On the estimation of transfer functions, regularizations and Gaussian processes—Revisited☆

Tianshi Chen [1], Henrik Ohlsson, Lennart Ljung

Division of Automatic Control, Department of Electrical Engineering, Linköping University, Linköping SE-58183, Sweden

A B S T R A C T

Intrigued by some recent results on impulse response estimation by kernel and nonparametric techniques, we revisit the old problem of transfer function estimation from input–output measurements. We formulate a classical regularization approach, focused on finite impulse response (FIR) models, and find that regularization is necessary to cope with the high variance problem. This basic, regularized least squares approach is then a focal point for interpreting other techniques, like Bayesian inference and Gaussian process regression. The main issue is how to determine a suitable regularization matrix (Bayesian prior or kernel). Several regularization matrices are provided and numerically evaluated on a data bank of test systems and data sets. Our findings based on the data bank are as follows. The classical regularization approach with carefully chosen regularization matrices shows slightly better accuracy and clearly better robustness in estimating the impulse response than the standard approach – the prediction error method/maximum likelihood (PEM/ML) approach. If the goal is to estimate a model of given order as well as possible, a low order model is often better estimated by the PEM/ML approach, and a higher order model is often better estimated by model reduction on a high order regularized FIR model estimated with careful regularization. Moreover, an optimal regularization matrix that minimizes the mean square error matrix is derived and studied. The importance of this result lies in that it gives the theoretical upper bound on the accuracy that can be achieved for this classical regularization approach.

© 2012 Elsevier Ltd. All rights reserved.

1. Introduction

Estimation of the transfer function, or impulse response, of a linear system is a problem that we feel that we have known "everything about" for at least a quarter of a century, e.g. Ljung (1985), based on well established theory and algorithms in statistics and the system identification community. Nevertheless, papers on the problem are still appearing. A recent, very inspiring, and thought-provoking, contribution is Pillonetto and Nicolao (2010a) (see also the follow-up, (Pillonetto, Chiuso, & Nicolao, 2011)), which shows rather remarkable results based on Gaussian processes and spline kernels. That has prompted the current wish to revisit the transfer function estimation problem from scratch.

1.1. Problem formulation

Consider a single-input–single-output linear stable system

$$y(t) = G_0(q)u(t) + v(t). \tag{1}$$

Here q is the shift operator, $qu(t) = u(t+1)$, $v(t)$ is additive noise, independent of the input $u(t)$, and the transfer function is

$$G_0(q) = \sum_{k=1}^{\infty} g_k^0 q^{-k}. \tag{2}$$

The coefficients g_k^0, $k = 1, \ldots, \infty$, form the *impulse response* of the system. The corresponding frequency response is defined as

$$G_0(e^{i\omega}) = \sum_{k=1}^{\infty} g_k^0 e^{-i\omega k}. \tag{3}$$

Given the input–output data $Z^N = \{u(t), y(t), t = 1, \ldots, N\}$, the goal is to find an estimate $\hat{G}_N(e^{i\omega})$ of $G_0(e^{i\omega})$ that is as good as possible. A related goal is to assess and quantify the error in the estimate.

The traditional way is to postulate a finite-dimensional parameterization

$$G(q, \theta) \tag{4}$$

in terms of θ and then estimate θ in some suitable way and deliver the estimate $\hat{G}_N(e^{i\omega}) = G(e^{i\omega}, \hat{\theta}_N)$. Many such parameterizations have been suggested and tested in the literature, e.g. Ljung (1999). A distinct difficulty is to determine the "size" of the parameter vector θ and to assess the error that stems from $G_0(e^{i\omega})$ being outside the set of functions that is covered within the parameterization. Partly for that reason, alternative approaches based on other ideas, like Gaussian process regression, and non-parametric descriptions of the function $G_0(e^{i\omega})$ (or the impulse response) have recently been suggested, e.g. Pillonetto and Nicolao (2010a) and Pillonetto et al. (2011). Related methods for assessing the quality of $\hat{G}_N(e^{i\omega})$ have been discussed in the 1990s and early 2000s, Goodwin, Gevers, and Ninness (1992), Gustafsson and Hjalmarsson (1995) and Goodwin, Braslavsky, and Seron (2002) in connection with bias quantification.

1.2. Questions revisited

Suppose we are given a batch of input–output data. We have no information about the data, except that it is collected from a linear stable system with additive noise. The task is one of the following:

(a) Estimate, as well as possible, the impulse response of the unknown system.
(b) Estimate a model of given order that has an impulse response as close as possible to the unknown system.

The standard answers to these questions are

(for b) to use a prediction error method/maximum likelihood (PEM/ML) estimate for the given model structure.
(for a) to try several models of different orders, apply **(b)** and use model order/model selection techniques to pick the best model order, and finally get the PEM/ML estimate with the best model order.

We shall revisit these two questions with an emphasis on high order regularized FIR (finite impulse response) models that are simple, safe and robust ways of building linear models, directly focusing on the impulse response. This basic, regularized least squares approach is then a focal point for interpreting other techniques, like Bayesian inference and Gaussian process regression (Pillonetto et al., 2011; Pillonetto & Nicolao, 2010a). The main issue is how to determine a suitable regularization matrix (Bayesian prior or kernel). Several regularization matrices are provided and numerically evaluated on a data bank of test systems and data sets. A natural question is then if there exists an optimal regularization matrix. It turns out that there actually exists an optimal regularization matrix that minimizes the mean square error matrix and gives a theoretical upper bound on the accuracy that can be achieved.

Notations. Throughout the paper, let $\delta_{t,s}$ denote the Kronecker-delta function, i.e., if $t = s$, $\delta_{t,s} = 1$, otherwise $\delta_{t,s} = 0$, I_n denote the $n \times n$ identity matrix, $\mathrm{diag}(a_1, \ldots, a_n)$ denote a diagonal matrix with a_k as the (k, k)th element, $k = 1, \ldots, n$. Let $x|y \sim \mathcal{N}(m, P)$ denote that conditioned on y, x is a multivariate Gaussian random variable with mean vector m and covariance matrix P. For positive integers i, j, k with $i \le k$, let $i : j : k$ denote the vector $[i, i+j, i+2j, \ldots, i+mj]$, where m is the integer that rounds $(k-i)/j$ toward 0. For symmetric matrices A and B, let $A \ge B$ denote that $A - B$ is a positive semi-definite matrix.

2. A data bank of test systems and data sets

To test different techniques we generated a data bank of systems and data sets. They should be representative of real-life systems and data sets, in that the underlying system is not of low order (but could allow good low order approximations) and should correspond to different signal-to-noise ratios (SNR). We have done as follows:

- A number of 30th order random SISO continuous-time systems were generated using the command rss in MATLAB.
- These continuous-time systems were sampled at 3 times the bandwidth to yield the discrete-time systems using the following commands in MATLAB
  ```
  bw=bandwidth(m)
  f=bw*3*2*pi
  md = c2d(m,1/f,'zoh')
  ```
  where m is the continuous-time system and md is the corresponding discrete-time system.
- These discrete-time systems were split into 2500 "fast" systems S1 that have all their poles inside a circle with radius 0.95 and 2500 "slow" systems S2 which have at least one pole outside the circle with radius 0.95 (but inside the unit circle).
- The 5000 systems were simulated with an input which was white Gaussian noise with unit variance, and output additive white Gaussian noise with different variances:
  . low SNR: SNR = 1. The additive output noise has the same variance as the noise-free output. The number of data in these records is 375.
  . high SNR: SNR = 10. The additive output noise has a variance which is a tenth of the variance of the noise-free output. The number of data in these records is 500.
- This gives four collections of data sets.
  . S1D1: Fast systems with high SNR.
  . S2D1: Slow systems with high SNR.
  . S1D2: Fast systems with low SNR.
  . S2D2: Slow systems with low SNR.

All these data sets are accessible from http://www.rt.isy.liu.se/~tschen/research/regul_fir/systems_tested/.

To evaluate various methods, the estimates of the impulse response coefficients $\hat{g}_k$, $k = 1, \ldots, 125$, were compared to the true ones by the measure

$$W = 100 \left(1 - \left[\frac{\sum_{k=1}^{125} |g_k^0 - \hat{g}_k|^2}{\sum_{k=1}^{125} |g_k^0 - \bar{g}^0|^2} \right]^{1/2} \right), \qquad \bar{g}^0 = \frac{1}{125} \sum_{k=1}^{125} g_k^0. \tag{5}$$

The W in (5) corresponds to the "fit" in the compare command in the System Identification Toolbox, Ljung (2007). Note that $W = 100$ means a perfect fit between the true impulse response and the corresponding estimate for the first 125 coefficients. Each data set gives rise to a particular value of W, and in the tables below we give the average of W over all the sets in a certain collection.

3. A classical perspective

In the classical perspective $G_0(e^{i\omega})$ is unknown and estimated from the data. The estimate is a random variable (due to the noise $v(t)$) and the quality can be assessed by the "distance" between the estimate and the true value.

A reasonable measure is the mean square error (MSE)

$$M_N(\omega) = E|\hat{G}_N(e^{i\omega}) - G_0(e^{i\omega})|^2. \tag{6}$$

Here, the expectation E is with respect to the output noise process $v(t)$. Now, the MSE $M_N(\omega)$ is classically split into a bias part

$$B_N(\omega) = E\hat{G}_N(e^{i\omega}) - G_0(e^{i\omega}) \tag{7}$$

and a variance part

$$V_N(\omega) = E|\hat{G}_N(e^{i\omega}) - E\hat{G}_N(e^{i\omega})|^2 \tag{8}$$

so that

$$M_N(\omega) = V_N(\omega) + |B_N(\omega)|^2. \tag{9}$$

3.1. Trading variance for bias to minimize the MSE

In the expression for the MSE $M_N(\omega)$, the bias term $B_N(\omega)$ decreases and the variance term $V_N(\omega)$ increases, when the model becomes more flexible (contains more essential parameters). The MSE $M_N(\omega)$ is then often minimized for a model flexibility that does not give zero bias. In other words, a pragmatic choice of model flexibility allows some bias to reduce variance so that the MSE $M_N(\omega)$ is minimized.

3.2. OE-models

We will not be concerned with noise models in this contribution, so a natural numerator/denominator model is

$$G(q, \theta) = \frac{B(q, \theta)}{F(q, \theta)} \tag{10}$$

where $B(q, \theta)$ and $F(q, \theta)$ are polynomials of q^{-1}. The PEM/ML approach to the estimation of (10) would be

$$\hat{\theta}_N^{OE} = \arg \min_\theta \sum_{t=1}^N |y(t) - G(q, \theta)u(t)|^2. \tag{11}$$

The estimation involves a search for the solution of the non-convex problem (11), which may lead to local minima and possibly ill-conditioned calculations. An alternative is to fix the denominator $F(q, \theta)$ to 1 (or any fixed, stable, polynomial) so that a linear regression problem is obtained.

3.3. FIR-models

The simplest approach to estimate $G(q, \theta)$ is to truncate the expansion (2) at a finite number of impulse response coefficients ("FIR" model, corresponding to fixing $F(q, \theta) = 1$ in (10))

$$G(q, \theta) = \sum_{k=1}^n g_k q^{-k}, \quad \theta = [g_1 \, g_2 \, \dots \, g_n]^T \tag{12}$$

where n is the order of the FIR model. The vector θ is then easily estimated by the least squares method. Write the model as

$$y(t) = \varphi^T(t)\theta + v(t), \qquad \varphi(t) = [u(t-1) \, \dots \, u(t-n)]^T \tag{13a}$$

or $Y_N = \Phi_N^T \theta + \Lambda_N \tag{13b}$

where $Y_N = [y(n+1) \, y(n+2) \, \dots \, y(N)]^T \tag{13c}$

$$\Phi_N = [\varphi(n+1) \, \varphi(n+2) \, \dots \, \varphi(N)] \tag{13d}$$

$$\Lambda_N = [v(n+1) \, v(n+2) \, \dots \, v(N)]^T. \tag{13e}$$

The least squares solution is well known:

$$\hat{\theta}_N^{LS} = [\hat{g}_1^{LS} \, \hat{g}_2^{LS} \, \dots \, \hat{g}_n^{LS}]^T = \arg \min v_N(\theta) \tag{14a}$$

$$v_N(\theta) = \|Y_N - \Phi_N^T \theta\|^2 = \sum_{t=n+1}^N (y(t) - \varphi^T(t)\theta)^2 \tag{14b}$$

$$\hat{\theta}_N^{LS} = (\Phi_N \Phi_N^T)^{-1} \Phi_N Y_N = R_N^{-1} F_N \tag{14c}$$

$$F_N = \Phi_N Y_N = \sum_{t=n+1}^N \varphi(t)y(t) \tag{14d}$$

$$R_N = \Phi_N \Phi_N^T = \sum_{t=n+1}^N \varphi(t)\varphi(t)^T. \tag{14e}$$

Remark 1. Since $u(-n + 1), \dots, u(0)$ are not known, the summation in (14b) starts at $n + 1$ to allow $\varphi(t)$ to be formed. This is known as the 'non-windowed' case. As can be seen from (13c), this means that the first n outputs, $y(1), y(2), \dots, y(n)$ in the data set $Z^N = \{u(t), y(t), t = 1, \dots, N\}$ are not used.

How good is the resulting FIR model? Let us assume that

$$Ev(t) = 0, \qquad Ev(t)v(s) = \sigma^2 \delta_{t,s}. \tag{15}$$

The input $u(t)$ (and thus $\varphi(t)$) is seen as a deterministic variable, and for the conceptual analysis here, for simplicity we will assume that there exists $\mu > 0$ such that

$$\frac{1}{N-n} R_N \to \mu I_n \quad \text{as } N \to \infty. \tag{16}$$

This will hold w.p. 1 if $u(t)$ is chosen as white noise with variance μ but may be true under many other choices of input (PRBS, certain multi-sine input etc.). This means that for reasonably large N,

$$\frac{1}{N-n} R_N \approx \mu I_n. \tag{17}$$

Then it is immediate to show that

$$E\hat{\theta}_N^{LS} = \theta_0 = [g_1^0 \, g_2^0 \, \dots \, g_n^0]^T \tag{18}$$

$$E(\hat{\theta}_N^{LS} - \theta_0)(\hat{\theta}_N^{LS} - \theta_0)^T = \sigma^2 R_N^{-1} \approx \frac{\sigma^2}{(N-n)\mu} I_n \tag{19}$$

which gives the bias, variance, and MSE, corresponding to (7) to (9), as follows

$$B_N(\omega) = \sum_{k=n+1}^\infty g_k^0 e^{i\omega k} \tag{20a}$$

$$V_N(\omega) \approx \frac{n\sigma^2}{(N-n)\mu} \tag{20b}$$

$$M_N(\omega) \approx \frac{n\sigma^2}{(N-n)\mu} + \left| \sum_{k=n+1}^\infty g_k^0 e^{i\omega k} \right|^2. \tag{20c}$$

It is well known from Ljung and Wahlberg (1992) that by letting the order n increase to infinity with the number of data N, sufficiently slowly, the model (12) will converge to the true transfer function (2). To minimize the MSE $M_N(\omega)$ with respect to the order n for a given data size N requires some idea on the size of $B_N(\omega)$ as a function of n. Assume that the system has all poles inside a circle with radius $\bar{\lambda}$. Then there exists a $\bar{c} > 0$ such that

$$|g_k^0| < \bar{c}\bar{\lambda}^k \tag{21a}$$

$$|B_N(\omega)| < \frac{\bar{c}\bar{\lambda}^{n+1}}{1 - \bar{\lambda}}. \tag{21b}$$

So, since the squared bias decreases like $\bar{\lambda}^{2n}$ as a function of n and the variance increases like n (for large N) an upper bound on the MSE $M_N(\omega)$ is minimized by an order n that increases with N like

$$n_{\text{opt}} \sim \log N. \tag{22}$$

As a result, the upper bound on the MSE $M_N(\omega)$ is minimized at relatively low orders compared to the data size.

3.4. Regularization

Still, we see that the variance increases linearly with the FIR model order n so for higher order FIR models it is important to counteract the increasing variance by *regularization*. This is an example of pragmatic bias–variance trade-off, cf. Section 3.1. Regularization means that we replace the criterion $v_N(\theta)$ in (14) by

$$v_N^R(\theta, D) = \sum_{t=n+1}^{N} (y(t) - \varphi^T(t)\theta)^2 + \theta^T D\theta \tag{23a}$$

where D is positive semi-definite and called a regularization matrix. That changes the estimate to be

$$\hat{\theta}_N^R = [\hat{g}_1^R \, \hat{g}_2^R \dots \hat{g}_n^R]^T \tag{23b}$$

$$= (R_N + D)^{-1} F_N = (R_N + D)^{-1} R_N \hat{\theta}_N^{LS}. \tag{23c}$$

We now disregard the tail of the impulse response g_k^0, $k > n$ and assume that the true system is given by a FIR model of order n.

How to select D? We have (all expectations are with respect to $v(t)$)

$$E\hat{\theta}_N^R = (R_N + D)^{-1} R_N \theta_0 \tag{24a}$$

$$\theta_{bias}^R = E\hat{\theta}_N^R - \theta_0 = -(R_N + D)^{-1} D\theta_0 \tag{24b}$$

$$\tilde{\theta} = \hat{\theta}_N^R - E\hat{\theta}_N^R = (R_N + D)^{-1} R_N(\hat{\theta}_N^{LS} - \theta_0) \tag{24c}$$

$$E\tilde{\theta}\tilde{\theta}^T = (R_N + D)^{-1} \sigma^2 R_N (R_N + D)^{-1} \tag{24d}$$

$$\begin{aligned}
MSE(\hat{\theta}_N^R) &= E(\hat{\theta}_N^R - \theta_0)(\hat{\theta}_N^R - \theta_0)^T \\
&= E\tilde{\theta}\tilde{\theta}^T + \theta_{bias}^R(\theta_{bias}^R)^T \\
&= (R_N + D)^{-1}(\sigma^2 R_N + D\theta_0\theta_0^T D^T)(R_N + D)^{-1}
\end{aligned} \tag{24e}$$

where $MSE(\hat{\theta}_N^R)$ is the *MSE* matrix of $\hat{\theta}_N^R$ with respect to the true impulse response coefficients vector θ_0 in (18).

Suppose that $D = diag(d_1, d_2, \dots, d_n)$ and (17) is used for R_N. The (k, k)th element of $MSE(\hat{\theta}_N^R)$ has the form

$$MSE(\hat{g}_k^R) \approx \frac{\sigma^2 \mu(N - n) + d_k^2(g_k^0)^2}{(\mu(N - n) + d_k)^2} \tag{25}$$

which is minimized with respect to d_k by $d_k = \sigma^2/(g_k^0)^2$. Therefore this gives a clue how to choose the regularization matrix D: If the system is stable as in (21a), the diagonal of D should increase exponentially:

$$d_k = \frac{\sigma^2}{c\lambda^k}, \quad k = 1, \dots, n \tag{26}$$

where $\lambda = \bar{\lambda}^2$ and $c = \bar{c}^2$.

Remark 2. The LS solution of the nth order FIR model (12) can be seen as a special case of regularization for a higher mth order FIR model: If we choose the diagonal regularization $D = diag(d_1, d_2, \dots, d_m)$ with $m > n$ and

$$d_k = \begin{cases} 0 & \text{if } k \leq n \\ \infty & \text{if } k > n \end{cases} \tag{27}$$

then the regularized LS estimate of the m order model is equal to the usual LS estimate of the n-th order model.

Remark 3. Regularization as in (23a) is often used in a Tikhonov sense, Tikhonov and Arsenin (1977), where the objective is to make an ill-conditioned problem have better numerical properties. Here, however, the main aspect of regularization is to better deal with the bias–variance trade-off (9). But the concepts are closely linked. Indeed, if the input u for example is band-limited, then the matrix R_N will become very ill-conditioned for large n. At the same time, the variance of the regular LS estimate, $\sigma^2 R_N^{-1}$ (cf. (19)) will be very large, and the need for a careful bias–variance trade-off is obvious.

Remark 4. From (24b) we see that for the regularized estimate, the bias is linear in the true parameter. This means that the estimation problem is of the kind studied in Eldar (2006), to which we refer for general comments and insights. For the treatment here it is more convenient to directly infer the pertinent results.

Remark 5. A standard way in statistics to combine two unbiased parameter estimates θ_1 and θ_2 with covariance matrices P_1 and P_2 to yield an unbiased estimate with minimum variance is to form

$$\theta = (P_1^{-1} + P_2^{-1})^{-1}(P_1^{-1}\theta_1 + P_2^{-1}\theta_2). \tag{28}$$

In that perspective the regularized estimate (23b) can be seen as the combination of the un-regularized estimate $\hat{\theta}_N^{LS}$ and an estimate $\bar{\theta} = [0 \, 0 \, \dots \, 0]^T$ with variance $\sigma^2 D^{-1}$.

3.5. Using a base-line model

If the impulse response is decaying slowly, a high order FIR model will be required to capture that. It may then be beneficial to incorporate a "base-line model" that can take care of a dominating part of the impulse response. For example, a model with an additive base-line model can be like

$$G(q, \eta, \theta) = G_b(q, \eta) + G_r(q, \theta) \tag{29a}$$

$$\text{with } G_r(q, \theta) = \sum_{k=1}^{n} g_k q^{-k}. \tag{29b}$$

Here $G_b(q, \eta)$ is the base-line model, e.g. of the kind (10), with η the associated parameter vector and $G_r(q, \theta)$ a high order FIR model and θ is defined as in (12).

Given the input–output data $Z^N = \{u(t), y(t), t = 1, \dots, N\}$, a base-line model $G_b(q, \hat{\eta}_N)$ is first estimated separately using e.g. PEM/ML methods. With

$$y_b(t) = G_b(q, \hat{\eta}_N)u(t) \tag{30}$$

the residual output $y_r(t)$ is defined as

$$y_r(t) = y(t) - y_b(t). \tag{31}$$

So a new input–output data $Z_r^N = \{u(t), y_r(t), t = 1, \dots, N\}$ is formed and the FIR model $G_r(q, \theta)$ in (29) can then be estimated using the regularization method as in (23).

An interpretation of how the base-line model enters a general model description will be given in Section 4.2.

3.6. Cross-validation

Using the classical methods mentioned in Sections 3.2–3.4 for optimal MSE means that we must know certain variables (say β), like the best OE model order, the best FIR model order n in (22) or the optimal regularization parameters c, λ in (26). The necessary information to compute these are typically not known, which in the classical perspective typically is handled by *cross-validation*:

(1) Split the data record into two parts: an estimation data part and a validation data part.

(2) Estimate models $G(q, \hat{\theta}_N)$ using the estimation data for different values of β.
(3) Form the error between the measured and the model outputs for these models using the validation data:

$$\epsilon(t, \beta) = y(t) - G(q, \hat{\theta}_N)u(t) \tag{32}$$

$$W(\beta) = \sum_t |\epsilon(t, \beta)|^2 \tag{33}$$

and pick the value of β that minimizes $W(\beta)$. The model can then be re-estimated for this β using the whole data record.

3.7. Numerical illustration

Let us try these methods on our data bank of data sets as shown in Section 2.

Example 1 (*Fixed Order OE Models*). We estimate models (10) of different orders n (same order for $B(q, \theta)$ and $F(q, \theta)$) using the command $\mathtt{m = oe(data, [n, n, 1])}$ in the System Identification Toolbox, Ljung (2007), and compute the average fit (5) for all models in the corresponding data set.

The results are shown in the table below. It also contains the fits when the order n for each data set has been chosen by cross-validation (CV) testing orders 5:5:40.

	$n = 5$	$n = 15$	$n = 25$	$n = 35$	$n = 40$	CV
S1D1	86.3	86.4	74.2	54.9	42.6	89.4
S2D1	68.7	71.7	63.1	49.3	42.0	73.2
S1D2	71.9	56.1	34.5	10.2	−1.7	70.8
S2D2	50.8	42.3	20.4	−2.1	−8.5	49.6

Example 2 (*Fixed Order FIR Models*). We estimate models (12) of different orders n using the least squares method (14) and compute the average fit (5) for all models in the corresponding data set.

The results are shown in the table below. It also contains the fits when the order for each data set has been chosen by cross-validation (CV) testing orders 5:10:125.

	$n = 5$	$n = 35$	$n = 65$	$n = 95$	$n = 125$	CV
S1D1	32.2	83.1	85.8	81.7	76.9	86.1
S2D1	−0.7	47.1	60.0	64.0	65.3	67.4
S1D2	30.8	61.4	46.0	25.9	−0.1	59.6
S2D2	−1.8	30.5	24.2	8.0	−18.2	30.5

Example 3 (*FIR-Models of Order 125 with Regularization*). We estimate models (12) of order 125 using the regularization method (23) with diagonal D for different values of c and λ in (26), and compute the average fit (5) for all models in the corresponding data set. Throughout the simulations in this paper, the variance σ^2 is estimated from the sample variance of the estimated FIR model (12) of order 125 using the least squares method.

The results are shown in the table below. It also contains the fits when c and λ for each data set have been chosen by cross-validation (CV) testing the grid of 9 values, $c = 1, 5, 9$ and $\lambda = 0.5, 0.9, 0.95$.

	$c = 1$ $\lambda = 0.5$	$c = 1$ $\lambda = 0.9$	$c = 1$ $\lambda = 0.95$	$c = 9$ $\lambda = 0.5$	$c = 9$ $\lambda = 0.95$	CV
S1D1	51.0	84.8	79.2	58.2	77.5	84.8
S2D1	18.4	67.8	66.8	24.5	65.6	67.1
S1D2	37.4	54.9	36.3	44.7	17.1	55.6
S2D2	6.4	29.5	8.6	12.7	−7.5	23.3

Example 4 (*As Example 3, but with Base-Line Model (29)*). We estimate models (29) where an additive second order base-line model $G_b(q, \eta)$ is first identified using the command $\mathtt{m = oe(data, [2, 2, 1])}$, then the new data set $Z_r^N = \{u(t), y_r(t), t = 1, \ldots, N\}$ as described in Section 3.5 is formed, and finally an FIR model (12) of order 125 is estimated using the regularization method as in Example 3.

	$c = 1$ $\lambda = 0.5$	$c = 1$ $\lambda = 0.9$	$c = 1$ $\lambda = 0.95$	$c = 9$ $\lambda = 0.5$	$c = 9$ $\lambda = 0.95$	CV
S1D1	74.8	85.4	79.3	78.0	77.5	86.7
S2D1	56.5	72.2	69.6	58.7	68.4	74.1
S1D2	62.2	57.5	37.4	64.3	17.1	66.4
S2D2	42.2	32.6	9.8	42.7	−6.4	45.8

Findings. The "standard" approach (Example 1), works well. It is the best approach for all data sets except S2D1. Note that in the simulated data, the "true" order is 30, but this is normally not the best order choice for the OE models. The experiments in Example 2 also show that although the true impulse response is infinite, it is normally not the best choice to use maximum FIR model order. The high variance for such models overrides the low bias. Choosing the FIR model order by cross-validation gives a fit between 30%–85%. Using FIR models of order 125 and regularization (23) with diagonal D in (26) (Example 3) does not always improve the fit for all the c, λ tests, and the good affect is largely dependent on their values, so they should be chosen with care. The cross-validation choice of c, λ over the 9 point-grid gives a fit of about the same size as cross-validation over orders. Adding a second order base-line model, (Example 4), is beneficial, mostly so for the slow systems.

4. A Bayesian perspective

In the Bayesian view, the parameter to be estimated is itself a random variable, and we seek the posterior distribution of this parameter, given the observations.

The following well known and simple result about conditioning jointly Gaussian random variable is a key element in Bayesian calculations. Let

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} m_1 \\ m_2 \end{bmatrix}, \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} \right). \tag{34a}$$

Then

$$x_1 | x_2 \sim \mathcal{N}(m, P) \tag{34b}$$

$$m = m_1 + P_{12} P_{22}^{-1} (x_2 - m_2) \tag{34c}$$

$$P = P_{11} - P_{12} P_{22}^{-1} P_{21}. \tag{34d}$$

It is also good to recall the following simple matrix equality:

$$A(I_j + BA)^{-1} = (I_k + AB)^{-1} A \tag{35}$$

where A is an $k \times j$ matrix and B is a $j \times k$ matrix.

In the current setup, we regard the parameter of the nth order FIR model (12), i.e., the impulse response coefficients vector θ as a random variable, say of Gaussian distribution with zero mean and covariance matrix P_n:

$$\theta \sim \mathcal{N}(\theta^{ap}, P_n), \quad \theta^{ap} = 0. \tag{36}$$

If the input $u(t)$ (and $\varphi(t)$, see (13a)) is known and the noise $v(t)$ is independently Gaussian distributed with

$$v(t) \sim \mathcal{N}(0, \sigma^2) \tag{37}$$

then with

$$Y_N = \Phi_N^T \theta + \Lambda_N. \tag{38}$$

Y_N and θ will be jointly Gaussian variables:

$$\begin{bmatrix} \theta \\ Y_N \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} P_n & P_n \Phi_N \\ \Phi_N^T P_n & \Phi_N^T P_n \Phi_N + \sigma^2 I_{N-n} \end{bmatrix} \right). \tag{39}$$

The posterior distribution of θ given Y_N follows from (34)

$$\theta | Y_N \sim \mathcal{N}(\hat{\theta}_N^{apost}, P_N^{apost}) \tag{40a}$$

$$\hat{\theta}_N^{apost} = P_n \Phi_N (\Phi_N^T P_n \Phi_N + \sigma^2 I_{N-n})^{-1} Y_N \tag{40b}$$

$$= (P_n \Phi_N \Phi_N^T + \sigma^2 I_n)^{-1} P_n \Phi_N Y_N \tag{40c}$$

$$= (R_N + \sigma^2 P_n^{-1})^{-1} F_N \tag{40d}$$

$$= ((\sigma^2 R_N^{-1})^{-1} + P_n^{-1})^{-1} (\sigma^2 R_N^{-1})^{-1} \hat{\theta}_N^{LS} \tag{40e}$$

$$P_N^{apost} = P_n - P_n \Phi_N (\Phi_N^T P_n \Phi_N + \sigma^2 I_{N-n})^{-1} \Phi_N^T P_n \tag{40f}$$

$$= P_n - (P_n \Phi_N \Phi_N^T + \sigma^2 I_n)^{-1} P_n \Phi_N \Phi_N^T P_n$$

$$= ((\sigma^2 R_N^{-1})^{-1} + P_n^{-1})^{-1} \tag{40g}$$

where $F_N, R_N, \hat{\theta}_N^{LS}$ are defined in (14). Moreover, (40b) and (40f) are the expressions from (34) while the steps to (40e) and (40g) using (35) stress the link to (28) merging the models $\hat{\theta}_N^{LS}$ and $\theta^{ap} = 0$.

We notice that this *a posteriori* estimate $\hat{\theta}_N^{apost}$ is the same as the regularized estimate $\hat{\theta}_N^R$ if the regularization matrix D is chosen as

$$D = \sigma^2 P_n^{-1}. \tag{41}$$

This is just a restatement of the well-known fact that regularization is closely related to prior estimates.

So this gives an insight into how to choose the regularization matrix: Let it reflect the size and correlations of the impulse response coefficients. For the size, it is entirely in line with the choice of diagonal elements (26). If the impulse response is smooth (for example a fast sampled continuous system) it is also natural to let P_n reflect that, by letting the diagonals close to the main diagonal show high correlation. A simple choice is to let the correlation coefficient between g_k and g_j in (12) be $\rho^{|k-j|}$. With diagonal elements of P_n being $c\lambda^k$ as in (26) we then get a covariance matrix P_n whose (k, j)th element is

$$c\rho^{|k-j|}\lambda^{(k+j)/2} \tag{42}$$

where $c \geq 0$, $0 \leq \lambda \leq 1$ and $|\rho| \leq 1$. The estimates that we come up with are thus the same as in the classical, regularized estimate (23b), but the Bayesian perspective has given additional insights into the choice of D.

4.1. Estimating hyper-parameters

The Bayesian perspective gives one more insight. Suppose that prior knowledge does not give a definite choice of P_n, but it is natural to let it depend on unknown hyper-parameters β, $P_n(\beta)$ (like $\beta = [c \; \lambda]$ in (26)). From (39) we see that

$$Y_N \sim \mathcal{N}(0, \sigma^2 I_{N-n} + \Phi_N^T P_n(\beta) \Phi_N) \tag{43a}$$

so with a classical twist in this Bayesian framework we can form the likelihood function of the observation Y_N given β, and estimate β by the maximum likelihood (ML) method:

$$\hat{\beta} = \arg\min_{\beta} Y_N^T \Sigma(\beta)^{-1} Y_N + \log\det \Sigma(\beta) \tag{43b}$$

where $\Sigma(\beta) = \sigma^2 I_{N-n} + \Phi_N P_n(\beta) \Phi_N^T$. This method of estimating hyper-parameters in the prior distribution is known as the *empirical Bayes* methods (Carlin & Louis, 1996).

The noise variance σ^2 used in (43b) and (41) can of course be included among the hyper-parameters, but in the simulations in this paper we used the way as mentioned in Example 3.

4.2. Base-line model as a prior model

The prior mean θ^{ap} in (36) is usually set to 0. It is interesting to see that in the Bayesian perspective the impulse response coefficients vector of the base-line model in (29) can actually be seen as a nonzero prior mean θ^{ap} in (36).

Given the data $Z^N = \{u(t), y(t), t = 1, \ldots, N\}$, a base-line model $G_b(q, \hat{\eta}_N)$ is first estimated. Then from (30) and (31),

$$Y_N^r = Y_N - Y_N^b \tag{44}$$

where $Y_N^* = [y_*(n+1) \; y_*(n+2) \; \ldots \; y_*(N)]^T$ and "*" represents either "r" or "b". Using the Bayesian method as described above, the impulse response coefficient vector of the FIR model $G_r(q, \hat{\theta}_N^{apost})$ in (29) resulting from $Z_r^N = \{u(t), y_r(t), t = 1, \ldots, N\}$ can be written as

$$\hat{\theta}_N^{apost} = P_n(\beta) \Phi_N (\Phi_N^T P_n(\beta) \Phi_N + \sigma^2 I_{N-n})^{-1} Y_N^r. \tag{45}$$

Here the hyper-parameter β is determined by the maximum likelihood method:

$$\hat{\beta} = \arg\min_{\beta} (Y_N^r)^T \Sigma(\beta)^{-1} Y_N^r + \log\det \Sigma(\beta) \tag{46}$$

where $\Sigma(\beta)$ is defined in (43b).

On the other hand, let $\hat{g}_k^b$ and $\hat{g}_k^{apost} k = 1, \ldots, \infty$, denote the impulse response coefficients of the base-line model $G_b(q, \hat{\eta}_N)$ and the FIR model $G_r(q, \hat{\theta}_N^{apost})$ in (29), respectively. Then for sufficiently large n, the model (29) satisfies

$$G(q, \hat{\eta}_N, \hat{\theta}_N^{apost}) \approx \sum_{k=1}^{n} (\hat{g}_k^b + \hat{g}_k^{apost}) q^{-k}. \tag{47}$$

Now let $\hat{\theta}_N^{apost} = [\hat{g}_1^{apost}, \hat{g}_2^{apost}, \ldots, \hat{g}_n^{apost}]^T$ and $\hat{\theta}_N^b = [\hat{g}_1^b, \hat{g}_2^b, \ldots, \hat{g}_n^b]^T$. Then from (44) to (46), and $Y_N^b \approx \Phi_N^T \hat{\theta}_N^b$, it is straightforward to see that for sufficiently large n, the impulse response coefficients vector $\hat{\theta}_N^b + \hat{\theta}_N^{apost}$ of $G(q, \hat{\eta}_N, \hat{\theta}_{N apost})$ can be seen as the posterior mean of θ given Y_N with the prior distribution

$$\theta \sim \mathcal{N}(\theta^{ap}, P_n), \quad \theta^{ap} = \hat{\theta}_N^b. \tag{48}$$

4.3. Numerical illustration

Let us test, on the data bank of data sets as shown in Section 2, the Bayesian method (40) and (43) with the following prior covariances: the diagonal (26) and the correlation (42)

$$P_{DI}(k, j) = \begin{cases} c\lambda^k & \text{if } k = j \\ 0 & \text{else} \end{cases} \quad (\text{'Diagonal'}) \tag{49a}$$

$$P_{DC}(k, j) = c\rho^{|k-j|}\lambda^{(k+j)/2} \quad (\text{'Diagonal/correlated'}) \tag{49b}$$

where the hyper-parameters are $c \geq 0, 0 \leq \lambda \leq 1$ and $|\rho| \leq 1$. The complexity of the prior (49b) can be reduced by linking $\rho = f(\lambda)$, where $f(\cdot)$ could be, for example, either a nondecreasing function that satisfies $f(0) = 0$ and $f(1) = 1$ or a nonincreasing function that satisfies $f(0) = 0$ and $f(1) = -1$. Here, we test a special case of the prior (49b) by linking $\rho = \lambda^{1/2}$:

$$P_{TC}(k, j) = c \min(\lambda^j, \lambda^k) \quad (\text{'Tuned/correlated'}). \tag{49c}$$

Remark 6. It's interesting to note that the prior (49c) actually corresponds to the stable spline kernel of order 1 introduced using stochastic arguments in Pillonetto et al. (2011). We refer to Pillonetto, Chiuso, and De Nicolao (2010) for discussions regarding the comparison between the performance of the stable spline kernels of orders 1 and 2.

Example 5 (*Testing ML Estimation of Hyper-Parameters*)**.** We first estimate models (12) of order 125 using the Bayesian method (40) and (43) with the prior covariances (49). Then we estimate models (29) where an additive second order base-line model $G_b(q, \eta)$ is first identified using the command m = oe(data, [2, 2, 1]), then the new data set $Z_r^N = \{u(t), y_r(t), t = 1, \ldots, N\}$ as described in Section 3.5 is formed, and finally an FIR model (12) of order 125 is estimated using the Bayesian method (40) and (43) again.

The average fit (5) is calculated and the simulation results are shown in the table below, where an "e" is appended to the regularization matrix name if a base-line model is used.

	DI	DC	TC	DIe	DCe	TCe
S1D1	86.7	90.8	90.3	88.9	91.2	91.1
S2D1	68.6	78.0	77.8	75.6	81.6	81.6
S1D2	61.8	72.7	72.4	68.9	74.0	74.1
S2D2	33.2	60.7	60.8	50.6	62.2	61.8

Findings. We see that estimating the hyper-parameters for DI and DIe give about the same fit as the CV in Examples 3 *and* 4*. The ML estimates of the hyper-parameters are slightly better though, perhaps since the search is over a continuum of* c, λ *and not just the 9-point grid, used for CV. It is also clear that allowing and estimating correlation between the impulse response coefficients with DC, and TC gives a clear improvement. It should be noted that the criterion (43b) is not convex, so it requires some care to initialize the search and search for the minimum. This can be illustrated by the fact that TC actually behaves better than DC in some cases, though it is a special case of DC, but with fewer parameters. In all the tests, we initialize* $c = \exp(5)$*,* $\rho = 0.5$*. Since the optimization problem (43b) is sensitive to the initial value of* λ*, we solved (43b) twice with two initial values of* λ*, 1 and 0.5, respectively. The hyper-parameter estimate that gave a larger likelihood* $p(Y_N|\beta)$ *was chosen as the ultimate hyper-parameter estimate.*

5. Gaussian process regression to the transfer function estimation

Gaussian process regression (GPR) has become a widely spread and very popular method for inference in machine learning, see, e.g. Rasmussen and Williams (2006). In short, it is about inferring an unknown function $f(x)$ from measurements $y_k, k = 1, 2, \ldots, N$ that bear some information about $f(x)$. The argument x can either be a continuous or a discrete variable. The prior information about the function is that it is a Gaussian process, with a certain mean and covariance function. This means that the vector $[f(x_1), f(x_2), \ldots, f(x_n)]$, for any collection of points x_i is a jointly Gaussian random vector, with mean $m(x) = Ef(x)$ and covariances

$$Cov(f(x_i), f(x_j)) = P(x_i, x_j) \tag{50}$$

where $P(x_i, x_j)$ is often called a *kernel*. Often $m(x) \equiv 0$. Typically, the observation y_k is a linear functional of $f(x_i)$, measured in additive Gaussian noise. This causes $[f(x), y_1, \ldots, y_N]$ to be a jointly Gaussian vector, which means that the posterior distributions,

$$p(f(x_1), \ldots, f(x_n)|y_1, \ldots, y_N) \tag{51}$$

can be calculated by the rules for conditioning jointly Gaussian random variables, (34).

In Pillonetto and Nicolao (2010a) the GPR is applied to estimating the impulse response of a stable linear system. For a sampled model, the impulse response function is given by $g_k^0, k = 1, \ldots, \infty$ in (2). The observation y_k is the measured output in (1) at time $t = k$. Modeling the impulse response function as a Gaussian process means that, for any n,

$$[g_1, \ldots, g_n] \sim \mathcal{N}(0, P_n) \tag{52}$$

where P_n is the $n \times n$ upper left block matrix of the semi-infinite matrix P defined in (50). This is the same situation as in the Bayesian perspective (36)–(40). The Gaussian process estimate of any collections of impulse response coefficients is thus given by (40).

The only thing that remains to be discussed is the choice of prior covariances (52) (or (50)). Of course, the considerations for choosing P_n in (52) and in (36) must be the same, and the relation to the thoughts about the regularization matrix D in (41) still holds. But in GPR several standard choices for (50) exist.

In Pillonetto and Nicolao (2010a) the following kernels/covariance functions are discussed

$$P_{CS}(k, j) = \begin{cases} c\dfrac{k^2}{2}\left(j - \dfrac{k}{3}\right), & k \geq j \\[2mm] c\dfrac{j^2}{2}\left(k - \dfrac{j}{3}\right), & k < j \end{cases} \quad \text{('Cubic Spline')} \tag{53a}$$

$$P_{SE}(k, j) = ce^{-\frac{(k-j)^2}{2\lambda^2}} \quad \text{('Squared Exponential')} \tag{53b}$$

$$P_{SS}(k, j) = \begin{cases} c\dfrac{\lambda^{2k}}{2}\left(\lambda^j - \dfrac{\lambda^k}{3}\right), & k \geq j \\[2mm] c\dfrac{\lambda^{2j}}{2}\left(\lambda^k - \dfrac{\lambda^j}{3}\right), & k < j \end{cases}$$

('Stable Spline') $\tag{53c}$

where the hyper-parameters $c \geq 0, 0 \leq \lambda \leq 1$. There is also a MATLAB toolbox, Pillonetto and Nicolao (2010b) that implements the GPR, including estimating the hyper-parameters using (43).

Let us test the GPR approach with the kernels (53) on the data bank of data sets as shown in Section 2.

Example 6 (*D-Matrices Suggested in the GPR Approach*)**.** Similar to Example 5, let us estimate the models (12) of order 125 and (29) with the kernels (53).

The average fit (5) is calculated and the simulation results are shown in the table below, where an "e" is appended to the kernel name if a base-line model is used.

	CS	SE	SS	CSe	SEe	SSe
S1D1	78.0	80.8	90.3	81.6	84.2	90.4
S2D1	38.8	74.7	77.9	47.9	78.9	81.2
S1D2	16.6	44.4	70.1	60.7	65.7	71.6
S2D2	12.1	48.3	58.5	−44.3	58.6	59.6

Findings. The CS kernel, has difficulties with the slow systems, while the kernel SS shows a performance compatible with DC, DI and TC in Example 5*.*

Remark 7. The simulation results reported here are obtained using our own implementation. Similar results can be obtained using the stable spline toolbox for system identification (Pillonetto & Nicolao, 2010b). The difference lies in the estimation of σ^2 and on the other hand, in that two initial values of λ are used in solving (43b) as in Example 5.

Remark 8. For the test data bank, the DC, TC and SS priors/kernels give quite close results, while in some cases the DC prior is slightly better. This is perhaps because the DC prior uses an independent argument ρ to describe the correlations between impulse response coefficients. This extra flexibility enables the DC prior to capture more complicate impulse responses, although it also adds extra difficulty in solving the optimization problem (46).

Remark 9. It is fair to add that the theory around GPR and its relation to Bayesian estimation is much richer than shown here. The estimation of continuous time impulse responses can be handled in the same framework and there are interesting connections to Reproducing Kernel Hilbert Spaces (RKHS) and spline approximation. Our point here is that the actual resulting impulse response estimate is a regularized FIR estimate (23b) for certain choices of regularization matrix D. We refer to Pillonetto and Nicolao (2010a) for a more complete account of the theory.

Remark 10. It is interesting to note that a parametric model is also used in Pillonetto et al. (2011)(cf. Section 4.2). The parametric model therein is used to construct the prior covariance P_n of the impulse response coefficients θ so that it can capture impulse responses with more complicate behaviors. Moreover, the parametric model therein is estimated jointly with the hyperparameters by maximizing the marginal likelihood. In contrast, the parametric model $G_b(q, \hat{\eta}_N)$ here plays a role of a prior mean of θ, and estimated separately with the nonparametric model $\hat{\theta}_N^{apost}$. One benefit of the way used here is that two very low-dimensional optimization problems need to be solved instead of one low-dimensional one, making maybe the solution less exposed to local minima. These two different ways of using parametric models are tested on the test data bank. The simulation shows that they give quite close results, while in some cases the way introduced here is slightly better.

6. Optimal regularization matrix

We have seen that a focus in the discussions has been to find a proper regularization matrix D in (23a). This is the same problem as finding a proper prior covariance matrix P_n in (36), or as finding a good kernel $P(\cdot, \cdot)$ in the Gaussian process approach (50).

The algorithmic impact of all these choices is equivalent, and they lead to the same estimate (or posterior model). We have seen in the tables that the quality of the models depends quite a lot on these regularization matrices. Several regularization matrices can do very well in the model evaluations. So a natural question is: *Is there an optimal choice of regularization matrix for a certain true system θ_0?* Actually, there is, and we return to the classical perspective in Section 3 to analyze this.

We found in (24e) an expression for the MSE matrix. Let us first rewrite the expression (24e) using (41) that does not turn out to be ill-defined later on. Let

$$P_n = \sigma^2 D^{-1} \quad \text{and} \quad Q_0 = \theta_0\theta_0^T.$$

Then

$$(R_N + D)^{-1} = (P_n R_N + \sigma^2 I_n)^{-1} P_n$$

and

$$MSE(\hat{\theta}_N^R)(P_n) = (P_n R_N + \sigma^2 I_n)^{-1}(\sigma^2 P_n R_N P_n$$
$$+ \sigma^4 Q_0)(R_N P_n + \sigma^2 I_n)^{-1} \quad (54)$$

where we stressed how the MSE matrix depends on P_n for a given θ_0 (Q_0).

Again, is there a way to find a "best" P_n for a given system θ_0? We could first ask what the average MSE is if θ_0 is a random variable with zero mean and covariance $E\theta_0\theta_0^T = Q$. This average MSE is obtained by replacing Q_0 in (54) by Q, and corresponds in a Bayesian perspective to the MSE of the estimate $\hat{\theta}_N^{apost}$. Keeping in mind that $\hat{\theta}_N^{apost}$ is the posterior mean, we know from the Bayesian approach that the MSE of the estimate $\hat{\theta}_N^{apost}$ is minimized by picking the prior covariance P_n in (36) to be the true (prior) covariance Q of θ_0. Therefore, for a given θ_0, the choice

$$P_n = Q_0$$

minimizes (54), and we have the following result.

Theorem 1 (*Best Choice of P_n for Given θ_0*)**.** *The matrix in (54) obeys the following matrix inequality*

$$MSE(\hat{\theta}_N^R)(P_n) \geq MSE(\hat{\theta}_N^R)(\theta_0\theta_0^T), \quad \text{for any } P_n \geq 0. \quad (55)$$

Proof. An independent direct algebraic proof is given in Appendix. See also equation (14) in Eldar (2006).

That means that there is an optimal choice of regularization that is independent of both N and the input $u(t)$, that minimizes the MSE matrix in a matrix sense. Whatever the quadratic measure of fit for the model, the best choice of regularization is to use

$$P_n = \theta_0\theta_0^T \quad (56)$$

which yields the corresponding optimal regularized estimate

$$\hat{\theta}_N^{Opt} = (\theta_0\theta_0^T \Phi_N \Phi_N^T + \sigma^2 I_{N-n})^{-1}\theta_0\theta_0^T \Phi_N Y_N. \quad (57)$$

Let us try this regularization matrix on the data bank. $\quad\square$

Example 7 (*Best Regularization Compared to the Optimal One*)**.** Let us compute the estimates (57) corresponding to optimal regularization (56) and compare to the best performing ones (without the base-line model) in Examples 5 and 6.

	Best	Ideal
S1D1	90.8 (DC)	98.6
S2D1	78.0 (DC)	91.9
S1D2	72.7 (DC)	94.5
S2D2	60.8 (TC)	88.9

Findings. The optimal regularization (which requires system knowledge) clearly outperforms all the best choices from the previous sections.

We see that the performance of this regularization indeed is superior to all we have seen in the other tables. The figures in this example are indeed the upper bounds of what can be achieved for FIR models by regularization, Bayesian method and Gaussian process regression (both with zero prior mean). It is of independent interest to know such upper bounds.

The drawback is of course that the optimal choice (56) depends on the unknown system and cannot be used in practice.

As usual there are two approaches to this dilemma:

- Adaptive choice (choose P_n based on a preliminary estimate of θ).
- Robust choice (choose P_n as a min–max choice over a prior set of possible models).

It is an interesting topic for future research to find out how these approaches could be best implemented for more successful regularization.

7. Estimating the impulse response

Let us now sum up and consider the findings about question **(a)** in the introduction, to estimate the impulse of the unknown system, that has the best fit to the true impulse response.

The standard answer is to try the models (10) of different orders using PEM/ML methods, use the cross-validation in Section 3.6 to pick the best model order, and finally use the whole data record to estimate the model (10) with the best model order. Actually, the standard approach has been tested in Example 1 where its performance is shown in the column CV. Comparing the performance of the standard approach with that of the regularization methods based on the kernels/regularization matrices SS, TC and DC as shown in Examples 5 and 6 shows that
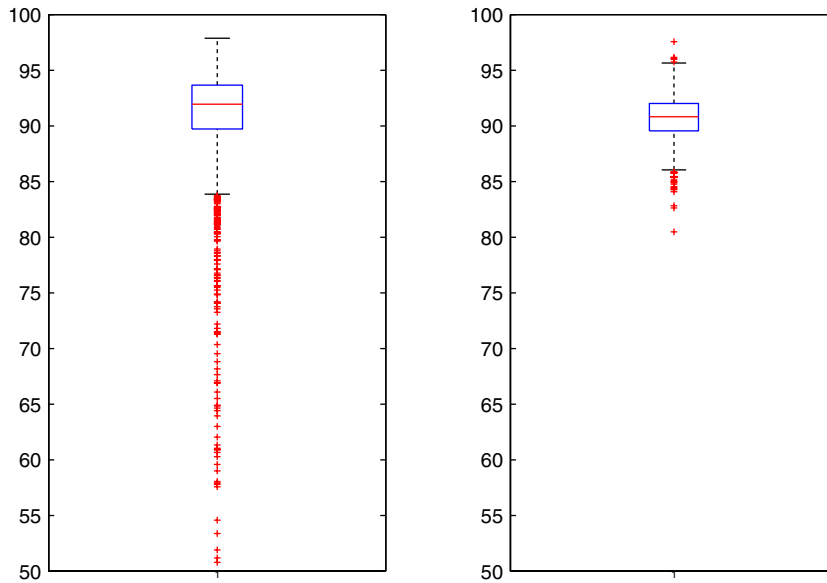
Fig. 1. Box plots of the 2500 fits for CV/S1D1 in Example 1 (left figure) and DC/S1D1 in Example 5 (right figure). The left figure has an additional 1.4% fits below 50.

the standard approach works rather well but the average fit can be improved.

Moreover, recall that each figure in the tables in Examples 1–6 is the average of 2500 fits. It is of course interesting to study the distribution of the fits over the different individual data sets. It turns out that the distributions in the CV column of Example 1 have better medians but long tails of poor fits, while the SS, TC and DC columns of Examples 5 and 6 are distributed much more compactly. For illustration, the box plots for the 2500 fits corresponding to CV/S1D1 in Example 1 and the DC/S1D1 in Example 5 are shown in Fig. 1. This observation indicates that the standard approach occasionally has problems and is actually less robust than the regularization methods based on the kernels/regularization matrices SS, TC and DC as shown in Sections 4 and 5.

8. Estimating a model of given order

Let us now turn to question **(b)** in the introduction, to find a model (10) of a given order, that has the best fit to the true impulse response.

The PEM/ML approach (11) has two good features, e.g. Ljung (1999):

(1) If the given model structure contains the true, unknown system, PEM/ML has the smallest possible variance (asymptotically) [among all unbiased estimates].
(2) If not, PEM/ML will converge, as $N \to \infty$, to the best possible approximation within the given structure.

Is there a catch? Yes, if the true system and model is of high order, the estimate will have rather high variance. It will be the smallest one possible for unbiased estimates, but just as shown in Section 3.1 it is conceivable that the MSE could be smaller if we allow some bias. There are several ways to achieve this. One would be to regularize the estimation criterion (11), just as in (23a). Another would be to use the best available impulse response estimate and fit it to the required model structure. That can be done by model reduction either by minimizing the L_2-fit, e.g. Tjärnström and Ljung (2002) or by balanced realization reduction (see `balred` in the System Identification Toolbox, Ljung (2007)), or any other model reduction technique.

Let us test the above methods on the data bank of data sets.

Example 8 (*Estimating Models of a Given Structure*)**.** We try the following methods:

- OE: m = oe(data, [n, n, 1])
- DC + BR:
    mf = DC(data, 125)
    m = balred(mf, n)

Here, the command `DC(data,125)` denotes the regularization method with the DC regularization (49b) (base-line model is not used). The average fit (5) is calculated and the simulation results are shown in the table below.

	$n = 2$	$n = 5$	$n = 10$	$n = 15$	$n = 20$	$n = 25$	$n = 30$
S1D1							
OE	57.4	86.3	89.2	86.4	81.5	74.2	61.5
DC + BR	28.2	83.0	90.6	90.8	90.8	90.8	90.8
S2D1							
OE	47.1	68.7	72.8	71.7	70.5	63.1	57.2
DC + BR	−47.0	53.2	73.5	76.2	77.0	77.4	77.6
S1D2							
OE	53.2	71.9	65.5	56.1	46.1	34.5	19.7
DC + BR	30.5	68.9	72.7	72.7	72.7	72.7	72.7
S2D2							
OE	40.2	50.8	43.0	42.3	30.7	20.5	10.5
DC + BR	−31.6	41.3	56.9	59.1	59.8	60.2	60.3

We also tried the regularization matrices/priors DI, TC and SS instead of DC, and they gave very similar or slightly inferior results.

Findings. For low order models (2nd and 5th order models), the PEM/ML method OE gives the best fit. In particular, the 5th order model gives a better fit than the 2nd order model. That's because the variance of the 5th order model is small enough so that further reducing the variance at the price of some bias gives a worse fit. In contrast, the balanced realization model reduction on the regularized FIR model has some difficulties for low order models. For higher order models, the PEM/ML method OE works badly due to the high variance caused by the increasing model flexibility. In contrast, the balanced realization model reduction on the regularized FIR model works rather well. That's because on the one hand the regularization can curb the flexibility and overcome the high variance, and on the other hand, the regularized FIR model allows good approximations with high order models. Therefore,

Table 1
The average fit (5) to test data bank as shown in Section 2 for the standard approach (cf. Example 1, CV column), the regularized FIR approach with the DC regularization (cf. Example 5, DC column), and the theoretical limit of the regularization (cf. Example 7, Ideal column).

	Standard	Reg. FIR (DC)	Opt. reg.
S1D1	89.4	90.8	98.6
S2D1	73.2	78.0	91.9
S1D2	70.8	72.7	94.5
S2D2	49.6	60.8	88.9

it is beneficial to first estimate a 125th order regularized FIR model and then reduce its order using balanced realization model reduction. We also see that the simple 5th order model for the OE gives not much worse performance than the best that can be achieved (within 5% for the "fast" systems and 10% for "slow" systems).

Remark 11. As mentioned in Section 2, the white noise input is used to simulate the 5000 systems. It is of course interesting to study the cases where the input is not white. We actually tested the case where the input is a band-limited Gaussian random signal. As a result, the balanced realization model reduction on the regularized FIR model gives a better fit than the PEM/ML method OE for the 5th order model. In contrast with the table in Example 8, the PEM/ML method OE has a significant drop in the fits for most cases, especially for model orders greater than 2. The balanced realization model reduction on the regularized FIR model gives a little bit worse fits and shows very good robustness. Nevertheless, for the band-limited Gaussian random input signal, the same finding as in Example 8 can be drawn. For low order models (2nd order model), the PEM/ML method OE gives the best fit. For high order models, the balanced realization model reduction on the regularized FIR model is preferred.

9. Conclusions

So, let us return to the two questions posed in the introduction and summarize our findings.

The first question is to estimate the impulse response of a linear system as well as possible. We have tried two basic techniques for that:

- The "standard" approach: estimate parametric models of different "sizes" by PEM/ML techniques and choose the model size by cross validation.
- A regularized FIR model approach: estimate a high order FIR model and regularize the estimate to suitably curb flexibility. We have reported several interpretations and paradigms for this approach, and a key feature is to select the regularization matrix (priors or kernels).

For convenience we repeat in Table 1 the bottom line of the results of these two techniques, when applied to the test data bank, described in Section 2. We have also tried a hybrid version of the two approaches by adding a second order base-line model. This gives a touch better fit. (See Example 5.) A box plot for the 2500 figures behind the first two entries in the table was given in Fig. 1.

The conclusion is that for the test data bank that has relatively short data records and high order systems, the regularization method with carefully chosen regularization matrices shows both better accuracy and robustness than the standard approach.

We have seen that the results depend significantly on the choice of regularization matrix, and how well it is tuned. That raises the question of how far regularization can bring us. The theoretical limits in the last column of Table 1 are therefore of interest. They show a substantial potential, but it is not clear how much of it can be achieved with practical algorithms.

The second question is to estimate a model (10) of given order that has an impulse response as close as possible to the unknown system. We have also tried two basic techniques for that:

- The "standard" approach: estimate model (10) by PEM/ML techniques.
- A regularized FIR model approach, together with a model reduction technique: fit a well estimated regularized FIR model to (10) by some model reduction techniques.

The conclusion is that for the data and systems in the test data bank, a low order model is often better estimated by the standard approach; a higher order model is often better estimated by model reduction on a high order regularized FIR model with careful regularization.

Appendix. Proof of Theorem 1

Define

$$M = -(P_n R + I_n)^{-1} \quad \text{and} \quad M_0 = -(Q_0 R + I_n)^{-1} \tag{A.1}$$

where for convenience we have let $R = \sigma^{-2} R_N$ and $Q_0 = \theta_0 \theta_0^T$. With (A.1), (55) can be rewritten as

$$M(P_n R P_n + Q_0) M^T \geq M_0 (Q_0 R Q_0 + Q_0) M_0^T. \tag{A.2}$$

Note that

$$I + M = -M P_n R, \qquad I + M_0 = -M_0 Q_0 R \tag{A.3}$$

thus (55) can be further rewritten as

$$(I + M) R^{-1} (I + M)^T + M Q_0 M^T$$
$$\geq (I + M_0) R^{-1} (I + M_0)^T + M_0 Q_0 M_0^T. \tag{A.4}$$

In the following, we show that

$$(I + M) R^{-1} (I + M)^T + M Q_0 M^T$$
$$\quad - (I + M_0) R^{-1} (I + M_0)^T - M_0 Q_0 M_0^T$$
$$= (M - M_0)(R^{-1} + Q_0)(M - M_0)^T. \tag{A.5}$$

Simple calculation shows that (A.5) is equivalent to

$$(I + M_0) R^{-1} M^T + M R^{-1} (I + M_0^T)$$
$$\quad - (I + M_0) R^{-1} M_0^T - M_0 R^{-1} (I + M_0^T)$$
$$= 2 M_0 Q_0 M_0^T - M_0 Q_0 M^T - M Q_0 M_0^T. \tag{A.6}$$

It follows from the second equation of (A.3) that

$$(I + M_0) R^{-1} = -M_0 Q_0. \tag{A.7}$$

Now inserting (A.7) into the left hand side of (A.6) shows that (A.6) and thus (A.5) holds. Moreover, since $(M - M_0)(R^{-1} + Q_0)(M - M_0)^T$ in (A.5) is positive semi-definite, Eq. (A.4) holds as well, which in turn implies (55) holds. So this completes the proof.

Remark 12. As mentioned in Remark 4, the bias of the regularized estimate $\hat{\theta}_N^R$ is linear in θ_0. Note from (23b) that

$$\hat{\theta}_N^R = (I + M)\hat{\theta}_N^{LS} \tag{A.8}$$

so the bias is equal to $M\theta_0$ and thus M corresponds to the bias gradient matrix in (14) of Eldar (2006).

References

Carlin, B. P., & Louis, T. A. (1996). *Bayes and empirical Bayes methods for data analysis*. London: Chapman & Hall.

Eldar, Y. C. (2006). Uniformly improving the Cramér–Rao bound and maximum-likelihood estimation. *IEEE Transactions on Signal Processing*, 54(8), 2943–2956.

Goodwin, G. C., Braslavsky, J. H., & Seron, M. M. (2002). Non-stationary stochastic embedding for transfer function estimation. *Automatica*, 38, 47–62.

Goodwin, G. C., Gevers, M., & Ninness, B. (1992). Quantifying the error in estimated transfer functions with application to model order selection. *IEEE Transactions Automatic Control*, 37(7), 913–929.

Gustafsson, F., & Hjalmarsson, H. (1995). Twenty-one ML estimators for model selection. *Automatica*, 31(10), 1377–1392.

Ljung, L. (1985). On the estimation of transfer functions. *Automatica*, 21(6), 677–696.

Ljung, L. (1999). *System identification — theory for the user* (2nd ed.). Upper Saddle River, N.J: Prentice-Hall.

Ljung, L. (2007). *System identification toolbox for use with* MATLAB. *Version 7* (7th ed.). Natick, MA: The MathWorks, Inc.

Ljung, L., & Wahlberg, B. (1992). Asymptotic properties of the least-squares method for estimating transfer functions and disturbance spectra. *Advances in Applied Probability*, 24, 412–440.

Pillonetto, G., Chiuso, A., & De Nicolao, G. (2010). Regularized estimation of sums of exponentials in spaces generated by stable spline kernels. In *American control conference* (pp. 498–503).

Pillonetto, G., Chiuso, A., & Nicolao, G. D. (2011). Prediction error identification of linear systems: a nonparametric Gaussian regression approach. *Automatica*, 47(2), 291–305.

Pillonetto, G., & Nicolao, G. D. (2010a). A new kernel-based approach for linear system identification. *Automatica*, 46(1), 81–93.

Pillonetto, G., & Nicolao, G.D. (2010b). The stable spline toolbox for system identification, Technical report. University of Padova, Padova, Italy.

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian processes for machine learning*. Cambridge, MA: MIT Press.

Tikhonov, A. N., & Arsenin, V. Y. (1977). *Solutions of Ill-posed problems*. Washington, D.C: Winston/Wiley.

Tjärnström, F., & Ljung, L. (2002). L-2 model reduction and variance reduction. *Automatica*, 38, 1517–1530.

Tianshi Chen was born in China in November 1978. He received the M.S. degree in July 2005 from the Harbin Institute of Technology and the Ph.D. degree in December 2008 from the Chinese University of Hong Kong. From April 2009 to March 2011, he was a postdoctoral researcher at the Automatic Control group of Linköping University, Linköping, Sweden. His research interests are mainly within the areas of system identification, statistical signal processing, and nonlinear control theory and applications. He is currently an Assistant Professor at Linköping University.

Henrik Ohlsson was born in Sweden in 1981. He received the M.Sc. degree in Applied Physics and Electrical Engineering in Oct. 2006 and his Ph.D. degree in Automatic Control in Nov. 2010, all from Linköping University, Sweden. He has held visiting positions at the University of Cambridge and the University of Massachusetts. His research interests are mainly within the areas of system identification and machine learning. Henrik is currently a visiting scholar at the University of California at Berkeley and an Assistant Professor at Linköping University.

Lennart Ljung received his Ph.D. in Automatic Control from Lund Institute of Technology in 1974. Since 1976 he is Professor of the chair of Automatic Control In Linkoping, Sweden, and is currently Director of the Strategic Research Center "Modeling, Visualization and Information Integration" (MOVIII). He has held visiting positions at Stanford and MIT and has written several books on System Identification and Estimation. He is an IEEE Fellow, an IFAC Fellow and an IFAC Advisor as well as a member of the Royal Swedish Academy of Sciences (KVA), a member of the Royal Swedish Academy of Engineering Sciences (IVA), an Honorary Member of the Hungarian Academy of Engineering and a Foreign Associate of the US National Academy of Engineering (NAE). He has received honorary doctorates from the Baltic State Technical University in St Petersburg, from Uppsala University, Sweden, from the Technical University of Troyes, France, from the Catholic University of Leuven, Belgium and from Helsinki University of Technology, Finland. In 2002 he received the Quazza Medal from IFAC, in 2003 he received the Hendrik W. Bode Lecture Prize from the IEEE Control Systems Society, and he was the recipient of the IEEE Control Systems Award for 2007.