_{n_z}, while the condition $\|W_{m[1:n,\,1:n]}(\theta) - W_n(\theta)\|_1^2 = O((m-n)/m)$, with $m \ge n$, limits the variation of $C_e^{-1}(\theta)$ as N increases. Note that both conditions are automatically satisfied for the simplified Markov estimators (17-12) and (17-13). Under Assumption 15.1, using

$$e(\theta, z) = e(\theta, z_0) + M_1(\theta)n_z \quad \text{and} \quad \operatorname{trace}(e^T C_e^{-1} e) = \operatorname{trace}(C_e^{-1} e e^T)$$

we find the expected value of the cost function (17-8)

$$V_{\text{Markov}}(\theta) = \mathcal{E}\{V_{\text{Markov}}(\theta, z)\} = \mathcal{E}\{V_{\text{Markov}}(\theta, z_0)\} + rN/2 \qquad (17\text{-}26)$$

Assumption 15.10 is satisfied because $\mathcal{E}\{V_{\text{Markov}}(\theta_0, z_0)\} = 0$ and $rN/2$ is θ independent. Hence, it follows from Theorem 15.11 that the Markov estimate (17-10) is strongly consistent.

Note that the expected values of the simplified cost functions (17-12) and (17-13) are also given by (17-26). We conclude that the Markov estimates of the (block) diagonal model (17-2) based on the simplified cost functions (17-12), (17-13) are still strongly consistent. Removing some parts of the nondiagonal elements of C_{n_z} does not influence the consistency property: the minimal requirement is that each residual is weighted with its variance. For frequency domain system identification, this means that the correlation of the errors n_z over the frequencies and between the different outputs can be neglected. However, the correlation of the errors n_z between an output and all the inputs may not be removed because it influences $\operatorname{var}(e_{k[i]}(\theta, z_k))$; otherwise consistency is lost.

17.4.1.2 Observations. In general, the estimates $\hat{z}$ (17-11) of the observations are inconsistent. This is due to the fact that the uncertainty of z in (17-11) is not decreased by making more observations (no averaging effect occurs in $C_{n_z}C_{n_z}^{+}z$). z cancels in (17-11) for signal models (17-3) and transfer function models (17-4) with known input (excitation) signals u_0. In these cases, strongly consistent estimates of the observations are obtained through, respectively, $\hat{z} = M_0(\hat{\theta}(z))$ and $\hat{y} = A^{-1}(\hat{\theta}(z))B(\hat{\theta}(z))u_0$ (see Appendix 17.C).

17.4.2 Strong Convergence

If model errors are present $(e(\tilde{\theta}(z_0), z_0) \neq 0)$, one can wonder why the Markov estimator should be preferred over, for example, the nonlinear least squares estimator

$$V_{\text{NLS}}(\theta, z) = \frac{1}{2}e^T(\theta, z)e(\theta, z) \tag{17-27}$$

The reason for this is that the Markov estimate $\hat{\theta}(z)$ converges to a value $\tilde{\theta}(z_0)$ that is independent of the signal-to-noise ratio. Indeed, replacing n_z by υn_z (and, hence, C_{n_z} by $\upsilon^2 C_{n_z}$) in the expected value of the cost function (17-26) gives

$$V_{\text{Markov}}(\theta) = \upsilon^2 \mathcal{E}\{V_{\text{Markov}}(\theta, z_0)\} + rN/2 \tag{17-28}$$

It shows that $\tilde{\theta}(z_0)$, the minimizing argument of (17-28), is independent of υ. The same is true for the simplified Markov estimators (17-12) and (17-13). This is not the case for the nonlinear least squares estimator (17-27). Indeed, the expected value of (17-27) equals

$$V_{\text{NLS}}(\theta) = \mathcal{E}\{V_{\text{NLS}}(\theta, z)\} = \mathcal{E}\{V_{\text{NLS}}(\theta, z_0)\} + \text{trace}(C_e(\theta)) \tag{17-29}$$

Replacing n_z by υn_z (and, hence, C_{n_z} by $\upsilon^2 C_{n_z}$) in (17-29) gives

$$V_{\text{NLS}}(\theta) = \mathcal{E}\{V_{\text{NLS}}(\theta, z)\} = \mathcal{E}\{V_{\text{NLS}}(\theta, z_0)\} + \upsilon^2\text{trace}(C_e(\theta)) \tag{17-30}$$

Clearly, the minimizer of (17-30) depends on υ.

17.4.3 Convergence Rate

Expression (15-14) for the difference $\hat{\theta}(z) - \tilde{\theta}(z_0)$ can be elaborated for the Markov estimator. It will be used to study the statistical properties of the residuals $\varepsilon(\hat{\theta}(z), z)$ and the global minimum of the cost function $V_{\text{Markov}}(\hat{\theta}(z), z)$.

Theorem 17.2 (Convergence Rate $\hat{\theta}(z)$ to $\tilde{\theta}(z_0)$): Under the assumptions of Theorem 15.21, large signal-to-noise ratios $(\upsilon \to 0)$ and small model errors $(\mu \to 0)$, the minimizer $\hat{\theta}(z)$ can be written as

$$\begin{aligned} \hat{\theta}(z) - \tilde{\theta}(z_0) &= \Delta_\theta(z) + \partial_\theta(z) + b_\theta(z) \\ \Delta_\theta(z) &= -\left[\left(\frac{\partial\varepsilon(\theta, z_0)}{\partial\tilde{\theta}(z_0)}\right)^T\left(\frac{\partial\varepsilon(\theta, z_0)}{\partial\tilde{\theta}(z_0)}\right)\right]^{-1}\left(\frac{\partial\varepsilon(\theta, z_0)}{\partial\tilde{\theta}(z_0)}\right)^T\left(\frac{\partial\varepsilon(\tilde{\theta}(z_0), z)}{\partial z}\right)n_z \end{aligned} \tag{17-31}$$

where $\mathcal{E}\{\Delta_\theta(z)\} = 0$, $\mathcal{E}\{\partial_\theta(z)\} = 0$ and

$$\begin{aligned}
\Delta_\theta(z) &= \upsilon O_{\mathrm{p}}(N^{-1/2})\\
\partial_\theta(z) &= (\upsilon^2 + \upsilon\mu + \mu\lambda(z_0))O_{\mathrm{p}}(N^{-1/2})\\
b_\theta(z) &= (\upsilon^2 + (\upsilon + \mu)\lambda(z_0))O_{\mathrm{p}}(N^{-1})
\end{aligned}$$

with $\lambda(z_0) = 1$ for random z_0 and $\lambda(z_0) = 0$ for deterministic z_0.

Proof. See Appendix 17.D. □

From (17-31), it follows that in the presence of model errors ($\mu \neq 0$), $\partial_\theta(z)$ and $b_\theta(z)$ do not decrease to zero for random z_0 as the noise level υ tends to zero. In the absence of model errors ($\mu = 0$), $\Delta_\theta(z)$ and $b_\theta(z)$ are, for deterministic z_0, an $\upsilon O_{\mathrm{p}}(N^{-1/2})$ and $\upsilon^2 O_{\mathrm{p}}(N^{-1})$, respectively. It shows that the bias error decreases as υ^2 while the stochastic error decreases as υ. Although $b_\theta(z) = (\upsilon^2 + \upsilon\lambda(z_0))O_{\mathrm{p}}(N^{-1})$ for random z_0, the conclusion remains valid because the expected value of $\upsilon\lambda(z_0)O_{\mathrm{p}}(N^{-1})$ in $b_\theta(z)$ is zero (see Appendix 17.D).

17.4.4 Asymptotic Normality

If the true model belongs to the model set, then expression (15-18) of the covariance matrix in Theorem 15.29 (asymptotic normality of $\sqrt{N}(\hat{\theta}(z) - \tilde{\theta}(z_0))$) can be elaborated.

Theorem 17.3 (Asymptotic Normality of $\sqrt{N}(\hat{\theta}(z) - \theta_0)$): Under the assumptions of Theorem 15.21 and Assumptions 15.1 ($P = \infty$) and 15.9, $\sqrt{N}(\hat{\theta}(z) - \theta_0)$ converges in law, at the rate $O(N^{-1/2})$, to a Gaussian random variable with zero mean and covariance matrix $\mathrm{Cov}(\sqrt{N}\delta_\theta(z))$

$$\begin{aligned}
\mathrm{Cov}(\sqrt{N}\delta_\theta(z)) &= V_N''^{-1}(\theta_0) + V_N''^{-1}(\theta_0)q_N(\theta_0)V_N''^{-1}(\theta_0)\\
q_N(\theta_0) &= N\mathcal{E}\{v_N'^{T}(\theta_0, n_z)v_N'(\theta_0, n_z)\}\\
&\quad + 2\mathrm{herm}(\mathcal{E}\{\left(\frac{\partial\varepsilon(\theta, z_0)}{\partial\theta_0}\right)^T\}\mathcal{E}\{\Delta(\theta_0, n_z)v_N'(\theta_0, n_z)\})
\end{aligned} \qquad (17\text{-}32)$$

with $v_N(\theta, n_z) = \frac{1}{2N}\Delta^T(\theta, n_z)\Delta(\theta, n_z)$ and $\Delta(\theta, n_z) = \Lambda(\theta)M_1(\theta)n_z$.

Proof. See Appendix 17.E. □

The expression (17-32) for $\mathrm{Cov}(\delta_\theta(Z))$ is not tractable and will be approximated. Replacing n_z by υn_z (and, hence, C_{n_z} by $\upsilon^2 C_{n_z}$), it can be seen that the second term in the expression of $\mathrm{Cov}(\sqrt{N}\delta_\theta(z))$ decreases to zero faster than $V_N''^{-1}(\theta_0)$ as the signal-to-noise ratio increases to infinity ($\upsilon \to 0$). It makes it possible to approximate (17-32) for "sufficiently large" signal-to-noise ratios as

$$\mathrm{Cov}(\delta_\theta(z)) = V_{\mathrm{Markov}}''^{-1}(\theta_0)(I_{n_\theta} + O(\upsilon)) \qquad (17\text{-}33)$$

with $V_{\mathrm{Markov}}''^{-1}(\theta_0) = \upsilon^2 O(N^{-1})$ (Exercise 17.9).

If modeling errors are present, $e(\tilde{\theta}(z), z_0) \neq 0$, then the full expression (15-18) of the covariance matrix should be used. Replacing $e(\tilde{\theta}(z), z_0)$ by $\mu e(\tilde{\theta}(z), z_0)$ and n_z by υn_z, an approximation for "small" model errors ($\mu \to 0$) and "large" signal-to-noise ratios ($\upsilon \to 0$) is given by (Exercise 17.10)

$$\begin{aligned} \mathrm{Cov}(\delta_\theta(z)) &= C_\theta(I_{n_\theta} + O(\upsilon) + O(\mu) + O(\mu^2\upsilon^{-2})\lambda(z_0)) \\ C_\theta &= \left[\mathcal{E}\left\{\left(\frac{\partial \varepsilon(\theta, z_0)}{\partial \tilde{\theta}(z_0)}\right)^T \left(\frac{\partial \varepsilon(\theta, z_0)}{\partial \tilde{\theta}(z_0)}\right)\right\}\right]^{-1} = \upsilon^2 O(N^{-1}) \end{aligned} \tag{17-34}$$

where $\lambda(z_0) = 1$ for random z_0 and $\lambda(z_0) = 0$ for deterministic z_0. Formula (17-34) shows that in the presence of model errors ($\mu \neq 0$), the uncertainty of the estimated model parameters does not decrease to zero for random z_0 as the noise level υ tends to zero. Intuitively, this can be understood as follows: in the absence of observation noise, $n_z = 0$, the model errors still depend on the particular realization of z_0. Hence, $\hat{\theta}(z)$ depends on z_0, and $\mathrm{Cov}(\delta_\theta(z)) = \mu^2 O(N^{-1})$.

17.4.5 Asymptotic Efficiency

Comparing the Cramér-Rao lower bound (17-21) for normally distributed errors n_z and deterministic z_0 with the asymptotic covariance matrix (17-32) shows that the Markov estimates are, in general, asymptotically inefficient ($V_N''(\theta_0) = V_N''(\theta_0, z_0)$ for deterministic z_0). The inefficiency term $V_N''^{-1}(\theta_0)q_N(\theta_0)V_N''^{-1}(\theta_0)$ is, however, small w.r.t. $V_N''^{-1}(\theta_0)$ for sufficiently large signal-to-noise ratios (Exercise 17.9). For some noise covariance matrices the inefficiency term is zero.

Theorem 17.4 (Asymptotic Efficiency of $\hat{\theta}(z)$): Under the assumptions of Theorem 17.3 and Assumption 17.1, the Markov estimates (17-10) are asymptotically efficient if $\mathrm{rank}(C_{n_z}) = rN$ for any $N \geq N_0$.

Proof. See Appendix 17.F. □

The condition $\mathrm{rank}(C_{n_z}) = rN$ is automatically satisfied for signal models (Exercise 17.11). In frequency domain identification of multivariable systems, it implies that the number of noncoherent noise sources must be equal to the number of outputs. Note that Theorem 17.4 is not valid for the estimates based on the simplified Markov cost functions (17-12) and (17-13).

17.4.6 Robustness

The consistency, asymptotic normality, convergence rate, and asymptotic bias properties of the Markov estimator (17-10) for (block) diagonal models (17-2) are robust w.r.t. to the knowledge of some nondiagonal parts of C_{n_z} (compare the simplified cost functions (17-12) and (17-13) with (17-8)). This is not the case for the asymptotic efficiency: removing the nondiagonal elements of C_{n_z} increases the uncertainty of the estimates.

17.4.7 Practical Calculation of Uncertainty Bounds

Theorems 15.29 and 17.4 and formulas (17-33), (17-34) require knowledge of the true observations z_0 and the (true) model parameters θ_0 or $\tilde{\theta}(z)$, which are not available. Approximations of the asymptotic covariance matrix are obtained by replacing z_0 by z and θ_0 or $\tilde{\theta}(z_0)$ by $\hat{\theta}(z)$. Mostly the following approximation is used:

$$\mathrm{Cov}(\underline{\hat{\theta}}(z)) \approx \left[\left(\frac{\partial\varepsilon(\theta, z)}{\partial\hat{\theta}(z)}\right)^T\left(\frac{\partial\varepsilon(\theta, z)}{\partial\hat{\theta}(z)}\right)\right]^{-1} \tag{17-35}$$

Note that the right-hand side of (17-35) is calculated in Newton-based minimization methods of the cost function (17-8).

Together with the results of Section 14.2, (17-35) allows the calculation of uncertainty bounds of any model-related quantity. For example, the uncertainty of $f(z, \hat{\theta}(z)) \in \mathbb{R}^m$ is found by linearizing $f(z, \hat{\theta}(z))$ at z_0, $\tilde{\theta}(z_0)$

$$f(z, \hat{\theta}(z)) \approx \frac{\partial f(z, \tilde{\theta}(z_0))}{\partial z_0}n_z + \frac{\partial f(z_0, \theta)}{\partial\tilde{\theta}(z_0)}(\hat{\theta}(z) - \tilde{\theta}(z_0)) \tag{17-36}$$

where $\hat{\theta}(z) - \tilde{\theta}(z_0)$ is given by (17-31). Calculating the covariance matrix of $f(z, \hat{\theta}(z))$ and replacing z_0 afterward by z and $\tilde{\theta}(z_0)$ by $\hat{\theta}(z)$ in this expression gives

$$\begin{aligned}\mathrm{Cov}(f(z, \hat{\theta}(z))) \approx{}& \left(\frac{\partial f(z, \hat{\theta}(z))}{\partial z}\right)C_{n_z}\left(\frac{\partial f(z, \hat{\theta}(z))}{\partial z}\right)^T + \left(\frac{\partial f(z, \theta)}{\partial\hat{\theta}(z)}\right)\mathrm{Cov}(\underline{\hat{\theta}}(z))\left(\frac{\partial f(z, \theta)}{\partial\hat{\theta}(z)}\right)^T\\ &+ 2\mathrm{herm}(\left(\frac{\partial f(z, \hat{\theta}(z))}{\partial z}\right)\mathrm{Cov}(n_z, \hat{\theta}(z) - \tilde{\theta}(z_0))\left(\frac{\partial f(z, \theta)}{\partial\hat{\theta}(z)}\right)^T)\end{aligned} \tag{17-37}$$

where, using (17-31) and (17-35), $\mathrm{Cov}(n_z, \hat{\theta}(z) - \tilde{\theta}(z_0))$ can be approximated as

$$\mathrm{Cov}(n_z, \hat{\theta}(z) - \tilde{\theta}(z_0)) \approx -C_{n_z}\left(\frac{\partial\varepsilon(\hat{\theta}(z), z)}{\partial z}\right)^T\left(\frac{\partial\varepsilon(\theta, z)}{\partial\hat{\theta}(z)}\right)\mathrm{Cov}(\underline{\hat{\theta}}(z))$$

17.5 RESIDUALS OF THE MODEL EQUATION

First we study the residuals for real observations and real model parameters. Afterward, the results are generalized to complex observations and complex model parameters.

17.5.1 Real Case

The weighted residual of the model equation, $\varepsilon(\hat{\theta}(z), z)$, is a random vector that depends directly on the errors n_z through the observations z and indirectly on these errors through the estimate $\hat{\theta}(z)$, which is a nonlinear function of n_z. To analyze its stochastic properties, we need assumptions on the square root $C_e^{-1/2}(\theta)[M_0(\theta)\ M_1(\theta)]$ of the weighting $W_N(\theta)$ (17-25) (convergence analysis) and on the true observations z_0 (existence of some moments).

Assumption 17.5 (Constraint on the Square Root of the Weighting): The $rN + 1$ by $sN + 1$ matrix $R_N(\theta) = \Lambda(\theta)[M_0(\theta)\ M_1(\theta)]$, with $\Lambda^T(\theta)\Lambda(\theta) = C_e^{-1}(\theta)$, satisfies $\|R_N(\theta)\|_p \le c < \infty$, with $p = 1, \infty$ and c an N-independent constant, for all N (∞ included) and any $\theta \in \Theta_r$. $R_N(\theta)$ is a continuous matrix function of θ in the compact set Θ_r.

Assumption 17.6 (Constraint on the Derivatives of the Square Root of the Weighting): The $rN+1$ by $sN+1$ matrix $R_N(\theta) = \Lambda(\theta)[M_0(\theta)\ M_1(\theta)]$, with $\Lambda^T(\theta)\Lambda(\theta) = C_e^{-1}(\theta)$, has continuous first- and second-order derivatives w.r.t. θ satisfying

$$\text{(a)}\ \left\|\frac{\partial R_N(\theta)}{\partial \theta_{[i]}}\right\|_p \le c_1 < \infty,\ i = 1, 2, \ldots, n_\theta$$

$$\text{(b)}\ \left\|\frac{\partial^2 R_N(\theta)}{\partial \theta_{[i]}\partial \theta_{[j]}}\right\|_p \le c_2 < \infty,\ i, j = 1, 2, \ldots, n_\theta$$

with $p = 1, \infty$ and c_1, c_2 N-independent constants, for $N = 1, 2, \ldots, \infty$ and $\theta \in \Theta_{\mathrm{r}}$.

Assumption 17.7 (True Observations): The true observations z_0 are uniformly bounded.

Lemma 17.8 (Convergence Rate Residuals): Under the assumptions of Theorem 15.21 and Assumptions 17.5 and 17.6(a), the residual $\varepsilon_{[i]}(\hat{\theta}(z), z)$ converges uniformly in prob. to $\varepsilon_{[i]}(\tilde{\theta}(z_0), z)$ at the rate $O_{\mathrm{p}}(N^{-1/2})$ in Θ_{r} as $N \to \infty$, $i = 1, 2, \ldots, rN$.

Proof. See Appendix 17.G. □

For large signal-to-noise ratios and small model errors, the convergence rate of the residuals can be refined.

Lemma 17.9 (Improved Convergence Rate Residual): Under the assumptions of Theorem 15.21 and Assumptions 17.5, 17.6, and 17.7, large signal-to-noise ratios $(\upsilon \to 0)$, and small model errors $(\mu \to 0)$, the residual $\varepsilon_{[i]}(\hat{\theta}(z), z)$, $i = 1, 2, \ldots, rN$, can be written as

$$\begin{aligned}
\varepsilon_{[i]}(\hat{\theta}(z), z) &= \varepsilon_{[i]}(\tilde{\theta}(z_0), z_0) + (Q_\varepsilon(z_0)\delta_\varepsilon(z))_{[i]} + O_{\mathrm{p}}(N^{-1/2})(\upsilon + \mu + \mu\upsilon^{-1}\lambda(z_0)) \\
Q_\varepsilon(z_0) &= I_{rN} - \left(\frac{\partial \varepsilon(\theta, z_0)}{\partial \tilde{\theta}(z_0)}\right)\left[\left(\frac{\partial \varepsilon(\theta, z_0)}{\partial \tilde{\theta}(z_0)}\right)^T\left(\frac{\partial \varepsilon(\theta, z_0)}{\partial \tilde{\theta}(z_0)}\right)\right]^{-1}\left(\frac{\partial \varepsilon(\theta, z_0)}{\partial \tilde{\theta}(z_0)}\right)^T \\
\delta_\varepsilon(z) &= \Delta(\tilde{\theta}(z_0), n_z) = \Lambda(\tilde{\theta}(z_0))M_1(\tilde{\theta}(z_0))n_z
\end{aligned} \tag{17-38}$$

where $\mathcal{E}\{\delta_\varepsilon(z)\} = 0$, $\mathrm{Cov}(\delta_\varepsilon(z)) = I_{rN}$ and

$$\begin{aligned}
\varepsilon_{[i]}(\tilde{\theta}(z_0), z_0) &= \mu\upsilon^{-1}O_{\mathrm{p}}(N^0) \\
(Q_\varepsilon(z_0)\delta_\varepsilon(z))_{[i]} &= O_{\mathrm{p}}(N^0) \\
Q_{\varepsilon[i,j]}(z_0) &= I_{rN[i,j]} + O_{\mathrm{p}}(N^{-1})
\end{aligned}$$

for $i = 1, 2, \ldots, rN$, with $\lambda(z_0) = 1$ for random z_0 and $\lambda(z_0) = 0$ for deterministic z_0. $Q_\varepsilon(z_0)$ is a symmetric idempotent matrix of rank $rN - n_\theta$.

Proof. See Appendix 17.I. □

If no model errors are present, $\varepsilon_{[i]}(\tilde{\theta}(z_0), z_0) = 0$, it follows from Lemma 17.8 that the residuals $\varepsilon(\hat{\theta}(z), z)$ are asymptotically white: $\mathrm{Cov}(\varepsilon(\theta_0, z)) = I_{rN}$. Therefore, we could think of verifying the presence of model errors through the sample correlation of the residuals

$$\hat{R}_{\varepsilon\varepsilon}(k) = \frac{1}{rN-k}\sum_{i=1}^{rN-k} \varepsilon_{[i]}(\hat{\theta}(z), z)\varepsilon_{[i+k]}(\hat{\theta}(z), z) \tag{17-39}$$

The following theorem shows that this makes sense, indeed.

Theorem 17.10 (Properties Sample Correlation): Under the assumptions of Theorem 15.21 and Assumptions 17.5, 17.6(a), the sample correlation $\hat{R}_{\varepsilon\varepsilon}(k)$ converges in prob. to

$$\frac{1}{rN-k}\sum_{i=1}^{rN-k} \mathcal{E}\{\, \varepsilon_{[i]}(\tilde{\theta}(z_0), z_0)\varepsilon_{[i+k]}(\tilde{\theta}(z_0), z_0)\} + \delta(k) \tag{17-40}$$

at the rate $O_p(N^{-1/2})$ as $N \to \infty$ ($\delta(k)$ is the Kronecker delta and k is fixed independent of N). If, in addition, Assumptions 15.1 ($P = \infty$) and 17.6(b) are valid, then $\hat{R}_{\varepsilon\varepsilon}(k)$ is asymptotically normally distributed. If no model errors are present (Assumptions 15.9, 15.10) and n_z is normally distributed (Assumption 17.1), then the variance of the truncated sample correlation $\underline{\hat{R}}_{\varepsilon\varepsilon}(k)$ equals, asymptotically,

$$\frac{1+\delta(k)}{rN-k} \tag{17-41}$$

Proof. See Appendix 17.J. □

Under the null hypothesis that no model errors are present, Theorem 17.10 makes it possible to verify whether or not the sample correlation is white within its uncertainty. This procedure is known as the whiteness test on the residuals.

Lemma 17.9 shows that in the absence of model errors, $\varepsilon_{[i]}(\tilde{\theta}(z_0), z_0) = 0$ and $\mu = 0$, n_θ linear dependences exist among the residuals $\varepsilon_{[i]}(\hat{\theta}(z), z)$, $i = 1, 2, \ldots, rN$ ($\mathrm{rank}(Q_\varepsilon(z_0)) = rN - n_\theta$). It explains why the expected value of $\underline{\hat{R}}_{\varepsilon\varepsilon}(0)$ approximately ($N \to \infty$, $\upsilon, \mu \to 0$) equals $\mathcal{E}\{\frac{1}{rN}\sum_{i=1}^{rN}((Q_\varepsilon(z_0)\delta_\varepsilon(z))_{[i]})^2\} = (rN - n_\theta)/(rN)$ (see Exercise 17.12) while Theorem 17.10 predicts the value 1. To compensate for this bias at lag zero, $\hat{R}_{\varepsilon\varepsilon}(k)$ and its standard deviation are often multiplied by $rN/(rN - n_\theta)$.

For deterministic z_0, the covariance matrix of the truncated residuals is given approximately by

$$\mathrm{Cov}(\underline{\varepsilon}(\hat{\theta}(z), z)) \approx \mathrm{Cov}(Q_\varepsilon(z_0)\delta_\varepsilon(z)) \approx I_{rN} - \left(\frac{\partial\varepsilon(\theta, z_0)}{\partial\tilde{\theta}(z_0)}\right)\mathrm{Cov}(\underline{\hat{\theta}}(z))\left(\frac{\partial\varepsilon(\theta, z_0)}{\partial\tilde{\theta}(z_0)}\right)^T \tag{17-42}$$

(see Exercise 17.13). It follows that the total uncertainty equals the uncertainty due to the observation noise n_z minus the uncertainty due to the estimated model parameters $\hat{\theta}$.

17.5.2 Complex Case

For *complex observations* $z \in \mathbb{C}^{sN}$ and *real or complex model parameters* the sample correlation of the residuals is defined as

$$\hat{R}_{\varepsilon\varepsilon}(k) = \frac{1}{rN-k}\sum_{i=1}^{rN-k} \varepsilon_{[i]}(\hat{\theta}(z), z)\bar{\varepsilon}_{[i+k]}(\hat{\theta}(z), z) \tag{17-43}$$

Theorem 17.10 is still valid for *circular complex distributed errors* $n_z \in \mathbb{C}^{sN}$ (see (14-12) and (14-13)) with the following modifications (proof: all formulas of Section 17.5.1 are valid for z_{re}, $n_{z\text{re}}$, and ε_{re}). $\hat{R}_{\varepsilon\varepsilon}(k)$ is asymptotically circular complex normally distributed except at lag zero, where it is asymptotically normally distributed. If no model errors are present, then $\hat{R}_{\varepsilon\varepsilon}(k)$ $(k \neq 0)$ is asymptotically circular complex normally distributed and the variance of the truncated sample correlation $\hat{R}_{\varepsilon\varepsilon}(k)$ equals, asymptotically, $1/(rN-k)$ (see Exercise 17.14). The multiplicative bias correcting factor for $\hat{R}_{\varepsilon\varepsilon}(k)$ (and its standard deviation) equals $rN/(rN-n_\theta/2)$ for real model parameters and $rN/(rN-n_\theta)$ for complex model parameters (see Exercise 17.15).

17.6 MEAN AND VARIANCE OF THE COST FUNCTION

First, the case of real observations and real model parameters is handled. Afterward, these results are generalized to complex observations and complex model parameters.

17.6.1 Real Case

This section studies the stochastic properties of the minimum of the cost function $V_{\text{Markov}}(\hat{\theta}(z), z)$. In particular, the contribution of the model errors $(e(\tilde{\theta}(z_0), z_0) \neq 0)$ and the noise n_z to $V_{\text{Markov}}(\hat{\theta}(z), z)$ is analyzed. The following lemma gives the properties of $V_{\text{Markov}}(\hat{\theta}(z), z)$ for a large number of observations $(N \to \infty)$, large signal-to-noise ratios $(\upsilon \to 0)$, and small model errors $(\mu \to 0)$.

Lemma 17.11 (Convergence Rate Cost Function): Under the assumptions of Theorem 15.21 and Assumptions 17.5, 17.6, and 17.7, large signal-to-noise ratios $(\upsilon \to 0)$, and small model errors $(\mu \to 0)$, the minimum of the cost function $V_{\text{Markov}}(\hat{\theta}(z), z)$ can be written as

$$\begin{aligned} V_{\text{Markov}}(\hat{\theta}(z), z) &= L(\tilde{\theta}(z_0), z) + (\upsilon + \mu + \mu\upsilon^{-2}(\upsilon + \mu)\lambda(z_0))O_{\text{p}}(N^0) \\ L(\tilde{\theta}(z_0), z) &= V_{\text{Markov}}(\tilde{\theta}(z_0), z_0) + \varepsilon^T(\tilde{\theta}(z_0), z_0)Q_\varepsilon(z_0)\delta_\varepsilon(z) + \frac{1}{2}\delta_\varepsilon^T(z)Q_\varepsilon(z_0)\delta_\varepsilon(z) \end{aligned} \tag{17-44}$$

where

$$\begin{aligned} V_{\text{Markov}}(\tilde{\theta}(z_0), z_0) &= \mu^2\upsilon^{-2}O_{\text{p}}(N) \\ \delta_\varepsilon^T(z)Q_\varepsilon(z_0)\delta_\varepsilon(z) &= O_{\text{p}}(N) \\ \varepsilon^T(\tilde{\theta}(z_0), z_0)\delta_\varepsilon(z) &= \mu\upsilon^{-1}O_{\text{p}}(N^{1/2}) \end{aligned}$$

with $\lambda(z_0) = 1$ for random z_0 and $\lambda(z_0) = 0$ for deterministic z_0.

Proof. See Appendix 17.K. □

As is the case for the estimates $\hat{\theta}(z)$, in general it is very difficult or impossible to show the existence of the expected value and the variance of the cost function $V_{\text{Markov}}(\hat{\theta}(z), z)$. However, the first- and second-order moments of $L(\tilde{\theta}(z_0), z)$ exist.

Theorem 17.12 (Properties Cost Function): Under Assumption 15.1 $(P = \infty)$ and the assumptions of Lemma 17.11, $V_{\text{Markov}}(\hat{\theta}(z), z)$ is asymptotically normally distributed. Under the assumptions of Lemma 17.11, we have

$$\mathcal{E}\{L(\tilde{\theta}(z_0), z)\} = \mathcal{E}\{V_{\text{Markov}}(\tilde{\theta}(z_0), z_0)\} + (rN - n_\theta)/2 \tag{17-45}$$

with n_θ the number of identifiable model parameters. If, in addition, the errors n_z are normally distributed (Assumption 17.1), then

$$\begin{aligned}\operatorname{var}(L(\tilde{\theta}(z_0), z)) = {} & \mathcal{E}\{\varepsilon^T(\tilde{\theta}(z_0), z_0)Q_\varepsilon(z_0)\varepsilon(\tilde{\theta}(z_0), z_0)\} + (rN - n_\theta)/2 \\ & + \operatorname{var}(V_{\text{Markov}}(\tilde{\theta}(z_0), z_0))\end{aligned} \tag{17-46}$$

For deterministic z_0, (17-46) reduces to

$$\operatorname{var}(L(\tilde{\theta}(z_0), z)) = \varepsilon^T(\tilde{\theta}(z_0), z_0)\varepsilon(\tilde{\theta}(z_0), z_0) + (rN - n_\theta)/2 \tag{17-47}$$

Proof. See Appendix 17.L. □

Theorem 17.12 shows that the model errors $(\varepsilon(\tilde{\theta}(z_0), z_0) \neq 0)$ increase not only the expected value of the cost function but also its uncertainty. This increase in uncertainty is larger for random than for deterministic observations z_0 $(\operatorname{var}(V_{\text{Markov}}(\tilde{\theta}(z_0), z_0)) = 0)$.

Under the null hypothesis that no model errors are present $(\mu = 0)$, Theorem 17.12 shows that $V_{\text{Markov}}(\hat{\theta}(z), z)$ is asymptotically $N((rN - n_\theta)/2, (rN - n_\theta)/2)$ distributed. It makes it possible to verify whether or not the cost function $V_{\text{Markov}}(\hat{\theta}(z), z)$ equals $(rN - n_\theta)/2$ within a given confidence level.

In the case of deterministic observations z_0, Theorem 17.12 allows estimation of the uncertainty of the cost function in the presence of model errors. Indeed, from (17-45) and Lemma 17.11 it follows that the contribution of the model errors to the cost function $V_{\text{Markov}}(\tilde{\theta}(z_0), z_0)$ can be estimated as

$$V_{\text{Markov}}(\tilde{\theta}(z_0), z_0) \approx \begin{cases} V_{\text{Markov}}(\hat{\theta}(z), z) - \dfrac{(rN - n_\theta)}{2} & V_{\text{Markov}}(\hat{\theta}(z), z) \geq \dfrac{(rN - n_\theta)}{2} \\ 0 & \text{elsewhere} \end{cases} \tag{17-48}$$

Substituting this expression in (17-47) gives an estimate of the variance of the cost function

$$\operatorname{var}(L_{\text{Markov}}(\tilde{\theta}(z_0), z_0)) \approx 2V_{\text{Markov}}(\hat{\theta}(z), z) - (rN - n_\theta)/2 \tag{17-49}$$

As a null hypothesis test already makes it possible to verify the presence of model errors, one could wonder why it is useful to know the uncertainty of the cost function in the presence of model errors. Formula (17-49) is useful for comparing the cost functions of two independent experiments, for example, to decide whether or not the model errors are significantly different in both experiments.

The variance expression (17-46) becomes intractable for non-Gaussian observation errors n_z. However, for deterministic observations z_0 and non-Gaussian n_z, it is still possible to give upper and lower bounds on the variance (Pintelon et al., 1997a).

17.6.2 Complex Case

For *complex observations* $z \in \mathbb{C}^{sN}$ and *circular complex errors* n_z, Theorem 17.12 and formulas (17-48), (17-49) are still valid with the following modifications. Replace $rN - n_\theta$ by $2rN - n_\theta$ for *real model parameters* and $rN - n_\theta$ by $2(rN - n_\theta)$ for *complex model parameters* (see Exercise 17.16).

17.7 MODEL SELECTION AND MODEL VALIDATION

An identification procedure typically consists of applying iteratively model selection and parameter estimation. The model selection (order estimation) is still the most critical step in the identification process and consists of detecting overmodeling as well as undermodeling. Overmodeling occurs if the considered model set includes the true model and if it is described by too many parameters. Undermodeling is, for example, due to unmodeled dynamics and/or nonlinear distortions in system identification or, for example, due to too small a number of sine waves and/or nonperiodic deterministic disturbances in signal modeling. This section describes the properties of several (classical) model selection methods.

17.7.1 Real Case

17.7.1.1 Detection of Overmodeling. The Akaike information criterion (AIC) and minimum description length (MDL) method select the model $M(\hat{\theta}(z), \hat{z})$ out of the model set $\mathbb{M}$ that minimizes the sum of the negative log-likelihood function of the parameters and a function that penalizes the use of a large number of parameters (Akaike, 1974; Rissanen, 1978; Liang et al., 1993). For model (17-1) and Gaussian-distributed errors n_z, they take the form

$$\text{AIC: } V_{\text{Markov}}(\hat{\theta}(z), z) + n_\theta \tag{17-50}$$

$$\text{MDL: } V_{\text{Markov}}(\hat{\theta}(z), z) + \frac{n_\theta}{2}\ln(\text{rank}(C_{n_z})) \tag{17-51}$$

with n_θ the number of identifiable model parameters (Appendix 17.M). Minimizing (17-50) and (17-51) over the set of models $\mathbb{M}$ ($V_{\text{Markov}}(\hat{\theta}(z), z)$ and n_θ vary over $\mathbb{M}$) gives the optimal model according to the AIC and MDL criteria, respectively. The AIC criterion is inconsistent because it selects too complex models (Kashyap, 1980), while the MDL criterion gives strongly consistent estimates of the order of ARMA models (Hannan, 1980).

Example 17.13: Consider the identification of the amplitudes A_k, phases ϕ_k, and frequency f_0 of the sum of h harmonically related sine waves (signal model (17-3) with $s = 1$): $M_{0[n]}(\theta) = \sum_{k=1}^{h} A_k \sin(k\omega_0 nT_s + \phi_k)$ with $n = 0, 1, \ldots, N-1$ and $\theta^T = [A_1 \ldots A_h \phi_1 \ldots \phi_h f_0]$. Hence, (17-50) and (17-51) apply with $n_\theta = 2h + 1$ and

$\text{rank}(C_{n_z}) = N$. According to the AIC or MDL principle, the optimal value of h is found by minimizing (17-50) or (17-51) w.r.t. $h \in \mathbb{N}$ ($V_{\text{Markov}}(\hat{\theta}(z), z)$ is a function of h). □

The AIC and MDL criteria have been derived by assuming implicitly, or explicitly, that the true model belongs to the model set (see Ljung, 1999 for AIC) and, therefore, are unable to detect undermodeling.

17.7.1.2 Detection of Undermodeling. Undermodeling can be detected by a null hypothesis test on the cost function (see Section 17.6): if $V_{\text{Markov}}(\hat{\theta}(z), z) > (rN - n_\theta)/2 + 2\sqrt{(rN - n_\theta)/2}$ then, with 95% confidence, model errors are present.

17.7.1.3 Model Validation. The whiteness test of the residuals (see Section 17.5) can be used as a model validation tool. If the sample correlation is not a delta function within its uncertainty, then model errors are present. If it is white within its uncertainty, then the model passes the validation test. This does not, however, mean that no model errors are present. Indeed, the whiteness test is insensitive to errors that behave as white noise in the residuals. Nonlinear distortions in system identification are an example of such errors (see Chapter 9). The presence of model errors in a validated model can be detected by a null hypothesis test on the cost function (see paragraph 17.7.1.2 of this section).

17.7.1.4 Model Selection Procedure. The following iterative model selection procedure results.

1. Choose an initial model set (model order).
2. Estimate the model parameters.
3. Validate the model using a whiteness test on the residuals (see Section 17.5). If the residuals are white or a user-defined criterion is satisfied, then go to 4, else increase the model complexity and go to 2.
4. Detect undermodeling using a null hypothesis test on the cost function (see Section 17.6). If the cost function lies in the interval $(rN - n_\theta)/2 \pm 2\sqrt{(rN - n_\theta)/2}$, then go to 5, else stop.
5. Detect overmodeling using the MDL criterion (17-51).

Possible user-defined criteria in step 3 are, for example, that the estimated contribution of the model errors $V_{\text{Markov}}(\tilde{\theta}(z_0), z_0)$ to the cost function (see (17-48)) is below a given level C, or, in system identification, that the maximal (relative) transfer function error is less than a given value ε. The proposed procedure starts with simple models and gradually increases the model complexity. Practice has shown that in most identification problems the iterative procedure stops at step 4 (the validated model still contains some model errors). This is quite natural because the proposed model set reflects our belief in what reality is. This belief is mostly (always?) an approximation of the true behavior.

17.7.2 Complex Case

The results of the real case are still valid for *complex observations* $z \in \mathbb{C}^{sN}$ and *circular complex errors* n_z with the following modifications: replace $\text{rank}(C_{n_z})$ by $2\text{rank}(C_{n_z})$. For *complex model parameters* n_θ is replaced by $2n_\theta$. For example, (17-51) becomes

$$\text{MDL: } V_{\text{Markov}}(\hat{\theta}(z), z) + \frac{n_\theta}{2}\ln(2\text{rank}(C_{n_z})) \tag{17-52}$$

for *real model parameters* $\theta \in \mathbb{R}^{n_\theta}$ and

$$\text{MDL: } V_{\text{Markov}}(\hat{\theta}(z), z) + n_\theta \ln(2\text{rank}(C_{n_z})) \tag{17-53}$$

for *complex model parameters* $\theta \in \mathbb{C}^{n_\theta}$.

Example 17.14: Consider the frequency domain identification of a linear time-invariant MIMO system from periodic steady-state measurements, and assume that the input and output spectra are observed at F frequencies (model (17-5) with $r = n_y$, $q = n_u$, and $I_k(\theta) = 0$; see also Chapter 5, left matrix fraction description (5-56). In this case (17-52) applies with $n_\theta = n_a n_y^2 + (n_b + 1)n_y n_u$ and $\text{rank}(C_{n_z}) = n_{\text{nc}}F$ with n_{nc} the number of noncoherent noise sources. For example, $n_{\text{nc}} = n_u + n_y$ for errors-in-variables problems (all observations are disturbed by noise) and $n_{\text{nc}} = n_y$ for output error problems (the inputs are exactly known). According to the MDL principle, the optimal values of n_a and n_b are found by minimizing (17-52) w.r.t. $n_a, n_b \in \mathbb{N}$ ($V_{\text{Markov}}(\hat{\theta}(z), z)$ is a function of n_a, n_b). □

17.8 EXERCISES

17.1. Consider a scalar (SISO) discrete-time system and assume that N samples of the input and output signals are available. Show that the time domain model can be written in the form (17-4) with $r = q = 1$.

17.2. Consider a multivariable system with n_u inputs and n_y outputs. Assume that the input and output signals are periodic and that the DFT spectra are available at F frequencies. Show that the frequency domain model equations can be written in the form (17-2) with $r = 2n_y$, $s = 2(n_u + n_y)$, and $N = F$ (hint: use the left matrix fraction description (5-56) and apply Lemma 13.4).

17.3. Repeat Exercise 17.2 for arbitrary excitations (hint: use the left matrix fraction description (5-56) generalized for arbitrary excitations).

17.4. Solve the constraint minimization problem (17-6) assuming that C_{n_z} is a regular matrix.

17.5. Assume that $M_0(\theta)$ in (17-1) is linear in some of the model parameters, say $\psi \in \mathbb{R}^{n_\psi}$, and that $M_1(\theta)$ is independent of ψ: $M_0(\theta) = M(\xi)\psi$ and $M_1(\theta) = M_1(\xi)$ where $\xi \in \mathbb{R}^{n_\theta - n_\psi}$, $M(\xi) \in \mathbb{R}^{rN \times n_\psi}$ and $\theta^T = [\psi^T \xi^T]$. This form, with $M_1(\theta) = I_{rN}$, is typically encountered in signal models (Cadzow, 1990; Van den Bos and Swarte, 1993). First, show that the cost function (17-8), after elimination of ψ, becomes

$$V_{\text{Markov}}(\xi, z) = \frac{1}{2}(R(\xi)\Lambda(\xi)M_1(\xi)z)^T(R(\xi)\Lambda(\xi)M_1(\xi)z)$$

where $\Lambda(\xi) \in \mathbb{R}^{rN \times rN}$ satisfies $\Lambda^T(\xi)\Lambda(\xi) = C_e(\xi)$ and $R(\xi) \in \mathbb{R}^{rN \times rN}$ is an idempotent matrix of rank $rN - n_\psi$, $R(\xi) = I_{rN} - P(\xi)(P^T(\xi)P(\xi))^{-1}P^T(\xi)$ with $P(\xi) = \Lambda(\xi)M(\xi)$. Show that the Markov estimate of ψ is given by

$$\hat{\psi} = -(M^T(\hat{\xi})C_e^{-1}(\hat{\xi})M(\hat{\xi}))^{-1}M^T(\hat{\xi})C_e^{-1}(\hat{\xi})M_1(\hat{\xi})z$$

Next, show that the cost function $V_{\text{Markov}}(\xi, z)$ can be written as

$$V_{\text{Markov}}(\xi, z) = \frac{1}{2}\varepsilon^T(\xi, z)\varepsilon(\xi, z)$$

where $\varepsilon(\xi, z) \in \mathbb{R}^{rN - n_\psi}$ equals $\varepsilon(\xi, z) = [I_{rN - n_\psi} \; 0]V^T(\xi)\Lambda(\xi)M_1(\xi)z$, and $V(\xi)$ is the orthogonal matrix of the eigenvectors of $R(\xi)$. Finally, show that $\text{Cov}(\varepsilon(\xi, n_z)) = I_{rN - n_\psi}$.

17.6. Prove the Cramér-Rao lower bound (17-22) for complex observations z, circular complex distributed errors n_z, and real model parameters θ (hint: first show that $V_{\text{Markov}}(\theta, z) = \frac{1}{2}e_{\text{re}}^T(\theta, z)C_{e_{\text{re}}}^{-1}(\theta)e_{\text{re}}(\theta, z) = e^H(\theta, z)C_e^{-1}(\theta)e(\theta, z)$ using Lemmas 13.3 and 13.4).

17.7. Prove the Cramér-Rao lower bound (17-23) for complex observations z, circular complex distributed errors n_z, and complex model parameters θ, assuming that $e(\theta, z)$ is an analytic function of θ (hint: rewrite (17-22) using $\frac{\partial f(\theta)}{\partial \text{Re}(\theta)} = \frac{\partial f(\theta)}{\partial \theta}$, $\frac{\partial f(\theta)}{\partial \text{Im}(\theta)} = j\frac{\partial f(\theta)}{\partial \theta}$, $\text{Re}(jX) = \text{Im}(X)$ and definition (13-40)).

17.8. Consider the model of Exercise 17.5 and show that the CR bound of the parameters ξ is given by

$$CR^{-1}(\xi_0) = Fi(\xi_0) = V_{\text{Markov}}''(\xi_0, z_0) = \left(\frac{\partial \varepsilon(\xi, z_0)}{\partial \xi_0}\right)^T\left(\frac{\partial \varepsilon(\xi, z_0)}{\partial \xi_0}\right)$$

(hint: start from the Fisher information matrix $Fi(\psi_0, \xi_0)$, apply the inverse of block matrices (13-8), and use the results of Exercise 17.5).

17.9. Show that the asymptotic covariance matrix is given by (17-33) as the signal-to-noise ratio increases to infinity (hint: show that $V_N''^{-1}(\theta_0) = O(\upsilon^2)$, $q_N(\theta_0) = O(\upsilon^{-1})$).

17.10. Show that the asymptotic covariance matrix of the model parameters is given by (17-34) for "small" model errors and "large" signal-to-noise ratios (hint: use $\varepsilon(\theta, z) = \varepsilon(\theta, z_0) + \Delta(\theta, n_z)$, $\mathcal{E}\{\Delta(\theta, n_z)\} = 0$, $\mathcal{E}\{\Delta(\theta, n_z)\Delta^T(\theta, n_z)\} = I_{n_\theta}$, $V'(\tilde{\theta}(z_0), z_0) = 0$ for deterministic z_0, $\varepsilon(\tilde{\theta}(z_0), z_0) = O(\mu\upsilon^{-1})$, and $\Delta(\tilde{\theta}(z_0), n_z) = O(\mu^0\upsilon^0)$).

17.11. Show that the rank condition $\text{rank}(C_{n_z}) = rN$ in Theorem 17.4 is automatically satisfied for signal models.

17.12. Show that $\mathcal{E}\{\frac{1}{rN}\sum_{i=1}^{rN}((Q_\varepsilon(z_0)\delta_\varepsilon(z))_{[i]})^2\} = (rN - n_\theta)/(rN)$ (hint: use the properties of $Q_\varepsilon(z_0)$ and $\delta_\varepsilon(z)$ given in Lemma 17.9).

17.13. Show that the covariance matrix of $Q_\varepsilon(z_0)\delta_\varepsilon(z)$ is given by (17-42) (hint: use the properties of $Q_\varepsilon(z_0)$ and $\delta_\varepsilon(z)$ given in Lemma 17.9 and approximation (17-34)).

17.14. Consider the case of complex observations z and circular complex distributed errors n_z. Prove that in the absence of model errors $\hat{R}_{\varepsilon\varepsilon}(k)$, $k \neq 0$, is asymptotically circular complex normally distributed $N^c(0, 1/(rN - k))$ (hint: follow the lines of part 3 of Appendix 17.J and show that $\text{cum}(\tilde{R}_{\varepsilon\varepsilon}(k), \tilde{R}_{\varepsilon\varepsilon}(k)) = 1/(rN - k)$, $\text{cum}(\tilde{R}_{\varepsilon\varepsilon}(k), \tilde{R}_{\varepsilon\varepsilon}(k)) = \text{cum}(\tilde{R}_{\varepsilon\varepsilon}(k), \tilde{R}_{\varepsilon\varepsilon}(k)) = 0$ with $\tilde{R}_{\varepsilon\varepsilon}(k) = (rN - k)^{-1}\sum_{i=1}^{rN-k}\delta_{\varepsilon[i]}(z)\bar{\delta}_{\varepsilon[i+k]}(z)$, and where $\delta_\varepsilon(z)$ defined in Lemma 17.9 is circular complex normally distributed).

17.15. Consider the case of complex observations z and circular complex distributed errors n_z. Show, using Lemma 17.9, that the expected value of $\hat{R}_{\varepsilon\varepsilon}(0)$ is approximately ($N \to \infty$, $\upsilon, \mu \to 0$) $(rN - n_\theta/2)/rN$ and $(rN - n_\theta)/rN$ for, respectively, real and complex model parameters (hint: replace z by z_{re}, and θ by θ_{re} for complex model parameters, and show that $\varepsilon(\theta, z_{\text{re}}) = \sqrt{2}\varepsilon_{\text{re}}(\theta, z)$ $(\varepsilon(\theta_{\text{re}}, z_{\text{re}}) = \sqrt{2}\varepsilon_{\text{re}}(\theta, z))$, next follow the lines of Section 17.5).

17.16. Consider the case of complex observations z and circular complex distributed errors n_z. Show, using Lemma 17.11, that Theorem 17.12 and formulas (17-48), (17-49) are still valid, where $rN - n_\theta$ is replaced by $2rN - n_\theta$ for real model parameters and $rN - n_\theta$ by $2(rN - n_\theta)$ for complex model parameters (hint: use the hint of Exercise 17.15).

17.9 APPENDIXES

Appendix 17.A: Constrained Minimization (17-6)

Expressing the stationarity of the cost function (17-6) w.r.t. z_{p} and λ gives

$$-C_{n_z}^{+}(z-z_{\mathrm{p}})+M_1^T(\theta)\lambda = 0 \tag{17-54}$$

$$M_0(\theta)+M_1(\theta)z_{\mathrm{p}} = 0 \tag{17-55}$$

λ is solved from (17-54) by left multiplication with $M_1(\theta)C_{n_z}$

$$\lambda = (M_1(\theta)C_{n_z}M_1^T(\theta))^{-1}M_1(\theta)C_{n_z}C_{n_z}^{+}(z-z_{\mathrm{p}}) \tag{17-56}$$

Because C_{n_z} is a symmetric positive semidefinite matrix, it can be decomposed into singular values as $C_{n_z} = U\Sigma U^T$, where U is an orthogonal matrix ($U^TU = UU^T = I_{sN}$) and Σ a diagonal matrix ($\mathrm{rank}(\Sigma) = \mathrm{rank}(C_{n_z})$) containing the sorted singular values (see Section 13.4). Defining

$$U^Tz_{\mathrm{p}} = \begin{bmatrix} W_{1\mathrm{p}} \\ W_{20} \end{bmatrix} \tag{17-57}$$

where $\dim(W_{1\mathrm{p}}) = \mathrm{rank}(C_{n_z})$, it follows that

$$z_{\mathrm{p}} = U\begin{bmatrix} W_{1\mathrm{p}} \\ W_{20} \end{bmatrix} \text{ and } C_{n_z}C_{n_z}^{+}z_{\mathrm{p}} = U\begin{bmatrix} I_{r_c} & 0 \\ 0 & 0 \end{bmatrix}U^Tz_{\mathrm{p}} = U\begin{bmatrix} W_{1\mathrm{p}} \\ 0 \end{bmatrix} \tag{17-58}$$

with $r_c = \mathrm{rank}(C_{n_z})$. $W_{1\mathrm{p}}$ stands for the (linear combination of) measurements lying in the regular space of C_{n_z} and is estimated, while W_{20} represents the (linear combination of) measurements lying in the null space of C_{n_z} and is known exactly. Hence, we have

$$z = U\begin{bmatrix} W_1 \\ W_{20} \end{bmatrix} \quad \text{and} \quad C_{n_z}C_{n_z}^{+}z = U\begin{bmatrix} W_1 \\ 0 \end{bmatrix} \tag{17-59}$$

Rewriting (17-55) and (17-56) using (17-58) and (17-59) makes it possible to eliminate $W_{1\mathrm{p}}$ in (17-56)

$$\lambda = (M_1(\theta)C_{n_z}M_1^T(\theta))^{-1}(M_0(\theta)+M_1(\theta)z) \tag{17-60}$$

Substituting (17-60) into (17-54) and left multiplication of (17-54) by C_{n_z} gives (17-7). Using (17-58), it can be seen that (17-7) is independent of W_{20}, which is known exactly, and, hence, makes it possible to estimate the measurements $W_{1\mathrm{p}}$ lying in the regular space of C_{n_z} only. Because $C_{n_z}^{+} = (C_{n_z}^{+}C_{n_z})C_{n_z}^{+}(C_{n_z}C_{n_z}^{+})$ (see Section 13.5, properties 1 and 2) and $C_{n_z}^{+T} = C_{n_z}^{+}$, the cost function (17-6), where z_{p} satisfies (17-55), can be written as

$$\frac{1}{2}[C_{n_z}C_{n_z}^{+}(z-z_{\mathrm{p}})]^T C_{n_z}^{+}[C_{n_z}C_{n_z}^{+}(z-z_{\mathrm{p}})] \tag{17-61}$$

Substituting (17-7) into (17-61) using $C_{n_z}C_{n_z}^{+}C_{n_z} = C_{n_z}$ gives (17-8) directly. □

Appendix 17.B: Proof of the Cramér-Rao Lower Bound for Semilinear Models

The proof consists of two parts. In the first part we reduce the sN parameters z_{p} to the $r_c - rN$ parameters x_0, with $r_c = \operatorname{rank}(C_{n_z})$. In the second part the Fisher information matrix $Fi(x_0, \theta_0)$ of the observations x and the model parameters θ is reduced to the Fisher information matrix $Fi(\theta_0)$.

(i) The known observations lie in the null space of C_{n_z} ($sN - r_c$ parameters) and are separated from the unknown observations by decomposing the parameter vector z_{p} as in (17-58). Using (17-58), with $U = [U_1 U_2]$ and $U_1 \in \mathbb{R}^{sN \times r_c}$, the constraint (17-55) can be written as

$$M_1(\theta)U_1W_{1\mathrm{p}} = -M_0(\theta) - M_1(\theta)U_2W_{20} \tag{17-62}$$

$C_e(\theta) = M_1(\theta)C_{n_z}M_1^T(\theta)$ has, by assumption, full rank rN and, therefore, $r_c \geq rN$ and $\operatorname{rank}(M_1(\theta)U_1) = rN$. Because $r_c \geq rN$, (17-62) has, in general, infinitely many solutions for $W_{1\mathrm{p}}$. They can be found by adding one particular solution of (17-62) to the solution of the homogenous part of (17-62). The solution W_{1h} of the homogenous part lies in the null space of $M_1(\theta)U_1 \in \mathbb{R}^{rN \times r_c}$. Because $M_1(\theta)U_1$ has full rank rN, it can be written as

$$W_{1h} = Q(\theta)x \tag{17-63}$$

with $x \in \mathbb{R}^{r_c - rN}$ and where $Q(\theta) \in \mathbb{R}^{r_c \times (r_c - rN)}$ satisfies $M_1(\theta)U_1Q(\theta) = 0$ (see Section 13.4.1). It can easily be verified that

$$W_{1p}(\theta) = -M^T(\theta)(M(\theta)M^T(\theta))^{-1}(M_0(\theta) + M_1(\theta)U_2W_{20}) \tag{17-64}$$

where $M(\theta) = M_1(\theta)U_1$, is a particular solution of (17-62). The complete solution of (17-62) equals $W_{1\mathrm{p}} = W_{1h} + W_{1p}(\theta)$ so that in (17-58) z_{p} can be written as

$$z_{\mathrm{p}} = z_{\mathrm{p}}(x, \theta) = U_1Q(\theta)x + U_1\tilde{W}_{1p}(\theta) + U_2W_{20} \tag{17-65}$$

with $x \in \mathbb{R}^{r_c - rN}$, which concludes the first part of the proof. The following property of the function $z_{\mathrm{p}}(x, \theta)$ will be used in the second part of the proof:

$$M_1(\theta)\frac{\partial z_{\mathrm{p}}(x, \theta)}{\partial x} = M_1(\theta)U_1Q(\theta) = 0 \tag{17-66}$$

(ii) The second part of the proof starts with the calculation of the Fisher information matrix $Fi(x_0, \theta_0)$. Applying (14-85) to (17-16), using $z_{\mathrm{p}}(x_0, \theta_0) = z_0$, gives

$$\begin{aligned} F_{xx} &= \left(\frac{\partial z_p(x,\theta)}{\partial x}\right)^T C^+_{n_z}\left(\frac{\partial z_p(x,\theta)}{\partial x}\right)\Bigg|_{x = x_0,\ \theta = \theta_0} \\ F_{x\theta} &= \left(\frac{\partial z_p(x,\theta)}{\partial x}\right)^T C^+_{n_z}\left(\frac{\partial z_p(x,\theta)}{\partial \theta}\right)\Bigg|_{x = x_0,\ \theta = \theta_0} \\ F_{\theta\theta} &= \left(\frac{\partial z_p(x,\theta)}{\partial \theta}\right)^T C^+_{n_z}\left(\frac{\partial z_p(x,\theta)}{\partial \theta}\right)\Bigg|_{x = x_0,\ \theta = \theta_0} \end{aligned} \qquad (17\text{-}67)$$

C_{n_z} and $C^+_{n_z}$ can be decomposed into singular values as

$$\begin{aligned} C_{n_z} &= U\begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix} U^T = U_1 \Sigma_1 U_1^T \\ C^+_{n_z} &= U\begin{bmatrix} \Sigma_1^{-1} & 0 \\ 0 & 0 \end{bmatrix} U^T = U_1 \Sigma_1^{-1} U_1^T \end{aligned} \qquad (17\text{-}68)$$

where Σ_1 contains the nonzero singular values of C_{n_z}. Putting (17-67) into (17-20) by using (17-68) gives, after some calculations,

$$Fi(\theta_0) = G^T[I_{r_c} - F(F^TF)^{-1}F^T]G \qquad (17\text{-}69)$$

with $G = \Sigma_1^{-1/2} U_1^T \partial z_p(x_0, \theta)/\partial\theta_0$ and $F = \Sigma_1^{-1/2} Q(\theta_0)$. Defining $E = M_1(\theta_0) U_1 \Sigma_1^{1/2}$, it follows from (17-66) that $EF = 0$. Because the matrices $E \in \mathbb{R}^{rN \times r_c}$ and $F \in \mathbb{R}^{r_c \times (r_c - rN)}$ have, respectively, full rank rN and $r_c - rN$ and $EF = 0$, it follows that the conditions of Theorem 13.2 are fulfilled for the symmetric idempotent matrices $F(F^TF)^{-1}F^T$ and $E^T(EE^T)^{-1}E$. Hence,

$$I_{r_c} - F(F^TF)^{-1}F^T = E^T(EE^T)^{-1}E \qquad (17\text{-}70)$$

so that (17-69) can be simplified as

$$Fi(\theta_0) = (EG)^T(EE^T)^{-1}(EG) \qquad (17\text{-}71)$$

Working out EG gives

$$\begin{aligned} EG &= M_1(\theta_0)\frac{\partial U_1 U_1^T z_p(x_0, \theta)}{\partial \theta_0} && (U_1 \text{ is independent of } \theta) \\ &= M_1(\theta_0)\frac{\partial (z_p(x_0, \theta) - U_2 W_{20})}{\partial \theta_0} && ((17\text{-}65) \text{ with } U_1^T U_1 = I_{r_c} \text{ and } U_1^T U_2 = 0) \\ &= -M_1(\theta_0)\frac{\partial (z_0 - z_p(x_0, \theta))}{\partial \theta_0} && (U_2,\ W_{20},\ z_0 \text{ are independent of } \theta) \end{aligned}$$

$$= -\frac{\partial M_1(\theta)(z_0 - z_{\mathrm{p}}(x_0, \theta))}{\partial \theta_0} \qquad (z_0 = z_{\mathrm{p}}(x_0, \theta_0))$$

$$= -\frac{\partial(M_0(\theta) + M_1(\theta)z_0)}{\partial \theta_0} \qquad ((17\text{-}55)\text{: } M_1(\theta)z_{\mathrm{p}}(x_0, \theta) = -M_0(\theta))$$

Substituting this result into (17-71), taking into account that $EE^T = C_e(\theta_0)$ (see (17-9)), gives

$$Fi(\theta_0) = \left(\frac{\partial e(\theta, z_0)}{\partial \theta_0}\right)^T C_e^{-1}(\theta_0)\left(\frac{\partial e(\theta, z_0)}{\partial \theta_0}\right) \tag{17-72}$$

Using $e(\theta_0, z_0) = 0$, the two other expressions in (17-21) follow directly. □

Appendix 17.C: Markov Estimates of the Observations for Signal Models andTransfer Function Models with Known Input

For signal models, we have $M_1(\theta) = -I_{sN}$, $C_e(\theta) = C_{n_z}$ (see (17-9)), and $C_{n_z}^{+} = C_{n_z}^{-1}$ (Assumption 15.2 implies that $C_e(\theta)$ is regular), and (17-11) becomes

$$\hat{z} = M_0(\hat{\theta}(z)) \tag{17-73}$$

For transfer function models with known input, we have

$$C_{n_z} = \begin{bmatrix} C_{n_y} & 0 \\ 0 & 0 \end{bmatrix} \Rightarrow C_{n_z}C_{n_z}^{+} = \begin{bmatrix} I_{rN} & 0 \\ 0 & 0 \end{bmatrix} \tag{17-74}$$

Taking into account (17-4), (17-11) becomes

$$\begin{bmatrix} \hat{y} \\ 0 \end{bmatrix} = \begin{bmatrix} A^{-1}(\hat{\theta}(z))B(\hat{\theta}(z))u_0 \\ 0 \end{bmatrix} \tag{17-75}$$

which concludes the proof. □

Appendix 17.D: Proof of the Convergence Rate of the Markov Estimates for Large Signal-to-Noise Ratios and Small Model Errors (Theorem 17.2)

$\delta_\theta(z)$ defined in expression (15-14) will be elaborated for the Markov estimator, assuming large signal-to-noise ratios and small model errors. Therefore, n_z is replaced by υn_z (and, hence, C_{n_z} by $\upsilon^2 C_{n_z}$) with $\upsilon \to 0$, and $e(\tilde{\theta}(z), z_0)$ by $\mu e(\tilde{\theta}(z), z_0)$ with $\mu \to 0$. The proof consists of two parts: part one studies $\delta_\theta(z) = \Delta_\theta(z) + \partial_\theta(z)$ and part two $b_\theta(z)$.

(i) The Hessian of the expected value of the cost function can be written as

$$V_N''(\tilde{\theta}(z_0)) = \frac{1}{N}\left(\frac{\partial\varepsilon(\theta, z_0)}{\partial\tilde{\theta}(z_0)}\right)^T\left(\frac{\partial\varepsilon(\theta, z_0)}{\partial\tilde{\theta}(z_0)}\right) + \frac{1}{N}\sum_{k=1}^{rN} \varepsilon_{[k]}(\tilde{\theta}(z_0), z_0)\frac{\partial^2\varepsilon_{[k]}(\theta, z_0)}{\partial\tilde{\theta}(z_0)^2} \quad (17\text{-}76)$$

where the first and second terms in the right-hand side are, respectively, an $O(\upsilon^{-2})$ and $O(\upsilon^{-2}\mu)$. Hence,

$$V_N''^{-1}(\tilde{\theta}(z_0)) = \left(\frac{1}{N}\left(\frac{\partial\varepsilon(\theta, z_0)}{\partial\tilde{\theta}(z_0)}\right)^T\left(\frac{\partial\varepsilon(\theta, z_0)}{\partial\tilde{\theta}(z_0)}\right)\right)^{-1} + \upsilon^2 O(\mu) \quad (17\text{-}77)$$

where the first term in the right-hand side is an $O(\upsilon^2)$. Using $\varepsilon(\theta, z) = \varepsilon(\theta, z_0) + \Delta(\theta, n_z)$, with $\Delta(\theta, n_z) = \Lambda(\theta)M_1(\theta)n_z$ (see (17-9)), gives

$$\begin{aligned} V_N'^T(\tilde{\theta}v(z_0), z) = {} & \frac{1}{N}\left(\frac{\partial\varepsilon(\theta, z_0)}{\partial\tilde{\theta}(z_0)}\right)^T \varepsilon(\tilde{\theta}(z_0), z_0) + \frac{1}{N}\left(\frac{\partial\varepsilon(\theta, z_0)}{\partial\tilde{\theta}(z_0)}\right)^T \Delta(\tilde{\theta}(z_0), n_z) \\ & \frac{1}{N}\left(\frac{\partial\Delta(\theta, n_z)}{\partial\tilde{\theta}(z_0)}\right)^T \varepsilon(\tilde{\theta}(z_0), z_0) + \frac{1}{N}\left(\frac{\partial\Delta(\theta, n_z)}{\partial\tilde{\theta}v(z_0)}\right)^T \Delta(\tilde{\theta}(z_0), n_z) \end{aligned} \quad (17\text{-}78)$$

Note that the expected value of each of the terms in the right-hand side of (17-78) is zero. This is evident for the first three terms. For the transpose of the fourth term, we find

$$\mathcal{E}\{\Delta^T(\tilde{\theta}(z_0), n_z)\frac{\partial\Delta(\theta, n_z)}{\partial\tilde{\theta}(z_0)}\} = \frac{\partial\mathcal{E}\{\Delta^T(\theta, n_z)\Delta(\theta, n_z)\}}{\partial\tilde{\theta}(z_0)} = \frac{\partial rN/2}{\partial\tilde{\theta}(z_0)} = 0$$

Applying the law of large numbers for mixing sequences (see Section 14.9, version 3) to each of the terms in the right-hand side of (17-78) shows that they are, respectively, an $O_p(\mu\upsilon^{-2}N^{-1/2})$, $O_p(\upsilon^{-1}N^{-1/2})$, $O_p(\upsilon^{-1}\mu N^{-1/2})$, and $O_p(N^{-1/2})$. Because $V_N'(\tilde{\theta}(z_0), z_0) = 0$ for deterministic z_0, (17-78) becomes

$$V_N'^T(\tilde{\theta}(z_0), z) = \upsilon^{-2}(\upsilon + \mu\lambda(z_0))O_p(N^{-1/2}) \quad (17\text{-}79)$$

with $\lambda(z_0) = 1$ for random z_0 and $\lambda(z_0) = 0$ for deterministic z_0. Combining (17-77) and (17-78) with $\Delta(\theta, n_z) = (\partial\varepsilon(\tilde{\theta}(z_0), z)/\partial z)n_z$, gives $\delta_\theta(z) = \Delta_\theta(z) + \partial_\theta(z)$ with $\mathcal{E}\{\Delta_\theta(z)\} = 0$, $\mathcal{E}\{\partial_\theta(z)\} = 0$, $\Delta_\theta(z) = \upsilon O_p(N^{-1/2})$, and $\partial_\theta(z) = (\upsilon^2 + \upsilon\mu + \mu\lambda(z_0))O_p(N^{-1/2})$.

(ii) From the proof of Theorem 15.21 (see Appendix 15.E), it follows that the $b_\theta(z)$ term stems from the difference $V_N''(\widehat{\theta}, z) - V_N''(\tilde{\theta}(z_0)) = O_p(N^{-1/2})$,

$$b_\theta(z) = [V_N''^{-1}(\widehat{\theta}, z) - V_N''^{-1}(\tilde{\theta}(z_0))]V_N'^T(\tilde{\theta}(z_0), z) \quad (17\text{-}80)$$

This expression will be refined. Applying the mean value theorem to $V_N''(\widehat{\theta}, z)$ at the points $\widehat{\theta}$, $\tilde{\theta}(z_0)$ gives

$$V_N''(\widehat{\theta}, z) = V_N''(\tilde{\theta}(z_0), z) + \sum_{k=1}^{n_\theta} \frac{\partial V_N''(\theta, z)}{\partial \theta_{1[k]}}(\widehat{\theta}_{[k]} - \tilde{\theta}_{[k]}(z_0)) \qquad (17\text{-}81)$$

with $\theta_1 = t_1\widehat{\theta} + (1-t_1)\tilde{\theta}(z_0)$ and $t_1 \in [0, 1]$. Because $\widehat{\theta} - \tilde{\theta}(z_0) = t(\hat{\theta}(z) - \tilde{\theta}(z_0))$ (see (15-10)), and $\hat{\theta}(z) - \tilde{\theta}(z_0) = \Delta_\theta(z) + \partial_\theta(z) + b_\theta(z)$ with $b_\theta(z) = O_p(N^{-1})$ (see previous paragraph), we have $\widehat{\theta} - \tilde{\theta}(z_0) = (\upsilon + \mu\lambda(z_0))O_p(N^{-1/2})$. Combined with $\partial V_N''(\theta, z)/\partial\theta_{1[k]} = \upsilon^{-2}O_p(N^0)$ (see Appendix 15.E), the second term in (17-81) becomes

$$\sum_{k=1}^{n_\theta} \frac{\partial V_N''(\theta, z)}{\partial \theta_{1[k]}}(\widehat{\theta}_{[k]} - \tilde{\theta}_{[k]}(z_0)) = \upsilon^{-2}(\upsilon + \mu\lambda(z_0))O_p(N^{-1/2}) \qquad (17\text{-}82)$$

The first term in (17-81) can be written as

$$V_N''(\tilde{\theta}(z_0), z) = V_N''(\tilde{\theta}(z_0), z_0) + \frac{1}{N}\frac{\partial^2 \varepsilon^T(\theta, z_0)\Delta(\theta, n_z)}{\partial \tilde{\theta}(z_0)^2} + v_N''(\tilde{\theta}(z_0), n_z) \qquad (17\text{-}83)$$

with $v_N(\theta, n_z) = \frac{1}{2N}\Delta^T(\theta, z_0)\Delta(\theta, n_z)$. $V_N''(\tilde{\theta}(z_0), z_0)$ converges w.p. 1 to $V_N''(\tilde{\theta}(z_0))$ for random z_0 and $V_N''(\tilde{\theta}(z_0), z_0) = V_N''(\tilde{\theta}(z_0))$ for deterministic z_0 (Lemma 15.17):

$$V_N''(\tilde{\theta}(z_0), z_0) = V_N''(\tilde{\theta}(z_0)) + \upsilon^{-2}\lambda(z_0)O_p(N^{-1/2}) \qquad (17\text{-}84)$$

Using Lemma 15.17, it follows that the second and third terms in the right-hand side of (17-83) are, respectively, an $O_p(\upsilon^{-1}N^{-1/2})$ and $O_p(\upsilon^0 N^{-1/2})$. Hence,

$$V_N''(\tilde{\theta}(z_0), z) = V_N''(\tilde{\theta}(z_0)) + (\upsilon^{-1} + \upsilon^{-2}\lambda(z_0))O_p(N^{-1/2}) \qquad (17\text{-}85)$$

Combining (17-81), (17-82), and (17-85), using $V_N''(\tilde{\theta}(z_0)) = \upsilon^{-2}O_p(N^0)$, gives

$$\begin{aligned} V_N''(\widehat{\theta}, z) &= V_N''(\tilde{\theta}(z_0)) + (\upsilon^{-1} + \upsilon^{-2}\lambda(z_0))O_p(N^{-1/2}) \\ \Rightarrow V_N''^{-1}(\widehat{\theta}, z) &= V_N''^{-1}(\tilde{\theta}(z_0)) + \upsilon(\upsilon^2 + \upsilon\lambda(z_0))O_p(N^{-1/2}) \end{aligned} \qquad (17\text{-}86)$$

Collecting (17-79), (17-80), and (17-86) finally gives

$$b_\theta(z) = (\upsilon^2 + (\upsilon + \mu)\lambda(z_0))O_p(N^{-1}) \qquad (17\text{-}87)$$

The term $\upsilon\lambda(z_0)O_p(N^{-1})$ in $b_\theta(z)$ stems from the product of the second term in (17-78) with the term $\upsilon^2\lambda(z_0)O_p(N^{-1/2})$ in (17-86). As the latter depends only on z_0 (see (17-84)), the expected value of this product is zero (by assumptions, z_0 and n_z are independent). □

Appendix 17.E: Proof of the Asymptotic Distribution of the Markov Estimates without Model Errors

Expression (15-18) will be elaborated for the Markov estimator assuming that no model errors are present $(e(\theta_0, z_0) = 0)$. Using $\varepsilon(\theta, z) = \varepsilon(\theta, z_0) + \Delta(\theta, n_z)$, with $\Delta(\theta, n_z) = \Lambda(\theta)M_1(\theta)n_z$, and $\varepsilon(\theta_0, z_0) = 0$ (see (17-8)), we find

$$V_N'^T(\theta_0, z) = \frac{1}{N}\left(\frac{\partial\varepsilon(\theta, z_0)}{\partial\theta_0}\right)^T \Delta(\theta_0, n_z) + v_N'^T(\theta_0, n_z) \tag{17-88}$$

where $v_N(\theta, n_z) = \frac{1}{2N}\Delta^T(\theta, n_z)\Delta(\theta, n_z)$. Because $\mathrm{Cov}(\Delta(\theta, n_z)) = I_{rN}$ (see (17-8)), $\varepsilon(\theta_0, z_0) = 0$ and z_0, n_z are mutually independent (Assumption 15.1), $Q_N(\theta_0)$ (15-18) becomes

$$\begin{aligned} Q_N(\theta_0) &= N\mathscr{E}\{V_N'^T(\theta_0, z)V_N'(\theta_0, z)\} \\ &= V_N''(\theta_0) + N\mathscr{E}\{v_N'^T(\theta_0, n_z)v_N'(\theta_0, n_z)\} \\ &\quad + 2\mathrm{herm}(\mathscr{E}\{\left(\frac{\partial\varepsilon(\theta, z_0)}{\partial\theta_0}\right)^T\}\mathscr{E}\{\Delta(\theta_0, n_z)v_N'(\theta_0, n_z)\}) \end{aligned} \tag{17-89}$$

Putting (17-89) into (15-18) gives (17-32). □

Appendix 17.F: Proof of the Asymptotic Efficiency of the Markov Estimates (Theorem 17.4)

The second term in the expression of $q_N(\theta_0)$ (see (17-32)) is a function of the third-order moments of n_z and, hence, is zero for Gaussian errors n_z. The first term is positive semidefinite and is zero if and only if $v_N(\theta, n_z)$ is independent of θ for any n_z. From (17-68) and Assumption 17.1, it follows that we can write n_z as $n_z = U_1\Sigma_1^{1/2}\varepsilon_z$ with $\varepsilon_z \in N_{r_c}(0, I_{r_c})$ (see Exercise 14.9). Hence, $M_1(\theta)n_z = E(\theta)\varepsilon_z$, with $E(\theta) = M_1(\theta)U_1\Sigma_1^{1/2}$, and $C_e(\theta) = E(\theta)E^T(\theta)$, so that

$$v_N(\theta, n_z) = \frac{1}{2N}\varepsilon_z^T E^T(\theta)(E(\theta)E^T(\theta))^{-1}E(\theta)\varepsilon_z \tag{17-90}$$

The matrix $E(\theta) \in \mathbb{R}^{rN\times r_c}$ has full rank rN $(rN \le r_c \le sN)$ because $C_e(\theta)$ has full rank rN. If $r_c = rN$, then $(E(\theta)E^T(\theta))^{-1} = E^{-T}(\theta)E^{-1}(\theta)$ and $v_N(\theta, n_z) = \frac{1}{2N}\varepsilon_z^T\varepsilon_z$. This concludes the proof because ε_z is independent of θ. □

Appendix 17.G: Proof of the Convergence Rate of the Residuals (Lemma 17.8)

Applying the mean value theorem to $\varepsilon_{[i]}(\hat{\theta}(z), z)$ at the points $\hat{\theta}(z)$, $\tilde{\theta}(z_0)$ gives

$$\varepsilon_{[i]}(\hat{\theta}(z), z) = \varepsilon_{[i]}(\tilde{\theta}(z_0), z) + \frac{\partial\varepsilon_{[i]}(\theta, z)}{\partial\theta_1}(\hat{\theta}(z) - \tilde{\theta}(z_0)) \tag{17-91}$$

where $\theta_1 = t_1\hat{\theta}(z) + (1-t_1)\tilde{\theta}(z_0)$ with $t_1 \in [0, 1]$. Under Assumptions 17.5, 17.6, and 15.1 $(P = 2)$ $\varepsilon(\theta, z)$ and $\partial\varepsilon(\theta, z)/\partial\theta_{[j]}$ are both mixing of order 2 in $\mathbb{\Theta}_{\rm r}$ (Corollary 14.7), so that $\text{var}(\varepsilon_{[i]}(\theta, z)) = O(N^0)$ and $\text{Covar}(\partial\varepsilon_{[i]}(\theta, z)/\partial\theta) = O(N^0)$ uniformly in $\mathbb{\Theta}_{\rm r}$. Hence, $\varepsilon_{[i]}(\tilde{\theta}(z_0), z)$ is an $O_{\rm m.s.}(N^0)$ $(O_{\rm p}(N^0))$. From Theorem 15.21, it follows that $\hat{\theta}(z)$ converges in prob. to $\tilde{\theta}(z_0)$ at the rate $O_{\rm p}(N^{-1/2})$. The same is true for θ_1, so that $\partial\varepsilon_{[i]}(\theta, z)/\partial\theta_1 = O_{\rm p}(N^0)$ (Lemma 15.33). We conclude that $\varepsilon_{[i]}(\hat{\theta}(z), z) - \varepsilon_{[i]}(\tilde{\theta}(z_0), z) = O_{\rm p}(N^{-1/2})$, which proves the lemma. □

Appendix 17.H: Properties of the Projection Matrix in Lemma 17.9

Part one of this appendix studies the stochastic properties of the projection matrix $Q_\varepsilon(z_0)$ in Eq. (17-38) for random z_0. Part two calculates the covariance matrix of $Q_\varepsilon(z_0)\delta_\varepsilon(z)$.

(i) Under Assumptions 15.1 $(P = 4)$ and 17.6(a), $\partial\varepsilon(\theta, z_0)/\partial\theta_{[j]}$ is mixing of order 4. Therefore, $\partial\varepsilon_{[i]}(\theta, z_0)/\partial\theta_{[j]} = O_{\rm p}(N^0)$ and

$$\frac{1}{N}\left(\frac{\partial\varepsilon(\theta, z_0)}{\partial\theta}\right)^T\left(\frac{\partial\varepsilon(\theta, z_0)}{\partial\theta}\right) = \frac{1}{N}\mathcal{E}\left\{\left(\frac{\partial\varepsilon(\theta, z_0)}{\partial\theta}\right)^T\left(\frac{\partial\varepsilon(\theta, z_0)}{\partial\theta}\right)\right\} + O_{\rm p}(N^{-1/2}) \quad (17\text{-}92)$$

uniformly in $\mathbb{\Theta}_{\rm r}$ as $N \to \infty$ (proof of (17-92) is similar to Lemma 15.3). Because Assumption 15.18 is valid for any $\mu \to 0$, it guarantees that $\mathcal{E}\{\varepsilon'^T(\tilde{\theta}(z_0), z_0)\varepsilon'(\tilde{\theta}(z_0), z_0)\} = O(N)$. Putting these results in (17-38) finally gives $Q_{\varepsilon[i,j]}(z_0) = I_{rN[i,j]} + O_{\rm p}(N^{-1})$. As z_0 is uniformly bounded, there exists an N_0 s.t. for any $N \geq N_0$: $\mathcal{E}\{Q_{\varepsilon[i,j]}(z_0)\} = I_{rN[i,j]} + O(N^{-1})$.

(ii) Using $\mathcal{E}\{\delta_\varepsilon(z)\} = 0$, $\text{Cov}(\delta_\varepsilon(z)) = I_{rN}$, and the fact that n_z and z_0 are stochastically independent gives $\text{Cov}(Q_\varepsilon(z_0)\delta_\varepsilon(z)) = \mathcal{E}\{Q_\varepsilon(z_0)\} = O(N^0)$, where the last equality follows from part one. □

Appendix 17.I: Proof of the Improved Convergence Rate of the Residuals (Lemma 17.9)

Taylor series expansion of $\varepsilon_{[i]}(\hat{\theta}(z), z)$ at $\tilde{\theta}(z_0)$, z_0 gives

$$\begin{aligned}\varepsilon_{[i]}(\hat{\theta}(z), z) &= \varepsilon_{[i]}(\tilde{\theta}(z_0), z_0) + \frac{\partial\varepsilon_{[i]}(\theta, z_0)}{\partial\tilde{\theta}(z_0)}\Delta\theta + \frac{\partial\varepsilon_{[i]}(\tilde{\theta}(z_0), z)}{\partial z_0}n_z \\ &\quad + \frac{1}{2}\Delta\theta^T\frac{\partial^2\varepsilon_{[i]}(\theta, z_1)}{\partial\theta_1{}^2}\Delta\theta + \Delta\theta^T\frac{\partial^2\varepsilon_{[i]}(\theta, z_1)}{\partial\theta_1\partial z_1}n_z\end{aligned} \quad (17\text{-}93)$$

with $\Delta\theta = \hat{\theta}(z) - \tilde{\theta}(z_0)$, $\theta_1 = t_1\hat{\theta}(z) + (1-t_1)\tilde{\theta}(z_0)$, $z_1 = t_1 z + (1-t_1)z_0$, and $t_1 \in [0, 1]$. We analyze each term in the right-hand side of (17-93). Under Assumptions 15.1 $(P = 2)$ and 17.5, the first term of (17-93) is mixing of order two (Corollary 14.7) and, hence, it behaves as $\upsilon^{-1}\mu O_{\rm p}(N^0)$. Using (17-31), the sum of the second and third terms of (17-93) becomes

$$(Q_{\varepsilon}(z_0)\delta_{\varepsilon}(z))_{[i]} + (\upsilon + \mu + \mu\lambda(z_0))O_{\mathrm{p}}(N^{-1/2}) \tag{17-94}$$

with $\lambda(z_0) = 1$ for random z_0 and $\lambda(z_0) = 0$ for deterministic z_0. Because $\mathcal{E}\{(Q_{\varepsilon}(z_0)\delta_{\varepsilon}(z))_{[i]}\} = 0$ and $\mathrm{var}((Q_{\varepsilon}(z_0)\delta_{\varepsilon}(z))_{[i]}) = O(N^0)$ (see Appendix 17.H), we have $(Q_{\varepsilon}(z_0)\delta_{\varepsilon}(z))_{[i]} = O_{\mathrm{p}}(N^0)$. Using $\Delta\theta = (\upsilon + \mu\lambda(z_0))O_{\mathrm{p}}(N^{-1/2})$ (Theorem 17.2) and Assumption 17.6, it can be seen that the last two terms of (17-93) are, respectively, a $\upsilon^{-1}(\upsilon + \mu\lambda(z_0))^2 O_{\mathrm{p}}(N^{-1})$ and $(\upsilon + \mu\lambda(z_0))O_{\mathrm{p}}(N^{-1/2})$. Putting all these results in (17-93) proves the lemma. □

Appendix 17.J: Proof of the Properties of the Sample Correlation of the Residuals (Theorem 17.10)

The proof consists of three parts: part one shows the weak convergence and the convergence rate of $\hat{R}_{\varepsilon\varepsilon}(k)$, part two proves the asymptotic normality of $\hat{R}_{\varepsilon\varepsilon}(k)$, and part three gives an asymptotic expression $(N \to \infty)$ for the variance of the truncated sample correlation $\hat{R}_{\varepsilon\varepsilon}(k)$. In the proof, it is essential that $N - k = O(N)$. Hence, the results are valid for constant values of k or constant fractions $k = \alpha N$, with α independent of N.

(i) Define the following functions:

$$f_N(\theta, z, k) = \frac{1}{rN-k}\sum_{i=1}^{rN-k} \varepsilon_{[i]}(\theta, z)\varepsilon_{[i+k]}(\theta, z)$$

$$f_N(\theta, k) = \mathcal{E}\{f_N(\theta, z, k)\}$$

$$f_N{}'(\theta, z, k) = \frac{1}{rN-k}\sum_{i=1}^{rN-k}\left(\varepsilon_{[i+k]}(\theta, z)\frac{\partial \varepsilon_{[i]}(\theta, z)}{\partial\theta} + \varepsilon_{[i]}(\theta, z)\frac{\partial\varepsilon_{[i+k]}(\theta, z)}{\partial\theta}\right)$$

with $f_N(\hat{\theta}(z), z, k) = \hat{R}_{\varepsilon\varepsilon}(k)$. Under Assumptions 17.5, 17.6, and 15.1 $(P = 4)$ $\varepsilon(\theta, z)$ and $\partial\varepsilon(\theta, z)/\partial\theta_{[j]}$ are jointly mixing of order 4 in Θ_{r}. Hence, according to the weak law of large numbers (14-68), $f_N(\theta, z, k)$ and $f_N{}'(\theta, z, k)$ converge in prob. to their expected values at the rate $O_{\mathrm{p}}(N^{-1/2})$, and $\|f_N{}'(\theta, z, k)\|_2 = O_{\mathrm{p}}(N^0)$. From Theorem 15.21, it follows that $\hat{\theta}(z)$ converges in prob. to $\tilde{\theta}(z_0)$ at the rate $O_{\mathrm{p}}(N^{-1/2})$. All the conditions of Lemma 15.34 are fulfilled so that $f_N(\hat{\theta}(z), z, k)$ converges in prob. to $f_N(\tilde{\theta}(z_0), k)$ at the rate $O_{\mathrm{p}}(N^{-1/2})$. Using $\varepsilon_{[i]}(\tilde{\theta}(z_0), z) = \varepsilon_{[i]}(\tilde{\theta}(z_0), z_0) + \delta_{\varepsilon[i]}(z)$ in $f_N(\tilde{\theta}(z_0), k)$ gives (17-40).

(ii) Now we show that $\hat{R}_{\varepsilon\varepsilon}(k)$ is asymptotically normally distributed. Therefore, it is sufficient to verify that all the conditions of Lemma 15.38 are fulfilled. Under Assumptions 15.1 $(P = \infty)$ and 17.5, $\varepsilon(\theta, z)$ is mixing of order infinity. Hence, according the central limit theorem ((14-74), version 4), $f_N(\theta, z, k)$ is asymptotically normally distributed (condition 1 of Lemma 15.38). Condition 2 of Lemma 15.38 is already satisfied (see part one of the proof). Under Assumption 17.6(b) we have $\|f_N{}''(\theta, z, k)\|_2 = O_{\mathrm{p}}(N^0)$ uniformly in Θ_{r} (condition 3 of Lemma 15.38). The proof is similar to that of $f_N{}'(\theta, z, k)$ in part one. The assumptions of Theorem 15.29 are satisfied so that $\hat{\theta}(z)$ is asymptotically normally distributed (condition 4 of Lemma 15.38).

(iii) In the absence of model errors (Assumptions 15.9 and 15.10), $\varepsilon(\tilde{\theta}(z_0), z_0) = 0$ and

$$\varepsilon(\tilde{\theta}(z_0), z) = \Delta(\tilde{\theta}(z_0), n_z) = \delta_\varepsilon(z) \tag{17-95}$$

where $\delta_\varepsilon(z)$ is defined in Lemma 17.9. Using (17-95), (17-91) becomes

$$\varepsilon_{[i]}(\hat{\theta}(z), z) = \delta_{\varepsilon[i]}(z) + O_p(N^{-1/2}) \tag{17-96}$$

where $\delta_{\varepsilon[i]}(z)$ is an $O_p(N^0)$. The asymptotic variance ($N \to \infty$) of the truncated sample correlation $\underline{\hat{R}_{\varepsilon\varepsilon}}(k)$ is, hence, given by

$$\text{var}(\frac{1}{rN-k}\sum_{i=1}^{rN-k} \delta_{\varepsilon[i]}(z)\delta_{\varepsilon[i+k]}(z))$$

Under Assumption 17.1, $\delta_{\varepsilon[i]}(z)$ is normally distributed $N_{rN}(0, I_{rN})$, so that the cumulants of $\delta_{\varepsilon[i]}(z)$ of order 3 and higher are zero (see Example 14.2). Using this result together with that of Example 14.36 and $\text{var}(x) = \text{cum}(x, x)$ (see Section 14.1), we find

$$\text{var}(\frac{1}{rN-k}\sum_{i=1}^{rN-k} \delta_{\varepsilon[i]}(z)\delta_{\varepsilon[i+k]}(z)) = \frac{1}{rN-k}(1 + \delta(k)) \qquad \square$$

Appendix 17.K: Proof of the Convergence Rate of the Minimum of the Cost Function (Lemma 17.11)

Taylor series expansion with remainder of $V_N(\hat{\theta}(z), z) = \frac{1}{N}V_{\text{Markov}}(\hat{\theta}(z), z)$ at $\tilde{\theta}(z_0)$, z_0 gives

$$\begin{aligned} V_N(\hat{\theta}(z), z) = {} & V_N(\tilde{\theta}(z_0), z_0) + \frac{\partial V_N(\theta, z_0)}{\partial \tilde{\theta}(z_0)}\Delta\theta + \frac{\partial V_N(\tilde{\theta}(z_0), z)}{\partial z_0}n_z \\ & + \frac{1}{2}\Delta\theta^T\frac{\partial^2 V_N(\theta, z_0)}{\partial \tilde{\theta}(z_0)^2}\Delta\theta + \frac{1}{2}n_z^T\frac{\partial^2 V_N(\tilde{\theta}(z_0), z)}{\partial z_0^2}n_z + \Delta\theta^T\frac{\partial^2 V_N(\theta, z)}{\partial \tilde{\theta}(z_0)\partial z_0}n_z \\ & + R_1 + R_2 + R_3 + R_4 \end{aligned} \tag{17-97}$$

where

$$R_1 = \frac{1}{6}\left(\sum_{i=1}^{n_\theta}\Delta\theta_{[i]}\frac{\partial}{\partial \theta_{1[i]}}\right)^3 V_N(\theta, z_1)$$

$$R_2 = \frac{1}{6}\left(\sum_{i=1}^{sN}n_{z[i]}\frac{\partial}{\partial z_{1[i]}}\right)^3 V_N(\theta_1, z)$$

$$R_3 = \frac{1}{2}\left(\sum\nolimits_{i=1}^{n_\theta} \Delta\theta_{[i]}\frac{\partial}{\partial\theta_{1[i]}}\right)\left(\sum\nolimits_{i=1}^{sN} n_{z[i]}\frac{\partial}{\partial z_{1[i]}}\right)^2 V_N(\theta, z)$$

$$R_4 = \frac{1}{2}\left(\sum\nolimits_{i=1}^{n_\theta} \Delta\theta_{[i]}\frac{\partial}{\partial\theta_{1[i]}}\right)^2\left(\sum\nolimits_{i=1}^{sN} n_{z[i]}\frac{\partial}{\partial z_{1[i]}}\right) V_N(\theta, z)$$

and $\theta_1 = t_1\hat{\theta}(z) + (1-t_1)\tilde{\theta}(z_0)$, $z_1 = t_1 z + (1-t_1)z_0$ with $t_1 \in [0, 1]$. Each term in the right-hand side of (17-97) will be studied.

(i) Under Assumptions 15.1($P = 4$) and 15.2(a), the first term $V_N(\tilde{\theta}(z_0), z_0)$ is mixing of order two so that $\mathrm{var}(V_N(\tilde{\theta}(z_0), z_0)) = O(N^{-1})$ and, hence,

$$V_N(\tilde{\theta}(z_0), z_0) = \mu^2\upsilon^{-2}O_p(N^0) \tag{17-98}$$

(ii) Using Theorem 17.2, the sum of the second and the third term can be written as

$$\frac{1}{N}\varepsilon^T(\tilde{\theta}(z_0), z_0)Q_\varepsilon(z_0)\delta_\varepsilon(z) + V_N'(\tilde{\theta}(z_0), z_0)(\mu\upsilon + \upsilon^2 + \mu\lambda(z_0))O_p(N^{-1/2}) \tag{17-99}$$

For deterministic z_0 we have $V_N'(\tilde{\theta}(z_0), z_0) = V_N'(\tilde{\theta}(z_0)) = 0$ (see (17-26)). For random z_0, $V_N'(\tilde{\theta}(z_0), z_0)$ converges in prob. to $V_N'(\tilde{\theta}(z_0))$ at the rate $O_p(N^{-1/2})$ (Lemma 15.17). $V_N'(\tilde{\theta}(z_0)) = 0$ by definition of $\tilde{\theta}(z_0)$, so that $V_N'(\tilde{\theta}(z_0), z_0) = \mu\upsilon^{-2}O_p(N^{-1/2})$. Putting these results in (17-99) gives

$$\frac{1}{N}\varepsilon^T(\tilde{\theta}(z_0), z_0)Q_\varepsilon(z_0)\delta_\varepsilon(z) + \frac{1}{N}\mu^2\upsilon^{-2}\lambda(z_0)O_p(N^0) \tag{17-100}$$

with $\lambda(z_0) = 1$ for random z_0 and $\lambda(z_0) = 0$ for deterministic z_0. Note also that $\varepsilon^T(\tilde{\theta}(z_0), z_0)Q_\varepsilon(z_0) = \varepsilon^T(\tilde{\theta}(z_0), z_0)$ for deterministic z_0.

(iii) The second-order derivative in the fourth term of (17-97) can be written as

$$\frac{\partial^2 V_N(\theta, z_0)}{\partial\tilde{\theta}(z_0)^2} = \frac{1}{N}\left(\frac{\partial\varepsilon(\theta, z_0)}{\partial\tilde{\theta}(z_0)}\right)^T\left(\frac{\partial\varepsilon(\theta, z_0)}{\partial\tilde{\theta}(z_0)}\right) + \frac{1}{N}\sum_{k=1}^{rN}\varepsilon_{[k]}(\tilde{\theta}(z_0), z_0)\frac{\partial^2\varepsilon_{[k]}(\theta, z_0)}{\partial\tilde{\theta}(z_0)^2} \tag{17-101}$$

Under Assumptions 17.5, 17.6, and 15.1 ($P = 4$), both terms in the right-hand side of (17-101) converge in prob. to their expected value (proof: similar to Lemma 15.17). Assumption 15.18, which is by assumption valid for any $\mu \to 0$, guarantees that both terms behave as an $O_p(N^0)$. We conclude that the first and second terms are, respectively, an $\upsilon^{-2}O_p(N^0)$ and $\mu\upsilon^{-2}O_p(N^0)$.

(iv) The sixth term of (17-97) can be written as

$$\begin{aligned}
&\Delta\theta^T(g(z) + f_N(\tilde{\theta}(z_0), z)) \\
&g(z) = \frac{1}{N}\varepsilon'^T(\tilde{\theta}(z_0), z_0)\delta_\varepsilon(z) \\
&f_N(\tilde{\theta}(z_0), z) = \frac{1}{N}\sum\nolimits_{k=1}^{rN}\varepsilon_{[k]}(\tilde{\theta}(z_0), z_0)\left(\frac{\partial\delta_{\varepsilon[k]}(z)}{\partial\tilde{\theta}(z_0)}\right)^T
\end{aligned} \qquad (17\text{-}102)$$

where $\delta_\varepsilon(z)$ is defined in Lemma 17.9. Because $\mathcal{E}\{g(z)\} = 0$ and

$$\mathrm{Cov}(g(z)) = \frac{1}{N^2}\mathcal{E}\left\{\left(\frac{\partial\varepsilon(\theta, z_0)}{\partial\tilde{\theta}(z_0)}\right)^T\left(\frac{\partial\varepsilon(\theta, z_0)}{\partial\tilde{\theta}(z_0)}\right)\right\} = O(N^{-1})$$

we conclude that $g(z) = \upsilon^{-1}O_{\mathrm{p}}(N^{-1/2})$. Under Assumptions 17.5, 17.6, and 15.1($P = 4$), $\varepsilon(\theta, z_0)$ and $\partial\delta_\varepsilon(z)/\partial\theta_{[i]}$ are jointly mixing of order 4, so that $f_N(\theta, z)$ converges in prob. to its expected value $f_N(\theta)$ at the rate $O_{\mathrm{p}}(N^{-1/2})$, uniformly in Θ_{r} (proof: similar to Lemma 15.17). Because $f_N(\tilde{\theta}(z_0)) = 0$, it follows that $f_N(\tilde{\theta}(z_0), z)$ is a $\mu\upsilon^{-1}O_{\mathrm{p}}(N^{-1/2})$. Combining these results with Theorem 17.2 gives the following expression for the sum of the fourth, fifth, and sixth terms in (17-97):

$$\frac{1}{2N}\delta_\varepsilon^T(z)Q_\varepsilon(z_0)\delta_\varepsilon(z) + \frac{1}{N}(\mu + \upsilon + \mu\upsilon^{-2}(\mu + \upsilon)\lambda(z_0))O_{\mathrm{p}}(N^0) \qquad (17\text{-}103)$$

(v) The term R_1 is bounded above by

$$|R_1| \le \sum_{i,j,k=1}^{n_\theta}|\Delta\theta_{[i]}\Delta\theta_{[j]}\Delta\theta_{[k]}|\frac{(\|z_1\|_2^2 + 1)}{N}\left\|\frac{\partial^3 W_N(\theta)}{\partial\theta_{1[i]}\partial\theta_{1[j]}\partial\theta_{1[k]}}\right\|_2 \qquad (17\text{-}104)$$

where, by Assumption 15.20 ($\theta_1 \in \Theta_{\mathrm{r}}$), the 2-norm of the third-order derivative of $W_N(\theta)$ w.r.t. θ is an $\upsilon^{-2}O_{\mathrm{p}}(N^0)$. Under Assumption 15.1($P = 4$), $\|z\|_2^2/N$ and $\|z_0\|_2^2/N$ converge both in prob. to their expected values, which are both an $O(N^0)$ (weak law of large numbers, see Section 14.9). Hence, $\|z\|_2^2/N$ and $\|z_0\|_2^2/N$ are both an $O_{\mathrm{p}}(N^0)$. This is also true for $\|z_1\|_2^2/N$, because $z_1 = t_1 z + (1 - t_1)z_0$ with $t_1 \in [0, 1]$. Combining these results with Theorem 17.2 gives

$$|R_1| \le \frac{1}{N}\upsilon^{-2}(\upsilon + \mu\lambda(z_0))^3 O_{\mathrm{p}}(N^{-1/2}) \qquad (17\text{-}105)$$

(vi) Because $V_N(\theta, z)$ is a quadratic function of z, it follows directly that

$$R_2 = 0 \qquad (17\text{-}106)$$

(vii) The term R_3 can be written as

$$R_3 = v_N'(\theta_1, n_z)\Delta\theta \qquad (17\text{-}107)$$

with $v_N(\theta, n_z) = \frac{1}{2N}\Delta^T(\theta, n_z)\Delta(\theta, n_z)$ and $\Delta(\theta, n_z) = \Lambda(\theta)M_1(\theta)n_z$. We will verify that all the conditions of Lemma 15.34 are satisfied for $v_N'(\theta_1, n_z)$. Under Assumptions 15.1($P = 4$) and 15.16, $v_N'(\theta, n_z)$ converges in prob. to its expected value $\mathcal{E}\{v_N'(\theta, n_z)\} = \partial(rN/2)/\partial\theta = 0$ at the rate $O_p(N^{-1/2})$ uniformly in Θ_r (Lemma 15.17). $\hat{\theta}(z)$ converges in prob. to $\tilde{\theta}(z_0)$ at the rate $O_p(N^{-1/2})$. This is also true for θ_1 because $\theta_1 = t_1\hat{\theta}(z) + (1 - t_1)\tilde{\theta}(z_0)$ with $t_1 \in [0, 1]$. Under Assumptions 17.1($P = 4$) and 15.16, we have

$$\|v_N''(\theta, n_z)\|_2 \leq \frac{\|n_z\|_2^2}{N}\left\|\frac{\partial^2 W_N(\theta)}{\partial\theta_{[i]}\partial\theta_{[j]}\partial\theta_{[k]}}\right\|_2 = O_p(N^0)$$

uniformly in Θ_r. We conclude from Lemma 15.34 that $v_N'(\theta_1, n_z) = \upsilon^0 O_p(N^{-1/2})$. Combining this result with Theorem 17.2 gives

$$R_3 = \frac{1}{N}(\upsilon + \mu\lambda(z_0))O_p(N^0) \tag{17-108}$$

(viii) The term R_4 can be written as

$$R_4 = \sum_{i,j=1}^{n_\theta} \Delta\theta_{[i]}\Delta\theta_{[j]}\frac{\partial^2 F(\theta_1, z)}{\partial\theta_{1[i]}\partial\theta_{1[j]}} \tag{17-109}$$

with $F(\theta, z) = \frac{1}{2N}\Delta^T(\theta, n_z)\varepsilon(\theta, z_1)$. z_1 is mixing of order four because this is the case for z and z_0. Therefore, under Assumptions 15.1 ($P = 4$) and 15.16, $F''(\theta, z)$ converges uniformly in prob. to its expected value $F''(\theta)$, which is an $O(N^0)$ (proof similar to Lemma 15.17). Because θ_1 converges in prob. to $\tilde{\theta}(z_0)$ it follows that $F''(\theta_1, z)$ converges in prob. to $F''(\tilde{\theta}(z_0)) = O(N^0)$ (Lemma 15.31), so that $F''(\theta_1, z) = \upsilon^{-1}O_p(N^0)$. Combining this result with Theorem 17.2 gives

$$R_4 = \frac{1}{N}\upsilon^{-1}(\upsilon + \mu\lambda(z_0))^2 O_p(N^0) \tag{17-110}$$

(ix) Putting (17-98), (17-100), (17-103), (17-105), (17-106), (17-108), and (17-110) into (17-97), taking into account that $V_N(\hat{\theta}(z), z) = \frac{1}{N}V_{\text{Markov}}(\hat{\theta}(z), z)$, proves (17-44).

Appendix 17.L: Proof of the Properties of the Cost Function (Theorem 17.12)

To prove the asymptotic normality of $V_N(\hat{\theta}(z), z)$, we show that all conditions of Lemma 15.38 are satisfied. In fact, it is sufficient to show the asymptotic normality of $V_N(\theta, z)$ because all the other conditions are satisfied under the assumptions of Lemma 17.11. $\varepsilon_{[k]}^2(\theta, z)$ is mixing of order infinity under Assumptions 15.1 and 17.5 and, therefore, the asymptotic normality of $V_N(\theta, z)$ follows directly from version 4 of the central limit theorem (see Section 14.10).

Using $\delta_\varepsilon^T(z)Q_\varepsilon(z_0)\delta_\varepsilon(z) = \text{trace}(Q_\varepsilon(z_0)\delta_\varepsilon(z)\delta_\varepsilon^T(z))$ and Assumption 15.1($P = 2$), we find

$$\mathcal{E}\{\, L_{\text{Markov}}(\tilde{\theta}(z_0), z_0)\,\} = \mathcal{E}\{\, V_{\text{Markov}}(\tilde{\theta}(z_0), z_0)\,\} + \mathcal{E}\{\,\text{trace}(Q_\varepsilon(z_0))\,\}$$

$Q_\varepsilon(z_0)$ is an idempotent matrix of rank $rN - n_\theta$ so that $\text{trace}(Q_\varepsilon(z_0)) = rN - n_\theta$. Because $\delta_\varepsilon(z) \in N^{rN}(0, I_{rN})$ under Assumption 17.1, we have $\delta_\varepsilon^T(z)Q_\varepsilon(z_0)\delta_\varepsilon(z) \in \chi^2(rN - n_\theta)$ (see Exercise 14.10), so that $\text{var}(\delta_\varepsilon^T(z)Q_\varepsilon(z_0)\delta_\varepsilon(z)) = 2(rN - n_\theta)$ (Stuart and Ord, 1987). Formula (17-46) follows directly from this result.

For deterministic z_0, we have $\text{var}(V_{\text{Markov}}(\tilde{\theta}(z_0), z_0)) = 0$ and $\varepsilon^T(\tilde{\theta}(z_0), z_0)Q_\varepsilon(z_0) = \varepsilon^T(\tilde{\theta}(z_0), z_0)$, so that (17-46) reduces to (17-47). □

Appendix 17.M: Model Selection Criteria

For model (17-1), the AIC and MDL criteria have the form

$$-\ln f_z(z, x, \theta) + g(k, m) \tag{17-111}$$

with $f_z(z, x, \theta)$ the likelihood function, x the number of unknown, independent variables in z_{p} (see Section 17.3), k the number of free parameters in the model to get the estimates $\hat{\theta}(z)$ and $\hat{x}$, m the number of independent noisy measurements, $g(k, m) = k$ for AIC, and $g(k, m) = 0.5k\ln m$ for MDL. The number of free parameters equals the total number of identifiable parameters $\dim(\hat{\theta}(z)) + \dim(\hat{x})$, so that $k = n_\theta + \dim(\hat{x})$. The number of independent noisy measurements equals the number of measurements $r_c = \text{rank}(C_{n_z})$ lying in the regular space of C_{n_z}. Taking into account that $\dim(\hat{x}) = r_c - rN$ is constant over the model set $\mathbb{M}$, (17-111) reduces for Gaussian-distributed errors n_z to (17-50) and (17-51). □

18

Identification of Invariants of (Over)Parameterized Models

Abstract: This chapter deals with the identification of invariants of (over)parameterized models. First, it is shown that the generalized Cramér-Rao lower bound on the estimate of invariants of (over)parameterized models is independent of the particular (over)parameterization chosen and equals that of the identifiable form. This result is useful for (asymptotically) efficient estimators. Next, it is shown that a certain class of estimates of invariants of (over)parameterized models are with probability one, independent of the particular (over)parameterization chosen. The result is nonasymptotic and the estimators considered minimize a cost function that is invariant with respect to the same parameter transformation as the overparameterized model.

18.1 INTRODUCTION

In many identification problems, one is faced with the choice of the parameterization of a model out of a large number of possibilities (see, for example, Guidorzi, 1975; Van Overbeek and Ljung, 1982). Some of these representations contain a redundant number of parameters and lead to the so-called overparameterized models. Such models result in singular Fisher information matrices (Shapiro, 1986). This situation is often encountered in practical parameter estimation problems.

Example 18.1 (Rational Transfer Function Model for SISO Systems): Consider the identification of the numerator and denominator coefficients of a rational transfer function model of a single input, single output (SISO) continuous-time system

$$G(s, \theta) = \frac{B(s, \theta)}{A(s, \theta)} = \frac{\sum_{r=0}^{n_b} b_r s^r}{\sum_{r=0}^{n_a} a_r s^r} \qquad (18\text{-}1)$$

where $\theta^T = (a_0, a_1, \ldots, a_{n_a}, b_0, b_1, \ldots, b_{n_b})$. Transfer function model (18-1) is overparameterized because θ is unidentifiable: $G(s, \lambda\theta) = G(s, \theta)$ for any $\lambda \in \mathbb{R}_0$. Assuming that the true model order is (n_a, n_b), the dimension of the null space of the corresponding Fisher

System Identification: A Frequency Domain Approach. Rik Pintelon and Johan Schoukens
Copyright © 2001 Institute of Electrical and Electronics Engineers. ISBN: 0-780-36000-1

information matrix (14-85) equals 1. Identifiable parameterizations ψ are obtained by fixing one coefficient of the numerator or denominator, for example, $a_{n_a} = 1$ for a monic denominator polynomial. Invariants of model (18-1) are, for example, the poles and zeros of the transfer function or the value of the transfer function itself. □

Example 18.2 (State Space Models for Multivariable Systems): Consider the identification of a proper multivariable discrete-time system parameterized by its state space representation (A, B, C, D)

$$G(z^{-1}, \theta) = z^{-1}C(I_n - z^{-1}A)^{-1}B + D \tag{18-2}$$

with n the order of the state space model, $A \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{n \times n_u}$, $C \in \mathbb{R}^{n_y \times n}$, and $D \in \mathbb{R}^{n_y \times n_u}$. The overparameterized model parameter θ is related to the (A, B, C, D) matrices as

$$\theta^T = [\text{vec}^T(A)\ \text{vec}^T(B)\ \text{vec}^T(C)\ \text{vec}^T(D)] \tag{18-3}$$

where vec(.) transforms a matrix into a column vector by stacking the columns of the matrix on top of each other. Transfer function (18-2) is invariant w.r.t. a regular transformation $T \in \mathbb{R}^{n \times n}$: replacing (A, B, C, D) by $(TAT^{-1}, TB, CT^{-1}, D)$ in (18-2), with $\det(T) \neq 0$, leaves $G(z^{-1}, \theta)$ unchanged. Hence, n^2 dependences exist between the entries of θ. Assuming that the true model order is n, the dimension of the null space of the Fisher information matrix equals n^2. Identifiable parameterizations ψ are obtained by constraining the matrices (A, B, C) (Van Overbeek and Ljung, 1982). Invariants of model (18-2) are, for example, the eigenvalues of A. □

First, we define the considered (over)parameterized models and their invariants. Next, the Cramér-Rao lower bound on the estimates of the invariants is analyzed. Further, the finite sample behavior of a certain class of estimates of invariants is studied. Finally, the chapter concludes with a numerical example of identification methods whose estimates are, respectively, dependent on and independent of the particular (over)parameterization chosen. Although throughout this chapter the theory is mainly illustrated on the identification of continuous-time single input, single output systems, it is also valid for discrete-time, multivariable, and nonlinear systems.

18.2 (OVER)PARAMETERIZED MODELS AND THEIR INVARIANTS

The system model is described as a general vector function $M(\theta, z_0)$ of the true signal $z_0 \in \mathbb{R}^N$ and the overparameterized model parameters $\theta \in \mathbb{R}^{n_\theta}$ with $n_\theta < N$. The model $M(\theta, z_0)$ is defined for every $\theta \in \mathbb{D}_\theta$, a subset of $\mathbb{R}^{n_\theta}$. The complement of $\mathbb{D}_\theta$ in $\mathbb{R}^{n_\theta}$ is $\mathbb{S}_\theta$, the set of singular points. $\mathbb{S}_\theta = \mathbb{R}^{n_\theta} \setminus \mathbb{D}_\theta$ has topological dimensions less than n_θ $(\dim(\mathbb{S}_\theta) < n_\theta)$. If no system model errors are present, then $M(\theta_0, z_0) = 0$ with θ_0 the true model parameters. Note that for overparameterized models, θ_0 is not a single point but a subspace of $\mathbb{R}^{n_\theta}$.

In practice, only noisy observations $z = z_0 + n_z$ of the true signal z_0 are available. If a parametric noise model for the observation noise n_z is identified, then the system model $M(\theta, z_0)$ must be extended with the noise model. To simplify the notations, the discussion is, without any loss of generality, limited to the case without a parametric noise model.

Assumption 18.3 (Invariance Model): The overparameterized model $M(\theta, z_0)$ is invariant with respect to a parameter transformation $g(\theta, \lambda) \in \mathbb{R}^{n_\theta}$ with $\lambda \in \mathbb{R}^{n_\lambda}$ and $0 < n_\lambda < n_\theta$: for any $\theta \in \mathbb{D}_\theta$ and $\lambda \in \mathbb{D}_\lambda \subset \mathbb{R}^{n_\lambda}$, with $\dim(\mathbb{R}^{n_\lambda} \setminus \mathbb{D}_\lambda) < n_\lambda$, we have $\operatorname{rank}(\partial g(\theta, \lambda)/\partial\lambda) = n_\lambda$ and $M(g(\theta, \lambda), z_0) = M(\theta, z_0)$.

Assumption 18.4 (Identifiable Models): The model $M(\theta, z_0)$ can be parameterized in β overlapping identifiable parameter sets $\psi_k \in \mathbb{R}^{n_\psi}$, satisfying

1. $\forall\theta \in \mathbb{D}_k, \exists\ \lambda_k(\theta) \in \mathbb{D}_\lambda : h_k(\psi_k) = g(\theta,\ \lambda_k(\theta))$
2. $\mathbb{D}_\theta = \bigcup_{k=1}^{\beta} \mathbb{D}_k$
3. $\dim(\mathbb{D}_k) = n_\theta$ and $\dim(\mathbb{D}_k \setminus \mathbb{D}_l) < n_\theta$

with $n_\psi = n_\theta - n_\lambda$ and $k, l = 1, 2, \ldots, \beta$. The function $\psi_k(\theta)$ and its derivative w.r.t. θ, $\psi_k'(\theta)$, are continuous in $\mathbb{D}_k$. $\psi_k'(\theta)$ has full rank n_ψ in $\mathbb{D}_k$.

Note that Assumptions 18.3 and 18.4 define a manifold with boundary given by an equivalence relationship. Assumptions 18.4 (1) and (2) guarantee that $M(\theta, z_0)$ can be parameterized in at least one identifiable parameter set ψ_k, $k = 1, 2, \ldots, \beta$, for any value of $\theta \in \mathbb{D}_\theta$. Assumption 18.4 (3) implies that the parameterizations are overlapping: each identifiable parameter set ψ_k can represent any model $M(\theta, z_0)$, except those corresponding to θ-values lying in some lower dimensional ($< n_\theta$) subspaces of $\mathbb{D}_\theta$.

Definition 18.5: An invariant of the model $M(\theta, z_0)$ is each model-related quantity $I(\theta)$ that is invariant w.r.t. the same parameter transformation $g(\theta, \lambda)$ as the model itself: $I(g(\theta, \lambda)) = I(\theta)$ (see Assumption 18.3).

The following two examples show that Assumptions 18.3 and 18.4 are satisfied in many practical identification problems.

Example 18.6 (Rational Transfer Function Model): Consider transfer function model (18-1) of Example 18.1 at angular frequencies ω_f, $f = 1, 2, \ldots, N$. Clearly, $G(j\omega_f, \lambda\theta) = G(j\omega_f, \theta)$. If, for example, coefficient a_0 is fixed to one, then the resulting identifiable parameter set ψ_0 can describe all models of the form (18-1), except those for which $a_0 = 0$. If a_1 is fixed to one, then the resulting identifiable set ψ_1 covers the subspace $\{a_0 = 0\}$ but cannot describe models with $a_1 = 0$. These observations are now stated more formally in the framework of Assumptions 18.3 and 18.4.

Assumption 18.3 is valid with

$$M(\theta, z_0) = \left[G(j\omega_1, \theta) - G_0(j\omega_1)\ G(j\omega_2, \theta) - G_0(j\omega_2)\ \ldots\ G(j\omega_N, \theta) - G_0(j\omega_N)\right]^T$$

$z_0 = [G_0(j\omega_1)G_0(j\omega_2)\ldots G_0(j\omega_N)]^T$, $G_0(j\omega)$ the true frequency response function, $n_\theta = n_a + n_b + 2$, $g(\theta, \lambda) = \lambda\theta$, $n_\lambda = 1$, $\mathbb{D}_\lambda = \mathbb{R}_0$, and

$$\mathbb{S}_\theta = \{\theta | \exists f \in \{1, 2, \ldots, N\} \text{ s.t. } G(j\omega_f, \theta) = \infty \text{ or } 0/0\}$$

Assumption 18.4 is valid with $\beta = n_\theta$, $n_\psi = n_\theta - 1$, $\lambda_k(\theta) = 1/\theta_k$,

$$h_k^T(\psi_k) = [\psi_{k[1]} \cdots \psi_{k[k-1]}\, 1\, \psi_{k[k+1]} \cdots \psi_{k[n_\psi]}]$$

and $\mathbb{D}_k = \mathbb{D}_\theta \setminus \{\theta_{[k]} = 0\}$. The relationship between the identifiable parameters ψ_k and the overparameterized set θ is given by

$$\psi_k^T(\theta) = \frac{1}{\theta_{[k]}}\left[\theta_{[1]} \;\dots\; \theta_{[k-1]}\; \theta_{[k+1]} \;\dots\; \theta_{[n_\theta]}\right] \qquad \square$$

Example 18.7 (State Space Model): Consider transfer function model (18-2) at the frequencies $z_f = e^{2\pi jf/N}$, $f = 0, 1, \dots, N-1$. The model $M(\theta, z_0)$ can be defined in exactly the same way as in Example 18.6. Assumption 18.3 is valid with $\lambda = \text{vec}(T)$, $n_\lambda = n^2$, and $\mathbb{D}_\lambda = \mathbb{R}^{n_\lambda}\setminus\{\lambda|\det(T) = 0\}$. Applying property (13-38) of the Kronecker product to (18-3), where (A, B, C, D) is replaced by $(TAT^{-1}, TB, CT^{-1}, D)$, gives an explicit expression for $g(\theta, \lambda)$

$$g(\theta, \lambda) = \text{diag}(T^{-T} \otimes T, I_{n_u} \otimes T, T^{-T} \otimes I_{n_y}, I_{n_y n_u})\theta$$

with diag(.) a block diagonal matrix. The existence and the properties of overlapping identifiable parameterizations, as required in Assumption 18.4, have been shown for linear multivariable systems in Hazewinkel (1977), Delchamps and Byrnes (1982), and Delchamps (1985), while Van Overbeek and Ljung (1982) discuss the numerical aspects when switching from one identifiable form to another. □

18.3 CRAMÉR-RAO LOWER BOUND FOR INVARIANTS OF (OVER)PARAMETERIZED MODELS

Consider the identification of a particular model $M(\theta, z_0)$ using noisy observations $z \in \mathbb{R}^N$, and assume that the true model belongs to the considered model set. The Cramér-Rao lower bounds on the covariance matrix of an estimator $\hat{I}(z)$ using (over)parameterization θ and ψ_k are given respectively by (see Theorem 14.18)

$$\begin{aligned} &\text{Cov}(\hat{I}(\hat{\theta}(z))) \geq \left(\frac{\partial I_0}{\partial \theta_0} + \frac{\partial b_I}{\partial \theta_0}\right) Fi^{+}(\theta_0) \left(\frac{\partial I_0}{\partial \theta_0} + \frac{\partial b_I}{\partial \theta_0}\right)^H &&\text{(a)} \\ &\text{Cov}(\hat{I}(h_k(\hat{\psi}_k(z)))) \geq \left(\frac{\partial I_0}{\partial \psi_{k0}} + \frac{\partial b_I}{\partial \psi_{k0}}\right) Fi^{-1}(\psi_{k0}) \left(\frac{\partial I_0}{\partial \psi_{k0}} + \frac{\partial b_I}{\partial \psi_{k0}}\right)^H &&\text{(b)} \end{aligned} \qquad (18\text{-}4)$$

with $I_0 = I(\theta_0)$, $b_I = \mathscr{E}\{\hat{I}(z)\} - I_0$ the bias that might be present in the estimate, $Fi(\theta_0)$ the singular Fisher information matrix of the overparameterized parameter vector θ_0, and $Fi(\psi_{k0})$ the regular Fisher information matrix of the identifiable parameter vector ψ_{k0}. Because we will compare (18-4a) with (18-4b), the Fisher information matrix $Fi(\theta_0)$ should be constrained to exclude the $\hat{\theta}(z)$ -values lying in the lower dimensional subspace $\mathbb{R}^{n_\theta} \setminus \mathbb{D}_k$. In Gorman and Hero (1990) it has been proved under some suitable regularity conditions that the constrained Cramér-Rao lower bound equals the unconstrained case if the constraints are not active in θ_0. We conclude that we can compare (18-4a) with (18-4b) if $\theta_0 \notin \mathbb{R}^{n_\theta} \setminus \mathbb{D}_k$.

Theorem 18.8 (Cramér-Rao Bound of Overparameterized Models): Under Assumptions 18.3 and 18.4, the Cramér-Rao lower bounds (18-4a) and (18-4b) of any estimator $\hat{I}(z)$ of an invariant $I(\theta)$ of the model $M(\theta, z_0)$ are independent of the particular (over)parameterization (θ) ψ_k chosen.

Proof. See Appendix 18.A. □

The theorem has been proved by comparing an overparameterized parameter set with an identifiable parameter set. By applying the same reasoning twice, the conclusions are also valid when comparing two different overparameterized parameter sets.

Intuitively, the theorem can be understood as follows. The null space of the Fisher information matrix spanned by the redundant model parameters is not affected by the noise. Hence, the redundant parameters will not increase the variance of identifiable model parameters.

18.4 ESTIMATES OF INVARIANTS OF (OVER)PARAMETERIZED MODELS— FINITE SAMPLE RESULTS

Consider the identification of a particular model $M(\theta, z_0)$ using noisy observations $z \in \mathbb{R}^N$ of the true signal z_0. The basic question now arises whether estimates of invariants $\hat{I}(\hat{\theta}(z))$, $\hat{I}(h_k(\hat{\psi}_k(z)))$, of model $M(\theta, z_0)$, calculated from estimates $\hat{\theta}(z)$, $\hat{\psi}_k(z)$, depend on the particular (over)parameterization θ, ψ_k. In general, the answer is yes: not only each realization but also the statistical properties (bias, uncertainty, ...) of the estimated invariants strongly depend on this choice (De Moor et al., 1994). In this section we show that estimators whose corresponding (equivalent) cost functions are invariant with respect to the same parameter transformation as the overparameterized model lead to estimates of invariants that *do not* depend on this choice. While in Section 18.3 it has been proved that the uncertainty of (asymptotically) efficient estimators is independent of the (over)parameterization θ, ψ_k, the result presented in this section is valid for any finite sample property (the distribution function, the sample mean, the sample variance, ..., and if the moments exist, the bias, variance, ...) of the estimated invariants and is not restricted to the class of (asymptotically) efficient estimators.

18.4.1 The Estimators

The estimators considered in this section minimize a cost function $V(\theta, z)$ that is defined for any $\theta \in \mathbb{D}_\theta$, a subspace of $\mathbb{R}^{n_\theta}$, and any $z \in \mathbb{D}_z$, a subspace of $\mathbb{R}^N$. The complements of $\mathbb{D}_\theta$ in $\mathbb{R}^{n_\theta}$ and $\mathbb{D}_z$ in $\mathbb{R}^N$ are, respectively, $\mathbb{S}_\theta$ and $\mathbb{S}_z$, the sets of singular points. $\mathbb{S}_\theta = \mathbb{R}^{n_\theta} \setminus \mathbb{D}_\theta$ and $\mathbb{S}_z = \mathbb{R}^N \setminus \mathbb{D}_z$ have topological dimensions less than, respectively, n_θ and N.

Assumption 18.9 (Invariance Cost Function): The (equivalent) cost function $V(\theta, z)$ minimized by the estimator is invariant with respect to the same parameter transformation $g(\theta, \lambda)$ as the model itself (see Assumption 18.3): $V(g(\theta, \lambda), z) = V(\theta, z)$, for any $\lambda \in \mathbb{D}_\lambda$, $\theta \in \mathbb{D}_\theta$, and $z \in \mathbb{D}_z$.

Very often the cost function is a function of some transformed form of the system model $M(\theta, z_0)$ and, therefore, Assumption 18.9 is, in general, not true. This is illustrated in the following example.

Example 18.10 (Frequency Domain Identification): Consider the identification of model (18-1) starting from frequency response function measurements $G(j\omega_f)$, $f = 1, 2, \ldots, N$. The linear least squares estimate minimizes (see Section 7.8)

$$V_{\text{LS}}(\theta, z) = \sum_{f=1}^{N} |A(j\omega_f, \theta)G(j\omega_f) - B(j\omega_f, \theta)|^2 \tag{18-5}$$

w.r.t. θ. Because $V_{LS}(\lambda\theta, z) = \lambda^2 V_{LS}(\theta, z)$, it follows directly that Assumption 18.9 is not fulfilled. Often, the frequency response function $G(j\omega_f)$ is obtained as the ratio of the output to the input DFT spectra $G(j\omega_f) = Y(f)/(U(f)$. The cost function is infinitely large for $U(f) = 0$ and, hence, $\mathbb{S}_z = \{z|U(f) = 0\}$. □

For each particular observation $z \in \mathbb{D}_z$, θ-values in $\mathbb{S}_\theta$ may exist for which the cost function $V(\theta, z)$ is not defined or infinitely large. At these singular points, the first- and second-order derivative of the cost function w.r.t. the model parameters either do not exist or are rank deficient. These θ-values should, therefore, be excluded during the minimization of $V(\theta, z)$. This can be done by introducing a regularization parameter μ in the cost function.

Assumption 18.11 (Regularized Cost Function): The cost $V(\theta, z)$ can be regularized as $V(\theta, z, \mu)$, with $\mu \in \mathbb{R}$, such that $V(\theta, z, 0) = V(\theta, z)$ and

1. $V(\theta, z, \mu)$ is a continuous function of μ, and $V(g(\theta, \lambda), z, \mu) = V(\theta, z, \mu)$
2. $H_{\psi_k\psi_k} = \partial^2 V(h_k(\psi_k), z, \mu)/\partial\psi_k^2$ has rank n_ψ and is a jointly continuous function of ψ_k, z
3. $H_{\psi_k z} = \partial^2 V(h_k(\psi_k), z, \mu)/\partial\psi_k\partial z$ has rank n_ψ and is a jointly continuous function of ψ_k, z

for any $\psi_k \in \mathbb{R}^{n_\psi}$, $k = 1, 2, \ldots, \beta$, for any $z \in \mathbb{R}_z \subset \mathbb{R}^N$ $(\dim(\mathbb{R}^N \setminus \mathbb{R}_z) < N)$, and for any $\mu \in \mathbb{R}_0$.

Due to Assumption 18.11(1), the regularization parameter μ can always be chosen such that the difference $V(\theta, z, \mu) - V(\theta, z)$ is arbitrarily small in the regular space $\mathbb{D}_\theta$. Proceeding in this way, μ is active only in the singular subspace $\mathbb{S}_\theta$ and this is exactly how the regularization is applied in practice. This motivates the following definition.

Definition 18.12: The estimates $\hat{\theta} = \hat{\theta}(z)$, $\hat{\psi}_k = \hat{\psi}_k(z)$ are the minimizing arguments of $V(\theta, z, \mu)$ and $V(h_k(\psi_k), z, \mu)$, respectively

$$\hat{\theta}(z) = \arg\min_{\theta} V(\theta, z, \mu) \quad \text{and} \quad \hat{\psi}_k(z) = \arg\min_{\psi_k} V(h_k(\psi_k), z, \mu)$$

Estimators satisfying Assumptions 18.9 and 18.11 are, for example, the nonlinear least squares (see Section 7.9), the generalized total least squares (see Section 7.10 and Van Huffel and Vandewalle, 1991), the one-step bootstrapped total least squares (see Section 7.10), and the maximum likelihood (see Section 7.11; Vandersteen et al., 1996a). Counterexamples in system identification are the linear least squares (see Section 7.8; Kalman, 1958; De Moor et al., 1994), the (iteratively) weighted linear least squares (see Section 7.8; Steigliz and McBride, 1965), and parametric time series analysis using, for example, the conditional maximum likelihood method (Box and Jenkins, 1976).

The z-values for which the cost function is singular ($\forall z \in \mathbb{S}_z$) and/or its second-order partial derivatives are not of full rank ($\forall z \in \mathbb{R}^N \setminus \mathbb{R}_z$, see Assumption 18.11) lie in lower dimensional subspaces of $\mathbb{R}^N$. To ensure that these values occur with probability zero, the following assumption is made.

Assumption 18.13 (pdf Noise): The probability density function of the disturbing errors $n_z = z - z_0$ is continuous.

Under Assumption 18.13, it is clear that $\text{Prob}(z \in \mathbb{S}_z) = \text{Prob}(z \in \mathbb{R}^N \setminus \mathbb{R}_z) = 0$. Assumptions 18.9 and 18.11 are illustrated in the following example.

Example 18.14 (Frequency Domain Identification): Consider the identification of model (18-1) starting from frequency response function measurements $G(j\omega_f)$, $f = 1, 2, \ldots, N$. The nonlinear least squares estimate minimizes (see Section 7.9)

$$V_{\text{NLS}}(\theta, z) = \sum_{f=1}^{N} |G(j\omega_f) - G(j\omega_f, \theta)|^2 \tag{18-6}$$

Because $G(j\omega_f, \lambda\theta) = G(j\omega_f, \theta)$, it follows directly that $V_{\text{NLS}}(\lambda\theta, z) = V_{\text{NLS}}(\theta, z)$. The cost function (18-6) is infinitely large for θ-values satisfying $a_k = 0$, $k = 0, 1, \ldots, n_a$. It is undefined for $\theta = 0$ and θ-values such that $G(j\omega_f, \theta)$ has a common pole-zero pair at $j\omega_f$. Hence,

$$\mathbb{D}_\theta = \mathbb{R}^{n_\theta} \setminus \{\theta | f \in \{1, 2, \ldots, N\} \text{ s.t. } G(j\omega_f, \theta) = \infty \text{ or } 0/0\}$$

Using $G(j\omega_f, \theta) = B(j\omega_f, \theta)/A(j\omega_f, \theta)$, the regularized cost function becomes

$$V_{\text{NLS}}(\theta, z, \mu) = \sum_{f=1}^{N} \frac{|A(j\omega_f, \theta)G(j\omega_f) - B(j\omega_f, \theta)|^2}{|A(j\omega_f, \theta)|^2 + \mu^2\theta^T\theta}$$ □

Quadratic cost functions satisfying Assumption 18.9 can always be written as $V(\theta, z) = \varepsilon^T(\theta, z)\varepsilon(\theta, z)/2$, where the residual $\varepsilon(\theta, z)$ is also invariant w.r.t. the parameter transformation $g(\theta, \lambda)$: $\varepsilon(g(\theta, \lambda), z) = \varepsilon(\theta, z)$. For such cost functions, it is possible to make statements about the null space of the Jacobian matrix $\partial\varepsilon(\theta, z)/\partial\theta$ w.r.t. the overparameterized θ.

Theorem 18.15 (Jacobian Matrix of Overparameterized Models): Under Assumptions 18.9 and 18.11, the Jacobian matrix $\partial\varepsilon(\theta, z)/\partial\theta$ of the quadratic cost function $V(\theta, z) = \varepsilon^T(\theta, z)\varepsilon(\theta, z)/2$ has a null space of dimension n_λ for any $\theta \in \mathbb{D}_k$ and $z \in \mathbb{D}_z$.

Proof. See Appendix 18.B. □

The dimension of the null space of the Jacobian matrix w.r.t. θ is independent of the noise on the observations z and of the model errors. For quadratic cost functions that violate Assumption 18.9, the dimension of the null space of the Jacobian matrix w.r.t. θ is affected by the noise and the model errors (see Exercises 18.2 and 18.4).

18.4.2 Main Result

Theorem 18.16: Under Assumptions 18.3, 18.4, 18.9, 18.11, and 18.13, the estimate of an invariant $I(\theta)$ of model $M(\theta, z_0)$ is with probability one independent of the particular (over)parameterization chosen: $\hat{I}(\hat{\theta}(z)) = \hat{I}(h_k(\hat{\psi}_k(z)))$ for $k = 1, 2, \ldots, \beta$.

Proof. See Appendix 18.C. □

Because $N = \dim(z)$ is fixed in the analysis, the theorem is valid for finite values of N. As it has not been assumed that the true model is within the model set, the theorem is also valid when there are model errors. As it has not been assumed that model $M(\theta, z_0)$ is linear in z_0, the theorem is also valid for nonlinear systems. An application of the nonlinear case can be found in Vandersteen et al. (1996a). The theorem also applies to estimators minimizing a series of cost functions where each cost function satisfies Assumptions 18.9 and 18.11. An example of such an estimator is the multistep bootstrapped total least squares method (see Section 7.12.3). Because the estimates of the invariants are independent of the (over)parameterization chosen, one should use the particular (over)parameterization that leads to the numerical best conditioned set of equations (see, for example, Van Overbeek and Ljung, 1982 and also Practical remark at the end of Section 18.5).

18.5 A SIMPLE NUMERICAL EXAMPLE

Consider again Example 18.14 with $N = 100$ data points, ω_f, $f = 1, 2, \ldots, N$ equally distributed in the band $[1, 1500]2\pi$ rad/s, and a true plant transfer function

$$G_0(s) = \frac{1 + 2\times10^{-4} s}{1 + 1\times10^{-3} s} \tag{18-7}$$

Independent, circular complex distributed noise $n_G(j\omega)$ with zero mean and variance 2×10^{-2} is added to the true transfer function $G(j\omega_f) = G_0(j\omega_f) + n_G(j\omega_f)$, $f = 1, 2, \ldots, N$. A hundred independent sets of 100 noisy transfer function values are generated. For each noisy data set the nonlinear least squares estimate (18-6), which satisfies Assumption 18.9 ($V_{\mathrm{NLS}}(\lambda\theta, z) = V_{\mathrm{NLS}}(\theta, z)$), and the linear least squares estimate (18-5), which violates Assumption 18.9 ($V_{\mathrm{LS}}(\lambda\theta, z) \neq V_{\mathrm{LS}}(\theta, z)$), are calculated for two models; one without model errors (18-8) and the other one with model errors (18-9).

$$G(s, \theta) = \frac{b_0 + b_1 s}{a_0 + a_1 s} \tag{18-8}$$

$$G(s, \theta) = \frac{b_0}{a_0 + a_1 s} \tag{18-9}$$

18.5.1.1 Model without Model Errors (18-8). The four identifiable forms of the overparameterized model (18-8) that satisfy Assumption 18.4 are

$$G(s, \psi_1) = \frac{\psi_{1[2]} + \psi_{1[3]} s}{1 + \psi_{1[1]} s},\ G(s, \psi_2) = \frac{\psi_{2[2]} + \psi_{2[3]} s}{\psi_{2[1]} + s},\ G(s, \psi_3) = \frac{1 + \psi_{3[3]} s}{\psi_{3[1]} + \psi_{3[2]} s},\ \text{and}$$

$$G(s, \psi_4) = \frac{\psi_{4[3]} + s}{\psi_{4[1]} + \psi_{4[2]} s}$$

For each of the 100 noisy data sets, the linear and nonlinear least squares estimates of the identifiable model parameters ψ_k, $k = 1, \ldots, 4$, and the overparameterized model parameter θ are calculated. The latter are obtained by solving the normal equations using the pseudoinverse. Three invariants of model (18-8) are the gain $K = b_0/a_0$ at $s = 0$ and the time constants $\tau_1 = b_1/b_0$ and $\tau_2 = a_1/a_0$. These invariants are calculated for each of the 100 least squares and nonlinear least squares estimates of $\hat{\psi}_k$, $k = 1, \ldots, 4$, and $\hat{\theta}$. The results are shown in Table 18-1. It follows that the linear least squares estimates of the invariants are strongly dependent on the particular (over)parameterization chosen. This is explained by the fact that the linear least squares cost function $V_{\text{LS}}(\theta, z)$ is not invariant w.r.t. the transformation $g(\theta, \lambda) = \lambda\theta$: $V_{\text{LS}}(\lambda\theta, z) \neq V_{\text{LS}}(\theta, z)$. Note that the estimates based on parameterizations ψ_2 (monic denominator polynomial) and θ perform much better than the other ones. From Table 18-1, it also follows that the nonlinear least squares estimates and their sample deviations are equal within the numerical precision of the calculations (12 digits). This is also true (but not shown in Table 18-1) for the estimates of each of the 100 data sets separately.

18.5.1.2 Model with Model Errors (18-9). The three identifiable forms of the overparameterized model (18-9) that satisfy Assumption 18.4 are

$$G(s, \psi_1) = \frac{\psi_{1[2]}}{1 + \psi_{1[1]}s}, \quad G(s, \psi_2) = \frac{\psi_{2[2]}}{\psi_{2[1]} + s}, \text{ and } G(s, \psi_3) = \frac{1}{\psi_{3[1]} + \psi_{3[2]}s}$$

For each of the 100 noisy data sets, the linear and nonlinear least squares estimates of the identifiable model parameters ψ_k, $k = 1, 2, 3$, and the overparameterized model parameter θ are calculated. Two invariants of model (18-9) are the gain $K = b_0/a_0$ at $s = 0$ and the time constant $\tau = a_1/a_0$. These invariants are calculated for each of the 100 least squares and nonlinear least squares estimates of $\hat{\psi}_k$, $k = 1, 2, 3$, and $\hat{\theta}$. From Table 18-2 it follows that the conclusions of the preceding section are also valid for model errors.

TABLE 18-1 Estimated Invariants K, τ_1, and τ_2 of Model (18-8)

Invariant	Least Squares		Nonlinear Least Squares	
	Sample Mean	Sample Sandard Deviation	Sample Mean	Sample Standard Deviation
$\tau_1(\hat{\psi}_1)$	-8.59e-07	7.0e-06	1.9956e-04	1.9e-05
$\tau_1(\hat{\psi}_2)$	2.236e-04	2.9e-05	1.9956e-04	1.9e-05
$\tau_1(\hat{\psi}_3)$	6.394e-05	8.5e-06	1.9956e-04	1.9e-05
$\tau_1(\hat{\psi}_4)$	4.805e-04	5.3e-05	1.9956e-04	1.9e-05
$\tau_1(\hat{\theta})$	1.879e-04	2.7e-05	1.9956e-04	1.9e-05
$\tau_2(\hat{\psi}_1)$	1.011e-04	1.4e-05	1.0055e-03	6.6e-05
$\tau_2(\hat{\psi}_2)$	1.327e-03	1.5e-04	1.0055e-03	6.6e-05
$\tau_2(\hat{\psi}_3)$	3.156e-04	3.2e-05	1.0055e-03	6.6e-05
$\tau_2(\hat{\psi}_4)$	7.7e-03	2.3e-01	1.0055e-03	6.6e-05
$\tau_2(\hat{\theta})$	8.84e-04	1.1e-04	1.0055e-03	6.6e-05
$K(\hat{\psi}_1)$	3.886e-01	1.6e-02	1.0024	3.8e-02
$K(\hat{\psi}_2)$	1.2198	7.8e-02	1.0024	3.8e-02
$K(\hat{\psi}_3)$	7.410e-01	3.2e-02	1.0024	3.8e-02
$K(\hat{\psi}_4)$	4	140	1.0024	3.8e-02
$K(\hat{\theta})$	9.417e-01	6.4e-02	1.0024	3.8e-02

TABLE 18-2 Estimated Invariants K and τ of Model (18-9)

Invariant	Least Squares		Nonlinear Least Squares	
	Sample Mean	Sample Standard Deviation	Sample Mean	Sample Standard Deviation
$\tau(\hat{\psi}_1)$	1.017e-04	1.1e-05	5.3026e-04	5.9e-05
$\tau(\hat{\psi}_2)$	6.068e-04	4.4e-05	5.3026e-04	5.9e-05
$\tau(\hat{\psi}_3)$	1.673e-04	1.9e-05	5.3026e-04	5.9e-05
$\tau(\hat{\theta})$	2.184e-04	3.0e-05	5.3026e-04	5.9e-05
$K(\hat{\psi}_1)$	3.892e-01	1.4e-02	7.6796e-01	4.4e-02
$K(\hat{\psi}_2)$	8.399e-01	3.2e-02	7.6796e-01	4.4e-02
$K(\hat{\psi}_3)$	6.395e-01	2.2e-02	7.6796e-01	4.4e-02
$K(\hat{\theta})$	4.920e-01	2.2e-02	7.6796e-01	4.4e-02

18.5.1.3 Practical Remark. From the many simulations that have been conducted, it can be concluded that the overparameterized models lead to one of the best, but not always the best, condition numbers of the normal equations. This is in agreement with the results of McKelvey and Helmersson (1997). It also followed that the overparameterized form often gives the best linear least squares estimate (see, for example, Table 18-1), which is in agreement with the results of De Moor et al. (1994). Hence, even for estimators that violate Assumption 18.9, the overparameterized models are strongly recommended.

18.6 EXERCISES

18.1. Consider Examples 18.1, 18.6, and 18.14. Show that the null space of the Jacobian matrix of the nonlinear least squares estimate (18-6) has dimension 1 (hint: show that $\text{rank}(\frac{\partial g(\theta, \lambda)}{\partial \lambda}) = 1$).

18.2. Consider Examples 18.1, 18.6, and 18.14. Show that the null space of the Jacobian matrix of the linear least squares estimate (18-5) has dimension 0 unless it is evaluated in the noiseless data z_0 and the true model parameters θ_0 (hint: use $\varepsilon(\theta, z) = \frac{\partial \varepsilon(\theta, z)}{\partial \theta}\theta$).

18.3. Consider Examples 18.2 and 18.7. Show that the null space of the Jacobian matrix of the nonlinear least squares estimate (18-6) has dimension n^2 (hint: show that $\text{rank}(\frac{\partial g(\theta, \lambda)}{\partial \lambda}) = n^2$).

18.4. Redo the simulations of Section 18.5. Verify that the Jacobian matrix $\partial e(\theta, z)/\partial\theta$ of the nonlinear least squares estimate has a null space of dimension 1 for each noise realization. Verify that the Jacobian matrix $\partial e(\theta, z)/\partial\theta$ of the linear least squares estimate is not rank deficient for noisy observations and/or in the presence of model errors.

18.7 APPENDIXES

Appendix 18.A: Proof of Theorem 18.8 (Cramér-Rao Bound of (Over)Parameterized Models)

Applying the chain rule for the partial derivatives on $Fi(\theta_0)$, (14-85) and $\partial I_0/\partial\theta_0$, makes it possible to rewrite (18-4a) in $\mathbb{D}_k$ as

$$\operatorname{Cov}(\hat{I}(\hat{\theta}(z))) \geq \left(\frac{\partial I}{\partial \psi_k}+\frac{\partial b_I}{\partial \psi_k}\right)\left(\frac{\partial \psi_k}{\partial \theta}\right)\left[\left(\frac{\partial \psi_k}{\partial \theta}\right)^T Fi(\psi_k)\left(\frac{\partial \psi_k}{\partial \theta}\right)\right]^+\left(\frac{\partial \psi_k}{\partial \theta}\right)^T\left(\frac{\partial I}{\partial \psi_k}+\frac{\partial b_I}{\partial \psi_k}\right)^H \qquad (18\text{-}10)$$

If we can show that

$$\left(\frac{\partial \psi_k}{\partial \theta}\right)\left[\left(\frac{\partial \psi_k}{\partial \theta}\right)^T Fi(\psi_k)\left(\frac{\partial \psi_k}{\partial \theta}\right)\right]^+\left(\frac{\partial \psi_k}{\partial \theta}\right)^T = Fi^{-1}(\psi_k) \qquad (18\text{-}11)$$

in $\mathbb{D}_k$, then the right-hand sides of (18-4a) and (18-4b) are equal in $\mathbb{D}_k$, which proves the theorem. Using the first Moore-Penrose condition of the pseudoinverse $AA^+A = A$ (see Section 13.5) with $A = \left(\frac{\partial \psi_k}{\partial \theta}\right)^T Fi(\psi_k)\left(\frac{\partial \psi_k}{\partial \theta}\right)$ gives

$$\left(\frac{\partial \psi_k}{\partial \theta}\right)^T Fi(\psi_k)\left(\frac{\partial \psi_k}{\partial \theta}\right)\left[\left(\frac{\partial \psi_k}{\partial \theta}\right)^T Fi(\psi_k)\left(\frac{\partial \psi_k}{\partial \theta}\right)\right]^+\left(\frac{\partial \psi_k}{\partial \theta}\right)^T Fi(\psi_k)\left(\frac{\partial \psi_k}{\partial \theta}\right) = \left(\frac{\partial \psi_k}{\partial \theta}\right)^T Fi(\psi_k)\left(\frac{\partial \psi_k}{\partial \theta}\right) \qquad (18\text{-}12)$$

Left multiplication with $\left(\frac{\partial \psi_k}{\partial \theta}\right)$ and right multiplication with $\left(\frac{\partial \psi_k}{\partial \theta}\right)^T$ of (18-12) give

$$QFi(\psi_k)\left(\frac{\partial \psi_k}{\partial \theta}\right)\left[\left(\frac{\partial \psi_k}{\partial \theta}\right)^T Fi(\psi_k)\left(\frac{\partial \psi_k}{\partial \theta}\right)\right]^+\left(\frac{\partial \psi_k}{\partial \theta}\right)^T Fi(\psi_k)Q = QFi(\psi_k)Q \qquad (18\text{-}13)$$

where $Q = \left(\frac{\partial \psi_k}{\partial \theta}\right)\left(\frac{\partial \psi_k}{\partial \theta}\right)^T$ is a regular n_ψ by n_ψ matrix in $\mathbb{D}_k$ (Assumption 18.4). Left division by $QFi(\psi_k)$ and right division by $Fi(\psi_k)Q$ of (18-13) give (18-11). □

Appendix 18.B: Proof of Theorem 18.15 (Jacobian Matrix of Overparameterized Models)

Because the residual $\varepsilon(\theta, z)$ is independent of λ, we have $\partial\varepsilon(\theta, z)/\partial\lambda = 0$. Using $\varepsilon(g(\theta, \lambda), z) = \varepsilon(\theta, z)$, this equality can be written as

$$\frac{\partial \varepsilon(g, z)}{\partial g}\frac{\partial g(\theta, \lambda)}{\partial \lambda} = 0 \qquad (18\text{-}14)$$

for any $\theta \in \mathbb{D}_\theta$ and $\lambda \in \mathbb{D}_\lambda$, and with $\partial\varepsilon(g, z)/\partial g \in \mathbb{R}^{N \times n_\theta}$ $(N \geq n_\theta)$ the Jacobian matrix. Because $\operatorname{rank}(\partial g(\theta, \lambda)/\partial\lambda) = n_\lambda$ with $n_\lambda < n_\theta$ (Assumption 18.3), it follows from (18-14) that $\dim(\operatorname{null}(\partial\varepsilon(g, z)/\partial g)) = n_\lambda$. □

Appendix 18.C: Proof of Theorem 18.16

The regularized cost function $V(\theta, z, \mu)$ is an invariant of model $M(\theta, z_0)$ (Assumption 18.11). This implies that for any $\hat{\theta} \in \mathbb{D}_k \cap \mathbb{D}_l$ $(k \neq l)$ $h_k(\hat{\psi}_k) = g(\hat{\theta}, \lambda_k(\hat{\theta}))$ and $g(\hat{\theta}, \lambda_l(\hat{\theta})) = h_l(\hat{\psi}_l)$; otherwise $\hat{\psi}_k$ and $\hat{\psi}_l$ would not be the minimizers of $V(h_k(\psi_k), z, \mu)$ and $V(h_l(\psi_l), z, \mu)$ respectively. Using $I(g(\theta, \lambda(\theta))) = I(\theta)$ (Definition 18.5), it follows that for any $\hat{\theta} \in \mathbb{D}_k \cap \mathbb{D}_l$,

$$\hat{I}(h_k(\hat{\psi}_k)) = \hat{I}(g(\hat{\theta}, \lambda_k(\hat{\theta}))) = \hat{I}(\hat{\theta}) \quad \text{and} \quad \hat{I}(h_l(\hat{\psi}_l)) = \hat{I}(g(\hat{\theta}, \lambda_l(\hat{\theta}))) = \hat{I}(\hat{\theta}) \tag{18-15}$$

and, hence, $\hat{I}(h_k(\hat{\psi}_k)) = \hat{I}(h_l(\hat{\psi}_l))$. The theorem is proved if it can be shown that the event '$\hat{\theta} \in \mathbb{D}_k \cap \mathbb{D}_l$' occurs with probability one. Therefore, it is sufficient to prove that the probability density functions (pdf's) of the estimates $\hat{\psi}_k$, $k = 1, 2, \ldots, \beta$, are continuous. Indeed, accumulation of probability mass in a lower dimensional space can occur only if the distribution function is degenerate (Papoulis, 1981). If the pdf of $\hat{\psi}_k$ is continuous, then the probability that $\hat{\psi}_k$ lies in a lower dimensional subspace of $\mathbb{R}^{n_\psi}$ equals zero. Under Assumption 18.4, the θ-values that cannot be represented by, for example, ψ_k, represent values of, for example, ψ_l, lying in a lower dimensional subspace of $\mathbb{R}^{n_\psi}$, so that $\text{Prob}(\hat{\theta} \in \mathbb{D}_k \setminus \mathbb{D}_l) = 0$. Similarly, $\text{Prob}(\hat{\theta} \in \mathbb{D}_l \setminus \mathbb{D}_k) = 0$ and, thus, $\text{Prob}(\hat{\theta} \in \mathbb{D}_k \cap \mathbb{D}_l) = 1$.

By Assumption 18.13, the pdf of the noisy data z is continuous. To prove that the probability density functions of the estimates $\hat{\psi}_k$, $k = 1, 2, \ldots, \beta$, are continuous it is, therefore, sufficient to show that $\hat{\psi}_k = \hat{\psi}_k(z)$ is a continuous function of z with continuous derivatives satisfying

$$\dim(\{z \,|\, \text{rank}(\frac{\partial \hat{\psi}_k(z)}{\partial z}) < n_\psi\}) < N \tag{18-16}$$

(Papoulis, 1981). If (18-16) is not satisfied, then the distribution function of $\hat{\psi}_k$ may be degenerate in the subspace $\mathbb{D}_k \setminus \mathbb{D}_l$, so that accumulation of probability mass may occur in $\mathbb{D}_k \setminus \mathbb{D}_l$. In this case, $\text{Prob}(\hat{\theta} \in \mathbb{D}_k \setminus \mathbb{D}_l) \neq 0$ and the theorem is no longer valid. The function $\hat{\psi}_k(z)$ is implicitly known by the definition of the minimizer $\hat{\psi}_k$

$$\left(\frac{\partial V(h_k(\hat{\psi}_k), z, \mu)}{\partial \hat{\psi}_k}\right)^T = 0 \text{ or } F(\hat{\psi}_k, z, \mu) = 0 \tag{18-17}$$

Under Assumption 18.11, $F(\hat{\psi}_k, z, \mu) \in \mathbb{R}^{n_\psi}$ has continuous, full rank, first-order partial derivatives w.r.t. $\hat{\psi}_k$ and z for any $\hat{\psi}_k \in \mathbb{R}^{n_\psi}$ and $z \in \mathbb{R}_z$ with $\dim(\mathbb{R}^N \setminus \mathbb{R}_z) < N$. Applying the implicit function theorem (Kaplan, 1993) to $F(\hat{\psi}_k, z, \mu) = 0$ shows that $\hat{\psi}_k = \hat{\psi}_k(z)$ is a continuous function of z with full rank, continuous derivative

$$\begin{aligned} \frac{\partial \hat{\psi}_k(z)}{\partial z} &= -\left(\frac{\partial F(\hat{\psi}_k, z, \mu)}{\partial \hat{\psi}_k}\right)^{-1}\left(\frac{\partial F(\hat{\psi}_k, z, \mu)}{\partial z}\right) \\ &= -\left(\frac{\partial^2 V(h_k(\hat{\psi}_k), z, \mu)}{\partial \hat{\psi}_k^2}\right)^{-1}\left(\frac{\partial^2 V(h_k(\hat{\psi}_k), z, \mu)}{\partial \hat{\psi}_k \partial z}\right) \end{aligned} \tag{18-18}$$

for any $\hat{\psi}_k \in \mathbb{R}^{n_\psi}$ and $z \in \mathbb{R}_z$ with $\dim(\mathbb{R}^N \setminus \mathbb{R}_z) < N$, which concludes the proof. □

References

Abel, J. P. (1993). A bound on mean-square-estimate error. *IEEE Trans. Information Theory,* vol. 39, no. 5, pp. 1675–1680.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automatic Control,* vol. AC-19, pp. 716–723.

Albertos, P., R. Sanchis, and A. Sala (1999). Output prediction under scarce data operation: control applications. *Automatica,* vol. 35, no. 10, pp. 1671–1681.

Anderson, B. D. O. and M. Deistler (1984). Identifiability in dynamic errors-in-variables models. *J. Time Series Analysis,* vol. 5, pp. 1–13.

Anderson, E., Z. Bai, C. Bishof, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammastry, A. McKenney, S. Ostrouchov, and D. Sorensen (1992). *LAPACK User's Guide.* SIAM Press, Philadelphia.

Anderson, T. W. (1958). *An Introduction to Multivariate Statistical Analysis.* Wiley, New York.

Åström, K. J., P. Hagander, and J. Sternby (1984). Zeros of sampled systems. *Automatica,* vol. 20, pp. 31–38.

Bai, Z., and J. Demmel (1993). Computing the generalized singular value decomposition. *SIAM J. Sci. Stat. Comput.,* vol. 14, no. 6, pp. 1464–1486.

Balabanian, N., and T. A. Bickart (1969). *Electrical Network Theory.* Wiley, New York.

Barker, H. A., and M. Zhuang (1997). Design of pseudo-random perturbation signals for frequency-domain identification of nonlinear systems. *Preprints of SYSID'97, 11th IFAC symposium on system identification,* Kitakyushu, Japan, pp. 1635–1640.

Beck, J. V., and K. J. Arnold (1977). *Parameter Estimation in Engineering and Science.* Wiley, New York.

Bendat, J. S. (1998). *Nonlinear Systems Techniques and Applications.* Wiley, New York.

Bendat, J. S., and A. G. Piersol (1980). *Engineering Applications of Correlations and Spectral Analysis.* Wiley, New York.

Ben-Israel, A., and T. N. E. Greville (1974). *Generalized Inverses: Theory and Applications.* Wiley, London.

Beya, K., R. Pintelon, J. Schoukens, P. Lataire, B. Mpanda-Mabwe, and M. Delhaye (1994). Identification of synchronous machines parameters using broadband excitation. *IEEE Trans. Energy Conversion,* vol. TEC-9, no. 2, pp. 270–280.

System Identification: A Frequency Domain Approach. Rik Pintelon and Johan Schoukens
Copyright © 2001 Institute of Electrical and Electronics Engineers. ISBN: 0-780-36000-1

Billings, S. A. (1980). Identification of nonlinear systems: A survey. *Proc. Inst. Elec. Eng.*, vol. 127, pt. D, pp. 272–285.

Billings, S. A., and S. Y. Fakhour (1982). Identification of systems containing linear dynamic and static nonlinear elements. *Automatica*, vol. 18, no. 1, pp. 15–26.

Billingsley, P. (1995). *Probability and Measure.* Wiley, New York.

Bohlin, T. (1971). On the problem of ambiguities in maximum likelihood identification. *Automatica,* vol. 7, pp. 199–210.

Box, G. E. P., and G. M. Jenkins (1976). *Time Series Analysis: Forecasting and Control.* Holden-Day, Oakland, CA.

Boyd, S., Y. S. Tang, and L. O. Chua (1983). Measuring Volterra kernels. *IEEE Trans. Circuits Systems,* vol. CAS-30, no. 9, pp. 648–651.

Brewer, J. W. (1978). Kronecker products and matrix calculus in system theory. *IEEE Trans. Circuits Systems,* vol. 25, pp. 772–781.

Brigham, E. O. (1974). *The Fast Fourier Transform.* Prentice-Hall, Englewood Cliffs, NJ.

Brillinger, D. R. (1981). *Time Series: Data Analysis and Theory.* McGraw-Hill, New York.

Broersen, P. M. T. (1995). A comparison of transfer function estimators. *IEEE Trans. Instrumentation Measurement,* vol. 44, pp. 657–661.

Brown, D., G. Carbon, and K. Ramsey (1977). Survey of excitation techniques applicable to the testing of automotive structures. *Int. Automotive Eng. Congress and Exposition,* Cobo Hall, Detroit, MI.

Cadzow, J. A. (1990). Signal processing via least squares error modeling. *IEEE ASSP Magazine*, pp. 12–31.

Cadzow, J. A., and O. M. Solomon (1987). Linear modeling and the coherence function. *IEEE Trans. Acoustics, Speech and Signal Processing,* vol. ASSP-35, no. 1, pp. 19–28.

Caines, P. E. (1988). *Linear Stochastic Systems.* Wiley, New York.

Chow, Y. S., and H. Teicher (1988). *Probability Theory: Independence, Interchangeability, Martingales* (2nd ed.). Springer-Verlag, New York.

Chua, L. O., and C.-Y. Ng (1979). Frequency domain analysis of nonlinear systems: General theory. *Electron. Circuits Syst.*, vol. 3, no. 4, pp. 165–185.

Crochiere, R. E., and L. R. Rabiner (1983). *Multirate Digital Signal Processing.* Prentice-Hall, Englewood Cliffs, NJ.

Delbaen, F. (1990). Optimizing the determinant of a positive definite matrix. *Bull. Soc. Math. Belg. Tijdschr. Belg. Wisk. Gen.*, vol. 42, no. 3, ser. B, pp. 333–346.

Delchamps, D. F. (1985). Global structure of families of multivariable linear systems with an application to identification. *Mathematical Systems Theory,* vol. 18, pp. 329–380.

Delchamps, D. F., and C. I. Byrnes (1982). Critical point behavior of objective functions defined on spaces of multivariable systems. *Proceedings of the 21st IEEE Conference on Decision and Control,* Orlando (Florida), vol. 2, pp. 937–943.

De Moor, B., M. Gevers, and G. Goodwin (1994). L_2-underbiased, and L_2-unbiased estimation of transfer functions. *Automatica,* vol. 30, no. 5, pp. 893–898.

Evans, C. (1998). *Identification of linear and nonlinear systems using multisine test signals.* Ph.D. dissertation, Dept. Electronics and IT, Univ. Glamorgan, Wales.

Evans, D. C., D. Rees, and D. L. Jones (1994). Identifying linear models of systems suffering nonlinear distortions. *Proceedings IEE Control'94,* Coventry, UK, pp. 288–296.

Ewins, D. J. (1991). *Modal Testing: Theory and Practice.* Wiley, New York.

Eykhoff, P. (1974). *System Identification, Parameter and State Estimation.* Wiley, New York.

Fedorov, V. V. (1972). *Theory of Optimal Experiments.* Academic Press, New York.

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications.* Wiley, London.

Fletcher, R. (1991). *Practical Methods of Optimization* (2nd ed.). Wiley, New York.

Forssell, U., and L. Ljung (1999). Closed-loop identification revisited. *Automatica,* vol. 35, no. 7, pp. 1215–1241.

Forssell, U., and L. Ljung (2000a). Identification of unstable systems using output error and Box-Jenkins model structures. *IEEE Trans. Automatic Control,* vol. 45, no. 1, pp. 137–141.

Forssell, U., and L. Ljung (2000b). A projection method for closed-loop identification. *IEEE Trans. Automatic Control,* vol. 45, no. 11, pp. 2101–2105.

Forsythe, E. G. (1957). Generation and use of orthogonal polynomials for data-fitting with a digital computer. *J. Soc. Indust. Appl. Math.,* vol. 5, no. 3, pp. 74–88.

Forsythe, E. G., and E. G. Straus (1955). On best conditioned matrices. *Proc. Am. Math. Soc.,* vol. 6, pp. 340–345.

Fuller, W. A. (1987). *Measurement Error Models.* Wiley, New York.

Gantmacher, F. R. (1990). *The Theory of Matrices.* Chelsea Publishing Company, New York.

Giri, N. (1965). On the complex analogues of T^2 and R^2 tests. *Annals of Mathematical Statistics,* vol. 36, pp. 664–670.

Godfrey, K. R. (1969). The theory of the correlation method of dynamic analysis and its application to industrial processes and nuclear power plant. *Measurement and Control,* vol. 2, pp. T65–T72.

Godfrey, K. R. (1980). Correlation methods. *Automatica,* vol. 16, pp. 527–534.

Godfrey, K. R., editor (1993a). *Perturbation Signals for System Identification.* Prentice-Hall, London.

Godfrey, K. R. (1993b). Introduction to perturbation signals for time-domain system identification. In K. R. Godfrey, editor, *Perturbation Signals for System Identification.* Prentice-Hall, Hemel Hempstead, England, pp. 1–59.

Golub, G. H., and C. F. Van Loan (1996). *Matrix Computations* (3rd ed.). John Hopkins University Press, Baltimore.

Goodman, N. R. (1963). Statistical analysis based upon a certain multivariate complex Gaussian distribution (an introduction). *Ann. Math. Statist.,* vol. 34, pp. 152–177.

Goodwin, G. C., and G. J. Adams (1994). Multi-rate techniques in non-zero-order hold identification. *Preprints of the 10th IFAC Symposium on System Identification,* Copenhagen, Denmark, vol. 3, pp. 125–130.

Goodwin, G. C., and R. L. Payne (1977). *Dynamic System Identification. Experimental Design and Data Analysis.* Academic Press, New York.

Gorman, J. D., and A. O. Hero (1990). Lower bounds for parametric estimation with constraints. *IEEE Trans. Inform. Theory,* vol. IT-26, no. 6, pp. 1285–1301.

Gradshteyn, L. S., and I. M. Ryzhik (1980). *Table of Integrals, Series, and Products* (corrected and enlarged edition). Academic Press, New York.

Guidorzi, R. (1975). Canonical structures in the identification of multivariable systems. *Automatica,* vol. 11, pp. 361–374.

Guillaume, P., I. Kollár, and R. Pintelon (1996a). Statistical analysis of nonparametric transfer function estimates. *IEEE Trans. Instrumentation and Measurement,* vol. IM-45, no. 2, pp. 594–600.

Guillaume, P., and R. Pintelon (1996). A Gauss-Newton-like optimization algorithm for "weighted" nonlinear least-squares problems. *IEEE Trans. Signal Processing,* vol. SP-44, no. 9, pp. 2222–2228.

Guillaume, P., R. Pintelon, and J. Schoukens (1992a). Parametric identification of two-port models in the frequency domain. *IEEE Trans. Instrumentation and Measurement,* vol. IM-41, no. 2, pp. 233–239.

Guillaume, P., R. Pintelon, and J. Schoukens (1992b). Non-parametric frequency response function estimators based on nonlinear averaging techniques. *IEEE Trans. Instrumentation and Measurement,* vol. IM-41, no. 6, pp. 739–746.

Guillaume, P., R. Pintelon, and J. Schoukens (1995). Robust parametric transfer function estimation using complex logarithmic frequency response data. *IEEE Trans. Automatic Control,* vol. AC-40, no. 7, pp. 1180–1190.

Guillaume, P., R. Pintelon, and J. Schoukens (1996b). Accurate estimation of multivariable frequency response functions. *Proceedings of the 13th IFAC Triennial World Conference,* San Francisco, pp. 423–428.

Guillaume, P., R. Pintelon, and J. Schoukens (1996c). Parametric identification of multivariable systems in the frequency domain—a survey, *Proceedings ISMA21—Noise and Vibration Engineering,* Leuven (Belgium), vol. II, pp. 1069–1082.

Guillaume, P., J. Schoukens, and R. Pintelon (1989). Sensitivity of roots to errors in the coefficients of polynomials obtained by frequency domain estimation methods. *IEEE Trans. Instumentation and Measurement,* vol. IM-38, no. 6, pp. 1050–1056.

Guillaume, P., J. Schoukens, R. Pintelon, and I. Kollár (1991). Crest-factor minimization using nonlinear chebyshev approximation methods. *IEEE Trans. Instrumentation and Measurement,* vol. IM-40, no. 6, pp. 982–989.

Gustafsson, F., and J. Schoukens (1998). Utilizing periodic excitation in prediction error based system identification. *Proceedings of the 37th IEEE Conference on Decision and Control,* Tampa, FL, pp. 3926–3931.

Haber, R. (1985). Nonlinearity tests for dynamic processes. *7th IFAC/IFORS Symposium on Identification and System Parameter Estimation,* York, UK, pp. 409–414.

Halvorsen, W. G., and D. L. Brown (1977). Impulse technique for structural frequency response testing. *Sound and Vibration,* vol. 11, pp. 8–21.

Hannan, E. J. (1980). The estimation of the order of an ARMA process. *Annals of Statistics*, vol. 8, no. 5, pp. 1071–1081.

Hannan, E. J., and M. Deistler (1988). *Linear Systems.* Wiley, New York.

Harris, F. (1978). On the use of windows for harmonic analysis with discrete Fourier transform. *Proceedings of the IEEE,* vol. 66, pp. 51–83.

Hazewinkel, M. (1977). Moduli and canonical forms for linear dynamical systems II: The topological case. *Mathematical Systems Theory,* vol. 10, pp. 363–385.

Henrici, P. (1974). *Applied and Computational Complex Analysis.* Wiley, New York, vol. 1.

Herlufsen, H. (1984). Dual channel FFT analysis (part I). *Tech. Rev.,* Brüel & Kjær, no. 1, Nærum, Denmark.

Heylen, W., S. Lammens, and P. Sas (1997). *Modal Analysis Theory and Testing.* Society for Experimental Mechanics, Bethel (USA).

Huber, P. J. (1981). *Robust Statistics.* Wiley, New York.

Isaksson, A. J. (1993). Identification of ARX-models subject to missing data. *IEEE Trans. Automatic Control,* vol. 38, no. 5, pp. 813–819.

Jazwinski, A. H. (1970). *Stochastic Processes and Filtering Theory.* Academic Press, London.

Kabaila, P. (1983). Parameter values of ARMA models minimizing the one-step-ahead prediction error when the true system is not in the model set. *J. Appl. Prob.,* vol. 20, no. 2, pp. 405–408.

Kahane, J. P. (1980). Sur les polynomes à coefficients unimodulaires. *Bull. London Math. Soc.,* no. 12, pp. 321–342.

Kahane, J. P. (1985). *Some Random Series of Functions.* University Press, Cambridge (UK).

Kailath, T. (1980). *Linear Systems.* Prentice-Hall, Englewood Cliffs, NJ.

Kalman, R. E. (1958). Design of a self-optimizing control system. *ASME Trans.*, vol. 80, no. 2, pp. 468–478.

Kaplan, W. (1993). *Advanced Calculus.* Addison-Wesley, Reading, MA.

Kashyap, R. L. (1980). Inconsistency of the AIC rule for estimating the order of autoregressive models. *IEEE Trans. Automatic Control,* vol. AC-25, no. 5, pp. 996–998.

Kay, S. M. (1988). *Modern Spectral Estimation: Theory and Application.* Prentice-Hall, Englewood Cliffs, NJ.

Kendall, M., and A. Stuart (1979). *Inference and Relationship,* vol. 2 of *The Advanced Theory of Statistics* (4th ed.). Charles Griffin, London.

Kollár, I. (1994). *Frequency-Domain System Identification Toolbox for Use with Matlab.* The Mathworks, Natick, MA.

Kollár, I., R. Pintelon, Y. Rolain, and J. Schoukens (1991). Correspondence: Another step towards an ideal data acquisition channel. *IEEE Trans. Instumentation and Measurement,* vol. IM-40, no. 3, pp. 659–660.

Kumaresan, R., C.S. Ramalingam, and D. Van Ormondt (1990). Estimating the parameters of NMR signals by transforming to the frequency domain. *Journal of Magnetic Resonance,* vol. 89, pp. 562–567.

Kwakernaak, H., and R. Sivan (1991). *Modern Signals and Systems.* Prentice-Hall, London.

Lancaster, P., and M. Tismenetsky (1985). *The Theory of Matrices.* Academic Press, Orlando, FL.

Lee, C. W. (1993). *Vibration Analysis of Rotors.* Kluwer Academic, Dordrecht.

Lee, C. W., and C.Y. Joh (1994). Development of the use of directional frequency response functions for the diagnosis of anisotropy and asymmetry in rotating machinery: Theory. *Mechanical Systems and Signal Processing,* vol. 8, no. 6, pp. 665–678.

Leonov, V. P., and A. N. Shiryaev (1959). On a method of calculation of semi-invariants. *Theory of Probability and its Applications,* vol. IV, no. 3, pp. 319–329.

Levi, E. C. (1959). Complex curve fitting. *IEEE Transactions Automatic Control,* vol. AC-4, pp. 37–43.

Liang, G., D. M. Wilkes, and J. A. Cadzow (1993). ARMA model order estimation based on the eigenvalues of the covariance matrix. *IEEE Trans. Signal Processing,* vol. SP-41, pp. 3003–3009.

Little, R. J. A., and D. B. Rubin (1987). *Statistical Analysis with Missing Data.* Wiley, New York.

Ljung, L. (1985). Asymptotic variance expressions for identified black-box transfer function models. *IEEE Trans. Automatic Control,* vol. AC-30, no. 9, pp. 834–844.

Ljung, L. (1993). Some results on identifying linear systems using frequency domain data. *Proc. 32nd Conf. Decis. Contr.,* San Antonio, TX, pp. 3534–3538.

Ljung, L. (1995). *System Identification Toolbox User's Guide.* The MathWorks, Natick, MA.

Ljung, L. (1999). *System Identification: Theory for the User* (2nd ed.). Prentice-Hall, Upper Saddle River, NJ.

Ljung, L., and T. Söderström (1983). *Theory and Practice of Recursive Identification.* MIT Press, Cambridge, MA

Lowen, S. B., and M. C. Teich (1990). Power-law shot noise. *IEEE Trans. Inform. Theory,* vol. 36, no. 6, pp. 1302–1318.

Lukacs, E. (1975). *Stochastic Convergence.* Academic Press, New York.

Mannetje: see 't Mannetje.

Mathai, A. M., and S. B. Provost (1992). *Quadratic Forms in Random Variables.* Marcel Dekker, New York.

McCormack, A. S., J. O. Flower, and K. R. Godfrey (1994a). The suppression of drift and transient effects for frequency-domain identification. *IEEE Trans. Instrumentation and Measurement,* vol. IM-43, no. 2, pp. 232–237.

McCormack, A. S., K. R. Godfrey, and J. O. Flower (1994b). The detection of and compensation for nonlinear effects for frequency-domain identification. *Proceedings IEE Control '94,* Coventry, UK, pp. 297-302.

McCormack, A. S., K. R. Godfrey, and J. O. Flower, (1995). Design of multilevel multiharmonic signals for system identification. *IEE Proceedings, Control Theory and Applications,* vol. 142, no. 3, pp. 247–252.

McKelvey, T., H. Akçay, and L. Ljung (1996). Subspace-based multivariable system identification from frequency response data, *IEEE Trans. Automatic Control,* vol. 41, no. 7, pp. 960–979.

McKelvey, T., and A. Helmersson (1997). System identification using an over-parametrized model class—Improving the optimization algorithm. *Proceedings of the 36th Conference on Decision and Control,* San Diego, CA, pp. 2984–2989.

Mendel, J. M. (1991). Tutorial on higher-order statistics (spectra) in signal processing and system theory: Theoretical results and some applications. *Proc. IEEE,* vol. 79, pp. 278–305.

Middleton, R. H., and G. C. Goodwin (1990). *Digital Control and Estimation.* Prentice-Hall, London.

Natke, H. G., J.-N. Juang, and W. Gawronski (1988). A brief review on the identification of nonlinear mechanical systems. *Proc. of the 6th International Modal Analysis Conference,* Kissimee, FL, pp. 1569–1574.

Nikias, C. L., and J.M. Mendel (1993). Signal processing with higher-order spectra. *IEEE Signal Processing Magazine,* July 1993, pp. 10–36.

Nikias, C. L., and A. P. Petropulu (1993). *Higher-Order Spectra Analysis.* Prentice-Hall, Englewood Cliffs, NJ.

Ninness, B. M., and G. C. Goodwin (1991). The relationship between discrete time and continuous time linear estimation. In N. K. Sinha and G. P. Rao, editors, *Identification of Continuous-Time Systems.* Kluwer Academic Publishers, Dordrecht, pp. 79–122.

Norton, J. P. (1986). *An Introduction to Identification.* Academic Press, London.

Oppenheim, A. V., and R. W. Schafer (1975). *Digital Signal Processing.* Prentice-Hall, New York.

Oppenheim, A. V., A. S Willsky, and S. H. Nawab (1997). *Signals and Systems.* Prentice-Hall, London.

Orey, S. (1958). A central limit theorem for m-dependent random variables. *Duke Math. J.,* vol. 25, pp. 543–546.

Paehlike, K. D., and H. Rake (1979). Binary multifrequency signals-synthesis and application. *Proc. 5th IFAC Symp.,* Darmstadt, FRG, pp. 589–597.

Paige, C. C. (1986). Computing the generalized singular value decomposition. *SIAM J. Sci. Stat. Comput.,* vol. 7, no. 4, pp. 1126–1146.

Papoulis, A. (1981). *Probability, Random Variables, and Stochastic Processes.* McGraw-Hill, New York.

Peeters, F., R. Pintelon, J. Schoukens, and Y. Rolain (2000). Parametric identification of rotor-bearing systems in the frequency domain. *Proceedings of the 18th International Modal Analysis Conference,* San Antonio, TX, pp. 1355–1361.

Peirlinckx, L., P. Guillaume, and R. Pintelon (1996). Accurate and fast estimation of the Fourier coefficients of periodic signals disturbed by a trend. *IEEE Trans. Instrumentation and Measurement,* vol. IM-45, no. 1, pp. 5–11.

Picinbono, B. (1993). *Random Signals and Systems.* Prentice-Hall, Englewood Cliffs, NJ.

Pintelon, R. (1990). Phase correction of linear time invariant systems with digital all-pas filters, *IEEE Trans. Instumentation and Measurement,* vol. IM-39, no. 2, pp. 324–330.

Pintelon, R. (1991). Comments on "Design of IIR filters in the complex domain." *IEEE Trans. Signal Processing,* vol. SP-39, no. 6, pp. 1454–1455.

Pintelon, R., P. Guillaume, Y. Rolain, J. Schoukens, and H. Van hamme (1994). Parametric identification of transfer functions in the frequency domain—A survey. *IEEE Trans. Automatic Control,* vol. AC-39, no. 11, pp. 2245–2260.

Pintelon, R., P. Guillaume, Y. Rolain, and F. Verbeyst (1992). Identification of linear systems captured in a feedback loop. *IEEE Trans. Instrumentation and Measurement,* vol. IM-41, no. 6, pp. 747–754.

Pintelon, R., P. Guillaume, and J. Schoukens (1996a). Measurement of noise (cross-) power spectra for frequency-domain system identification purposes: Large-sample results. *IEEE Trans. Instrumentation and Measurement,* vol. IM-45, no. 1, pp. 12–21.

Pintelon, R., P. Guillaume, G. Vandersteen, and Y. Rolain (1998). Analyses, development and applications of TLS algorithms in frequency-domain system identification. *SIAM J. Matrix Anal. Appl.,* vol. 19, no. 4, pp. 983–1004.

Pintelon, R., Y. Rolain, M. Vanden Bossche, and J. Schoukens (1990). Towards an ideal data acquisition channel. *IEEE Trans. Instumentation and Measurement,* vol. IM-39, no. 1, pp. 116–120.

Pintelon, R., and J. Schoukens (1990a). Robust identification of transfer functions in the s- and z-domains. *IEEE Trans. Instumentation and Measurement,* vol. IM-39, no. 4, pp. 565–573.

Pintelon, R., and J. Schoukens (1990b). Real-time integration and differentiation of analog signals by means of digital filtering. *IEEE Trans. Instrumentation and measurement,* vol. IM-39, no. 6, pp. 923–927.

Pintelon, R., and J. Schoukens (1996). An improved sine-wave fitting procedure for characterizing data acquisition channels. *IEEE Trans. Instrumentation and Measurement,* vol. IM-45, no. 2, pp. 588–593.

Pintelon, R., and J. Schoukens (1997a). Frequency-domain identification of linear time-invariant systems under nonstandard conditions. *IEEE Trans. Instrumentation and Measurement,* vol. IM-46, no. 1, pp. 65–71.

Pintelon, R., and J. Schoukens (1997b). Identification of continuous-time systems using arbitrary signals. *Automatica,* vol. 33, no. 5, pp. 991–994.

Pintelon, R., and J. Schoukens (1999a). Time series analysis in the frequency domain. *IEEE Trans. Signal Processing,* vol. 47, no. 1, pp. 206–210.

Pintelon, R., and J. Schoukens (1999b). Identification of continuous-time systems with missing data. *IEEE Trans. Instrumentation and Measurement,* vol. IM-48, no. 3, pp. 736–740.

Pintelon, R., and J. Schoukens (2000). Frequency domain system identification with missing data. *IEEE Trans. Automatic Control,* vol. AC-45, no. 5, pp. 364–369.

Pintelon, R., J. Schoukens, and P. Guillaume (1989). Parametric frequency domain modeling in modal analysis. *Mechanical Systems and Signal Processing,* vol. 3, no. 4, pp. 389–403.

Pintelon, R., J. Schoukens, T. McKelvey, and Y. Rolain (1996b). Minimum variance bounds for overparametrized models. *IEEE Trans. Automatic Control,* vol. AC-41, no. 5, pp. 719–720.

Pintelon, R., J. Schoukens, and J. Renneboog (1988). The geometric mean of power (amplitude) spectra has a much smaller bias than the classical arithmetic (rms) averaging. *IEEE Trans. Instumentation and Measurement,* vol. IM-37, no. 2, pp. 213–218.

Pintelon, R., J. Schoukens, and Y. Rolain (2000). Box-Jenkins continuous-time modeling, *Automatica,* vol. 36, no. 7, pp. 983–991.

Pintelon, R., and L. Van Biesen (1990). Identification of transfer functions with time delay and its application to cable fault location. *IEEE Trans. Instrumentation and Measurement,* vol. IM-39, no. 3, pp. 479–484.

Pintelon, R., J. Schoukens, and G. Vandersteen (1997a). Model selection through a statistical analysis of the global minimum of a weighted nonlinear least squares cost function. *IEEE Trans. Signal Processing,* vol. 45, no. 3, pp. 686–693.

Pintelon, R., J. Schoukens, and G. Vandersteen (1997b). Frequency domain system identification using arbitrary signals. *IEEE Trans. Automatic Control,* vol. AC-42, no. 12, pp. 1717–1720.

Pintelon, R., J. Schoukens, G. Vandersteen, and Y. Rolain (1999). Identification of invariants of (over)parametrized models: finite sample results. *IEEE Trans. Automatic Control,* vol. AC-44, no. 5, pp. 1073–1077.

Pyati, V. P. (1992). An exact expression for the noise voltage across a resistor shunted by a capacitor. *IEEE Trans. Circuits Systems-I,* vol. 39, no. 12, pp. 1027–1029.

Rabiner, L. R., and B. Gold (1975). *Theory and Application of Digital Signal Processing.* Prentice-Hall, New York.

Ralston, A., and P. Rabinowitz (1984). *A First Course in Numerical Analysis.* McGraw-Hill, New York.

Richardson, M. H., and D. L. Formenti (1982). Parameter estimation from frequency response measurements using rational fraction polynomials. *Proc. First Int. Modal Analysis Conf.,* Orlando, FL, vol. 1, pp. 167–181.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica,* vol. 14, pp. 465–471.

Rizzi, P. A. (1988). *Microwave Engineering.* Prentice-Hall, London.

Rolain, Y., and R. Pintelon (1999). Generating robust starting values for frequency-domain transfer function estimation. *Automatica,* vol. 35, pp. 965–973.

Rolain, Y., R. Pintelon, K. Q. Xu, and H. Vold (1995). Best conditioned parametric identification of transfer function models in the frequency domain. *IEEE Trans. Automatic Control,* vol. AC-40, no. 11, pp. 1954–1960.

Rolain, Y., and J. Schoukens (1990). Design and implementation of a fast logarithmic stepped sine for a fixed sample rate digital network analyzer. *IEEE Trans. Instrumentation and Measurement,* vol. IM-39, no. 1, pp. 151–156.

Rolain Y., J. Schoukens, and R. Pintelon (1997). Order estimation for linear time-invariant systems using frequency domain identification methods. *IEEE Trans. Automatic Control,* vol. 42, no. 10, pp. 1408–1417.

Rolain, Y., J. Schoukens, and G. Vandersteen (1998). Signal reconstruction for non-equidistant finite length sample sets: A "KIS" approach. *IEEE Transactions Instrumentation and Measurement,* vol. 47, no. 5, pp. 1046–1052.

Rosén, B. (1967). On the central limit theorem for sums of dependent random variables, *Z. Wahrsch. verw. Geb.,* vol. 7, pp. 48–82.

Sanathanan, C. K., and J. Koerner (1963). Transfer function synthesis as a ratio of two complex polynomials. *IEEE Trans. Automatic Control,* vol. AC-8, pp. 56–58.

Schetzen, M. (1980). *The Volterra and Wiener Theories of Nonlinear Systems.* Wiley, New York.

Schoukens, J. (1990). Modeling of continuous time systems using a discrete time representation. *Automatica,* vol. 26, no. 3, pp. 579–583.

Schoukens, J., T. Dobrowiecki, and R. Pintelon (1998a). Identification of linear systems in the presence of nonlinear distortions. A frequency domain approach. *IEEE Trans. Automatic Control,* vol. 43, no. 2, pp. 176–190.

Schoukens J., P. Guillaume, and R. Pintelon (1993). Design of broadband excitation signals. In K. R. Godfrey, editor, *Perturbation Signals for System Identification.* Prentice-Hall, Hemel Hempstead, pp. 126–160.

Schoukens, J., P. Guillaume, and R. Pintelon (1995). Generating piecewise-constant excitations with an arbitrary power spectrum. *IEE Proc. Control Theory Appl.,* vol. 142, no. 3, pp. 241–246.

Schoukens, J., F. Louage, and Y. Rolain (1996a). Study of the influence of clock instabilities in synchronized data acquisition systems. *IEEE Trans. Instrumentation and Measurement,* vol. IM-45, no. 2, pp. 601–604.

Schoukens, J., and R. Pintelon (1990). Measurement of frequency response functions in noisy environments. *IEEE Trans. Instrumentation and Measurement,* vol. IM-39, no. 6, pp. 905–909.

Schoukens, J., and R. Pintelon (1991). *Identification of Linear Systems: A Practical Guideline to Accurate Modeling.* Pergamon, Oxford.

Schoukens, J., R. Pintelon, and J. Renneboog (1988a). A maximum likelihood estimator for linear and nonlinear systems—A practical application of estimation techniques in measurement problems. *IEEE Trans. Instrumentation and Measurement,* vol. IM-37, no. 1, pp. 10–17.

Schoukens, J., R. Pintelon, and Y. Rolain (1997a). Maximum likelihood estimation of errors-in-variables models using a sample covariance matrix obtained from small data sets. In S. Van Huffel, editor, *Recent Advances in Total Least Squares Techniques and Errors-in-Variables Modeling.* SIAM, Philadelphia, pp. 59–68.

Schoukens, J., R. Pintelon, and Y. Rolain (1999a). Study of conditional ML estimators in time and frequency domain system identification. *Automatica,* vol. 35, no. 1, pp. 91–100.

Schoukens, J., R. Pintelon, and Y. Rolain (2000). Broadband versus stepped sine FRF measurements. *IEEE Trans. on Instrumentation and Measurement,* vol. IM-49, no. 2, pp. 275–278.

Schoukens, J., R. Pintelon, E. Van der Ouderaa, and J. Renneboog (1988b). Survey of excitation signals for F.F.T. based signal analyzers. *IEEE Trans. Instrumentation and Measurement,* vol. IM-37, no. 3, pp. 342–351.

Schoukens, J., R. Pintelon, and H. Van hamme (1994). Identification of linear dynamic systems using piecewise constant excitations: Use, misuse and alternatives. *Automatica,* vol. 30, no. 7, pp. 1153–1169.

Schoukens, J., R. Pintelon, G. Vandersteen, and P. Guillaume (1997b). Frequency-domain system identification using non-parametric noise models estimated from a small number of data sets. *Automatica,* vol. 33, no. 6, pp. 1073–1086.

Schoukens, J., and J. Renneboog (1986). Modeling the noise influence on the Fourier coefficients after a discrete Fourier transform. *IEEE Trans. Instrumentation and Measurement,* vol. IM-35, no. 3, pp. 278–286.

Schoukens, J., Y. Rolain, and P. Guillaume (1996b). Design of narrow band, high resolution multisines. *IEEE Trans, Instrumentation and Measurement,* vol. IM-45, no. 3, pp. 750–753.

Schoukens, J., Y. Rolain, F. Gustafsson, and R. Pintelon (1998b). Fast calculation of least-squares estimates for system identification. *Proceedings of the 37th IEEE Conference on Decision and Control,* Tampa, FL, pp. 3408–3410.

Schoukens, J., Y. Rolain, and R. Pintelon (1998c). Improved frequency response function measurements using random noise excitations. *IEEE Trans. Instrumentation and Measurement,* vol. IM-47, no. 1, pp. 322–326.

Schoukens, J., G. Vandersteen, R. Pintelon, and P. Guillaume (1999b). Frequency domain identification of linear systems using arbitrary excitations and a nonparametric noise model. *IEEE Transaction Automatic Control,* vol. 44, no. 2, pp. 343–347.

Schroeder, M. R. (1970). Synthesis of low peak factor signals and binary sequences with low autocorrelation. *IEEE Trans. Inform. Theory,* vol. IT-16, pp. 85–89.

Selby, S. M. (1973). *Standard Mathematical Tables.* The Chemical Rubber Company, Cleveland, OH.

Shapiro, A. (1986). Asymptotic theory of overparametrized structural models. *J. Am. Statistical Assoc.,* vol. 81, no. 393, pp. 142–149.

Sidman, M. D., F. E. DeAngelis, and G. C. Verghese (1991). Parametric system identification on logarithmic frequency response data. *IEEE Trans. Automatic Control,* vol. AC-36, no. 9, pp. 1065–1070.

Sinha, N. K., and G. P. Rao, editors (1991). *Identification of Continuous-Time Systems: Methodology and Computer Implementation.* Kluwer, Dordrecht.

Söderström, T. (1974). Convergence properties of the generalized least squares identification method. *Automatica,* vol. 10, pp. 617–626.

Söderström, T., and B. Carlsson (2000). Performance evaluation of methods for identifying continuous-time autoregressive processes. *Automatica,* vol. 36, pp. 53–59.

Söderström, T., H. Fan, B. Carlsson, and S. Bigi (1997a). Least squares parameter estimation of continuous-time ARX models from discrete-time data. *IEEE Transactions Automatic Control,* vol. AC-42, pp. 659–673.

Söderström, T., H. Fan, M. Mossberg, and B. Carlsson (1997b). Bias-compensation schemes for estimating continuous-time AR process parameters. *Proceedings of the 11th IFAC Symposium System Identification,* Kitakyushu, Japan, vol. 3, pp. 1337–1342.

Söderström, T., and P. Stoica (1981). Comparison of some instrumental variable methods—consistency and accuracy aspects. *Automatica,* vol. 17, pp. 101–115.

Söderström, T., and P. Stoica (1989). *System Identification.* Prentice-Hall, Englewood Cliffs, NJ, p. 256.

Sorenson, H. W. (1980). *Parameter Estimation: Principles and Problems.* Marcel Dekker, New York.

Souders, T. M., D. R. Flach, C. Hagwood, and G. L. Yang (1990). The effect of time jitter in sampling systems. *IEEE Trans. Instrumentation and Measurement,* vol. IM-39, pp. 80–85.

Spiegel, M. R. (1965). *Theory and Problems of Laplace Transforms.* McGraw-Hill, New York.

Steigliz, K. and L. E. McBride (1965). A technique for the identification of linear systems. *IEEE Trans. Automatic Control,* vol. AC-10, pp. 461–464.

Stoica, P., P. Eykhoff, P. Janssen, and T. Söderström (1986). Model-structure selection by cross-validation. *Int. J. Control,* vol. 43, no. 6, pp. 1841–1878.

Stoica, P., and R. L. Moses (1990). On biased estimators and the unbiased Cramér-Rao lower bound. *Signal Processing,* vol. 21, pp. 349–350.

Stout, W. F. (1974). *Almost Sure Convergence.* Academic Press, New York.

Stuart, A., and J. K. Ord (1987). *Distribution Theory,* vol. 1 of *Kendall's Advanced Theory of Statistics.* Charles Griffin, London.

Swevers, J., B. De Moor, and H. Van Brussel (1992). Stepped sine system identification, errors-in-variables and the quotient singular value decomposition. *Mech. Syst. Signal Processing,* vol. 6, no. 2, pp. 121–134.

Temes, G. C., and J. W. LaPatra (1977). *Circuit Synthesis and Design.* McGraw-Hill, New York.

't Mannetje, J. J. (1973). Transfer-function identification using a complex curve-fitting technique. *J. Mech. Eng. Sci.,* vol. 15, no. 5, pp. 339–345.

Tomlinson, G. R. (1987). Developments in the use of the Hilbert transform for detecting and quantifying non-linearity associated with frequency response functions. *Mechanical Systems and Signal Processing,* vol. 1, no. 2, pp. 151–171.

Torfs, D., R. Vuerinckx, J. Swevers, and J. Schoukens (1998). Comparison of two feedforward design methods aiming at accurate trajectory tracking of the end point of a flexible robot arm. *IEEE Transactions Control Systems Technology,* vol. 6, no. 1, pp. 2–14.

Van Barel, M., and A. Bultheel (1992). A parallel algorithm for discrete least squares rational approximation. *Numer. Math.,* vol. 63, pp. 99–121.

Van Barel, M., and A. Bultheel (1994). Discrete linearized least-squares rational approximation on the unit circle. *Journal of Computational and Applied Mathematics,* vol. 50, pp. 545–563.

Van Brussel, H. (1975). Comparative assessment of harmonic, random, swept sine and shock excitation methods for the identification of machine tool structures with rotating spindles. *Ann. CIRP,* pp. 291–296.

Van den Bos, A. (1974). *Estimation of Parameters of Linear Systems using Periodic Test Signals.* Doctoral thesis, T.U. Delft, The Netherlands, Delftse Universitaire Pers.

Van den Bos, A. (1985). Nonlinear least-absolute-values and minimax model fitting. *7th IFAC Symposium on Identification and System Parameter Estimation,* York, UK, pp. 173–177A.

Van den Bos, A. (1987). A new method for synthesis of low-peak-factor signals. *IEEE Trans. Acoustics, Speech and Signal Processing,* vol. ASSP-35, pp. 120–122.

Van den Bos, A. (1991). Identification of continuous-time systems using multiharmonic test signals. In N.K. Sinha and G.P. Rao, editors, *Identification of Continuous-Time Systems.* Kluwer Academic Publishers, Dordrecht, pp. 489–508.

Van den Bos, A., and R. G. Krol (1979). Synthesis of discrete-interval binary signals with specified Fourier amplitude spectra. *Int. J. Contr.,* vol. 30, no. 5, pp. 871–884.

Van den Bos, A., and J. H. Swarte (1993). Resolvability of the parameters of multiexponentials and other sum models, *IEEE Trans. Signal Processing,* vol. SP-41, pp. 313–322.

Vanden Bossche, M., J. Schoukens, and J. Renneboog (1986). Dynamic testing and diagnostics of A/D converters. *IEEE Trans. Circuits and Systems,* vol. CAS-33, no. 8, pp. 775–785.

Van den Hof, P. M. J., and R. J. P. Schrama (1995). Identification and control—Closed loop issues, *Automatica,* vol. 31, no. 12, pp. 1751–1770.

Van den Eijnde, E., and J. Schoukens (1991). On the design of optimal excitation signals. *Preprints of the 9th IFAC/IFORS Symposium on Identification and System Parameter Estimation,* Budapest, Hungary, pp. 827–832.

Van den Enden, A. W. M., G. C. Groendael, and E. Van de Zee (1977). An improved complex-curve fitting method. *Proc. Conf. Computer Aided Design of Electronic, Microwave Circuits and Syst.,* Hull, UK, pp. 53–58.

Van den Enden, A. W. M., and G. A. L. Leenknegt (1986). Design of optimal filters with arbitrary amplitude and phase requirements. In P. Young et al., editors, *Signal Processing III: Theories and Applications,* Elsevier Science, North Holland, pp. 183–186.

Van der Ouderaa, E., and J. Renneboog (1988). Logtone crest factors. *IEEE Trans. Instrumentation and Measurement,* vol. IM-37, pp. 656–657.

Van der Ouderaa, E., J. Schoukens, and J. Renneboog (1988a). Peak factor minimization using a time-frequency domain swapping algorithm. *IEEE Trans. Instrumentation and Measurement,* vol. IM-37, no. 1, pp. 145–147.

Van der Ouderaa, E., J. Schoukens, and J. Renneboog (1988b). Peak factor minimization of input and output signals of linear systems. *IEEE Trans. Instrumentation and Measurement,* vol. IM-37, no. 2, pp. 207–212.

Van hamme, H., R. Pintelon, and J. Schoukens (1991). Discrete-time modeling and identification of continuous time systems: a general framework. In M.K. Sinha and G. P. Rao, editors, *Identification of Continuous-Time Systems.* Kluwer Academic Publishers, Dordrecht, pp. 17–77.

Van Huffel, S., and J. Vandewalle (1991). *The Total Least Squares Problem: Computational Aspects and Analysis.* Frontiers in Applied Mathematics. SIAM, Philadelphia.

Van Overbeek, A. J. M., and L. Ljung (1982). On-line structure selection for multivariable state-space models. *Automatica,* vol. 18, no. 5, pp. 529–543.

Van Overschee, P., and B. DeMoor (1994). N4SID: subspace algorithms for the identification of combined deterministic-stochastic systems. *Automatica,* vol. 30, no. 1, pp. 75–93.

Van Overschee, P., and B. De Moor (1996a). Continuous-time frequency domain subspace system identification, *Signal Processing,* vol. 52, no. 2, pp. 179–194.

Van Overschee, P., and B. De Moor (1996b). *Subspace Identification of Linear Systems: Theory, Implementation, Applications.* Kluwer Academic Publishers, Dordrecht.

Vandersteen, G., Y. Rolain, J. Schoukens, and R. Pintelon (1996a). On the use of system identification for accurate parametric modelling of non-linear systems using noisy measurements. *IEEE Trans. Instrumentation and Measurement,* vol. IM-45, no. 2, pp. 605–609.

Vandersteen, G., H. Van hamme, and R. Pintelon (1996b). General framework for asymptotic properties of generalized weighted nonlinear least-squares estimators with deterministic and stochastic weighting. *IEEE Trans. Automatic Control,* vol. AC-41, no. 10, pp. 1501–1507.

Vanhoenacker, K., and J. Schoukens (1999). Frequency response function measurements in the presence of nonlinear distortions. *WISP'99, IEEE International Workshop on Intelligent Signal Processing,* Budapest, Hungary, pp. 87–92.

Verbeeck, J., R. Pintelon, and P. Guillaume (1999a). Determination of synchronous machine parameters using network synthesis techniques. *IEEE Trans. Energy Conversion,* vol. EC-14, no. 3, pp. 310–314.

Verbeeck, J., R. Pintelon, and P. Lataire (1999b). Identification of synchronous machine parameters using a multiple input multiple output approach. *IEEE Trans. Energy Conversion,* vol. EC-14, no. 4, pp. 909–917.

Verhaegen, M. (1994). Identification of the deterministic part of MIMO state space models given in innovations form from input-output data. *Automatica,* vol. 30, no. 1, pp. 61–74.

Verschueren, A., Y. Rolain, R.Vuerinckx, and G. Vandersteen (1998). Identifying S-parameter models in the Laplace domain for high frequency multiple port linear networks. *IEEE-MTT-S International Microwave Symposium Digest,* Baltimore, pp. 25–28.

Viberg, M., B. Wahlberg, and B. Ottersten (1997). Analysis of state space system identification methods based on instrumental variables and subspace fitting. *Automatica,* vol. 33, no. 9, pp. 1603–1616.

Vuerinckx, R., Y. Rolain, J. Schoukens, and R. Pintelon (1996). Design of stable IIR filters in the complex domain by automatic delay selection. *IEEE Trans. Signal Processing,* vol. SP-44, no. 9, pp. 2339–2344.

Vuerinckx, R., R. Pintelon, J. Schoukens, and Y. Rolain (1998). Obtaining accurate confidence regions for the estimated zeros and poles in system identification problems. *Proceedings of the 37th IEEE Conference on Decision and Control,* Tampa, FL, pp. 4464–4469.

Walter, E., and L. Pronzato (1997). *Identification of Parametric Models from Experimental Data.* Springer, Paris.

Wang, J. C. (1987). Realizations of generalized Warburg impedance with RC ladder networks and transmission lines. *J. Electrochem. Soc.*, vol. 134, no. 8, pp. 1915–1920.

Wilkinson, J. H. (1988). *The Algebraic Eigenvalue Problem.* Oxford University Press, Oxford.

Zarrop, M. B. (1979). Optimal experiment design for dynamic system identification. *Series of Lecture Notes in Control and Information Sciences,* vol. 21. Springer-Verlag, Berlin.

Subject Index

System Identification: A Frequency Domain Approach. Rik Pintelon and Johan Schoukens
Copyright © 2001 Institute of Electrical and Electronics Engineers. ISBN: 0-780-36000-1

Reference Index

System Identification: A Frequency Domain Approach. Rik Pintelon and Johan Schoukens
Copyright © 2001 Institute of Electrical and Electronics Engineers. ISBN: 0-780-36000-1

About the Authors

Rik Pintelon received both the degree of electrical engineer in 1982 and the degree of doctor in applied sciences in 1988 from the Vrije Universiteit Brussel (VUB), Brussels, Belgium.

From 1982 to 2000, Dr. Pintelon was a researcher of the Belgian National Fund for Scientific Research (FWO-Vlaanderen) at the Electrical Engineering (ELEC) Department of the VUB. From 1991 to 2000 he was a part-time lecturer at the same department, where he is currently a full-time professor in electrical engineering. His main research interests include system identification, signal processing, and measurement techniques. Dr. Pintelon is the coauthor of one book and the coauthor of more than 70 articles in refereed international journals. He has been a Fellow of IEEE since 1998.

Johan Schoukens received both the degree of electrical engineer in 1980 and the degree of doctor in applied sciences in 1985 from the Vrije Universiteit Brussel (VUB), Brussels, Belgium.

From 1981 to 2000, Dr. Schoukens was a researcher of the Belgian National Fund for Scientific Research (FWO-Vlaanderen) at the Electrical Engineering (ELEC) Department of the VUB. From 1986 to 2000 he was a part-time lecturer in the same department, where he is currently a full-time professor in electrical engineering. His main research interests include system identification, signal processing, and measurement techniques. Dr. Schoukens is the coauthor of one book and the coauthor of more than 70 articles in refereed international journals. He has been a Fellow of IEEE since 1997.

System Identification: A Frequency Domain Approach. Rik Pintelon and Johan Schoukens