Automatica 50 (2014) 657–682

Contents lists available at ScienceDirect

Automatica

journal homepage: www.elsevier.com/locate/automatica

Survey Paper

Kernel methods in system identification, machine learning and function estimation: A survey☆

Gianluigi Pillonetto [a,1], Francesco Dinuzzo [b], Tianshi Chen [c], Giuseppe De Nicolao [d], Lennart Ljung [c]

[a] Department of Information Engineering, University of Padova, Padova, Italy
[b] Max Planck Institute for Intelligent Systems, Tübingen, Germany
[c] Division of Automatic Control, Linköping University, Linköping, Sweden
[d] Department of Computer Engineering and Systems Science, University of Pavia, Pavia, Italy

ARTICLE INFO

Article history:
Received 16 November 2011
Received in revised form
13 January 2014
Accepted 15 January 2014
Available online 25 February 2014

Keywords:
Linear system identification
Prediction error methods
Model complexity selection
Bias-variance trade-off
Kernel-based regularization
Inverse problems
Reproducing kernel Hilbert spaces
Gaussian processes

ABSTRACT

Most of the currently used techniques for linear system identification are based on classical estimation paradigms coming from mathematical statistics. In particular, maximum likelihood and prediction error methods represent the mainstream approaches to identification of linear dynamic systems, with a long history of theoretical and algorithmic contributions. Parallel to this, in the machine learning community alternative techniques have been developed. Until recently, there has been little contact between these two worlds. The first aim of this survey is to make accessible to the control community the key mathematical tools and concepts as well as the computational aspects underpinning these learning techniques. In particular, we focus on kernel-based regularization and its connections with reproducing kernel Hilbert spaces and Bayesian estimation of Gaussian processes. The second aim is to demonstrate that learning techniques tailored to the specific features of dynamic systems may outperform conventional parametric approaches for identification of stable linear systems.

1. Preamble

System identification is about building mathematical models of dynamic systems from observed input–output data. It is a well established subfield of Automatic Control, with more than 50 years history of theoretical and algorithmic development as well as software packages and industrial applications.

General aspects. For time-invariant linear dynamical systems the output is obtained as a convolution between the input and the system's impulse response. This means that system identification is an example of an *inverse problem*: indeed, finding the impulse response from observed data is a *deconvolution problem*. Such problems are quite ubiquitous and appear in biology, physics, and engineering with applications e.g. in medicine, geophysics, and image restoration (Bertero, 1989; De Nicolao, Sparacino, & Cobelli, 1997; Hunt, 1970; Tarantola, 2005). The problem is non trivial as convolution is a continuous operator, e.g. on the space of square integrable functions, but its inverse may not exist or may be unbounded (Phillips, 1962).

The reconstruction of the continuous-time impulse response is always an ill-posed problem since such a function cannot be uniquely inferred from a finite set of observations. Also finite discretizations lead to an ill-conditioned problem, meaning that small errors in the data can lead to large estimation errors. Starting from the seminal works of Tikhonov and Phillips (Phillips, 1962; Tikhonov & Arsenin, 1977), a number of regularization methods have been proposed in the literature to solve the deconvolution problem, e.g. truncated singular value decompositions (Hansen, 1987) and gradient-based techniques (Hanke, 1995; Nemirovskii, 1986; Yao, Rosasco, & Caponnetto, 2007). This means that

☆ This research has been partially supported by the European Community under agreement no. FP7-ICT-223866-FeedNetBack, no. 257462-HYCON2 Network of excellence, by the MIUR FIRB project RBFR12M3AC-Learning meets time: a new computational approach to learning in dynamic systems as well as by the Linnaeus Center CADICS, funded by the Swedish Research Council, and the ERC advanced grant LEARN, no. 287381, funded by the European Research Council. The material in this paper was not presented at any conference. This paper was recommended for publication in revised form by Editor John Baillieul.

E-mail addresses: giapi@dei.unipd.it (G. Pillonetto), fdinuzzo@tuebingen.mpg.de (F. Dinuzzo), tschen@isy.liu.se (T. Chen), giuseppe.denicolao@unipv.it (G. De Nicolao), ljung@isy.liu.se (L. Ljung).

[1] Tel.: +39 3481212375; fax: +39 049 827 7699.

http://dx.doi.org/10.1016/j.automatica.2014.01.001
0005-1098/

regularization should be an important topic and area for system identification.

Identification techniques. The most widespread approach to identification of dynamic systems relies on parametric prediction error methods (PEMs), for which a large corpus of theoretical results is available (Ljung, 1999; Söderström & Stoica, 1989). The statistical properties of prediction error (and maximum likelihood) methods are well understood under the assumption that the model class is fixed. They show that such procedures are in some sense optimal, at least for large samples. However, within this parametric paradigm, a key point is the selection of the most adequate model structure. In the "classical, frequentist" framework, this is a question of trade-off between bias and variance, and can be handled by various model validation techniques. This is often carried out by resorting to complexity measures, such as the Akaike's criterion (AIC) (Akaike, 1974) or cross validation (CV), but some inefficiencies related to these classical approaches have been recently pointed out (Chen, Ohlsson, & Ljung, 2012; Pillonetto, Chiuso, & De Nicolao, 2011; Pillonetto & De Nicolao, 2010). In particular, it has been shown that sample properties of PEM approaches, equipped e.g. with AIC or CV, may be unsatisfactory when tested on experimental data, departing sharply from the properties predicted by standard (i.e. without model selection) statistical theory, which suggests that PEM should be asymptotically efficient for Gaussian innovations.

Parallel to this development in system identification, other techniques have been developed in the machine learning community. Until very recently, there has been little contact between these concepts and system identification.

Recent research has shown that the model selection problems can be successfully faced by a different approach to system identification that leads to an interesting cross fertilization with the machine learning field (Pillonetto & De Nicolao, 2010). Rather than postulating finite-dimensional hypothesis spaces, e.g. using ARX, ARMAX or Laguerre models, the new paradigm formulates the problem as function estimation possibly in an infinite-dimensional space. In the context of linear system identification, the elements of such space are all possible impulse responses. The intrinsical ill-posedness of the problem is circumvented using regularization methods that also admit a Bayesian interpretation (Rasmussen & Williams, 2006). In particular, the impulse response is modeled as a zero-mean Gaussian process. In this way, prior information is introduced in the identification process just assigning a covariance, named also kernel in the machine learning literature (Schölkopf & Smola, 2001). In view of the increasing importance of these kernel methods also in the general system identification scenario, the first aim of this survey is to make accessible to the control community some of the key mathematical tools and concepts underlying these learning techniques, e.g. reproducing kernel Hilbert spaces (Aronszajn, 1950; Cucker & Smale, 2001; Saitoh, 1988), kernel methods and regularization networks (Evgeniou, Pontil, & Poggio, 2000; Suykens, Gestel, Brabanter, De Moor, & Vandewalle, 2002; Vapnik, 1998), the representer theorem (Schölkopf, Herbrich, & Smola, 2001; Wahba, 1990) and the connection with the theory of Gaussian processes (Hengland, 2007; Rasmussen & Williams, 2006). It is also pointed out that a straight application of these techniques in the control field is doomed to fail unless some key features of the system identification problem are taken into account. First, as already recalled, the relationship between the unknown function and the measurements is not direct, as typically assumed in the machine learning setting, but instead indirect, through the convolution with the system input. This raises significant analogies with the literature on inverse problems (Bertero, 1989; Tikhonov & Arsenin, 1977). Furthermore, in system identification it is essential that the estimation process be informed on the stability of the impulse response. In this regard, a recent major advance has been the introduction of new kernels which include information on impulse response exponential stability (Chen et al., 2012; Pillonetto & De Nicolao, 2010). These kernels depend on some hyperparameters which can be estimated from data e.g. using marginal likelihood maximization. This procedure is interpretable as the counterpart of model order selection in the classical PEM paradigm but, as it will be shown in the survey, it turns out to be much more robust, appearing to be the real reason of success of these new procedures. Other research directions recently developed have been the justification of the new kernels in terms of Maximum Entropy arguments (Pillonetto & De Nicolao, 2011), the analysis of these new approaches in a classical deterministic framework leading to the derivation of the *optimal kernel* (Chen et al., 2012), as well as the extension of these new techniques to the estimation of optimal predictors (Pillonetto et al., 2011).

Outline of the survey. The present survey will deal with this meeting between conventional system identification of linear models and learning techniques. It is divided into three Parts with sections which are relevant, but can be skipped without interrupting the flow of the discussion, marked with a star ⋆.

Part I will describe the status in traditional parametric system identification in discrete-time with an account of how the bias-variance trade-off can be handled also by *regularization techniques*, including their Bayesian interpretation.

Part II is an account of general function estimation – or function learning – theory in a general and abstract setting. This includes the role of RKHS theory for this problem.

Part III treats linear system identification, mainly in continuous-time, as an application of learning the impulse response function from observed data, leaning on general function estimation and its adaptation to the specific properties of impulse responses of dynamic systems. This will link back to the regularizations techniques from the simplistic perspective in Part I. Considerations on computational issues are also included while some mathematical details are gathered in the Appendix.

In conclusion, the scope of this work is twofold. Firstly, our aim is to survey essential results in kernel methods for estimation, that are mostly published outside the control audience, and hence not so well known in this community. Secondly, we want to show that these results have much to offer for estimation problems in the control community, in particular for system identification.

Part I. Estimating system impulse responses in discrete time

In this part, we study the problem of estimating system impulse responses in discrete time.

2. "Classical" system identification

2.1. System identification

There is a very extensive literature on system identification, with many text books, like Ljung (1999) and Pintelon and Schoukens (2012a). Most of the techniques for system identification have their origins in estimation paradigms from mathematical statistics, and classical methods like Maximum Likelihood (ML) have been important elements in the area. In Part I the main ingredients of this "classical" view of system identification will be reviewed. For convenience, we will only focus on the single input–single output (SISO) linear time-invariant, stable and causal systems. We will also set the stage for the "kernel methods" for estimating the main characteristics of a system.

2.2. Parametric model structures

A model structure $\mathscr{M}$ is a parameterized collection of models that describe the relations between the input and output signal of the system. The parameters are denoted by θ so $\mathscr{M}(\theta)$ is a particular model. That model gives a rule to predict (one-step-ahead) the output at time t, i.e. $y(t)$, based on observations of previous input–output data up to time $t-1$ (denoted by $\mathscr{Z}^{t-1}$).

$$\hat{y}(t|\theta) = g(t, \theta, \mathscr{Z}^{t-1}). \tag{1}$$

For linear systems, a general model structure is given by the transfer function G from input to output and the transfer function H from a white noise source e to output additive disturbances:

$$y(t) = G(q, \theta)u(t) + H(q, \theta)e(t) \tag{2a}$$

$$\mathscr{E}e^2(t) = \sigma^2; \qquad \mathscr{E}e(t)e(k) = 0 \quad \text{if } k \neq t \tag{2b}$$

where $\mathscr{E}$ denotes mathematical expectation. This model is in discrete time and q denotes the shift operator $qy(t) = y(t+1)$. We assume for simplicity that the sampling interval is one time unit. The expansion of $G(q, \theta)$ and $H(q, \theta)$ in the inverse (backwards) shift operator gives the *impulse responses* of the two systems:

$$G(q, \theta) = \sum_{k=1}^{\infty} g_k(\theta)q^{-k} \tag{3}$$

$$H(q, \theta) = h_0(\theta) + \sum_{k=1}^{\infty} h_k(\theta)q^{-k} \tag{4}$$

where for normalization reasons, we assume $h_0(\theta) = 1$.

Under the assumption that $H(q, \theta)$ is inversely stable, see Ljung (1999, p. 64), the natural one-step-ahead predictor for (2a) is

$$\hat{y}(t|\theta) = \frac{H(q, \theta) - 1}{H(q, \theta)}y(t) + \frac{G(q, \theta)}{H(q, \theta)}u(t). \tag{5}$$

Since $h_0(\theta) = 1$, the numerator in the two terms starts with $h_1(\theta)q^{-1}$ and $g_1(\theta)q^{-1}$, respectively. So there is a delay in both y and u. The question is how to parameterize G and H.

Black-box models. Common *black box* (i.e. no physical insight or interpretation) parameterizations are to let G and H be rational in the shift operator:

$$G(q, \theta) = \frac{B(q)}{F(q)}; \qquad H(q, \theta) = \frac{C(q)}{D(q)} \tag{6}$$

where $B(q)$, $F(q)$, $C(q)$ and $D(q)$ are polynomials of q^{-1}.

A very common case is that $F(q) = D(q)$ and $C(q) = 1$ which gives the *ARX-model*:

$$y(t) = \frac{B(q)}{A(q)}u(t) + \frac{1}{A(q)}e(t), \quad \text{or}$$
$$A(q)y(t) = B(q)u(t) + e(t) \tag{7}$$

where $A(q) = 1 + a_1q^{-1} + \cdots + a_{n_a}q^{-n_a}$, $B(q) = b_1q^{-1} + \cdots + b_{n_b}q^{-n_b}$, and n_a, n_b are positive integers referred to as the orders of ARX-model.

Other common black/box structures of this kind are FIR-model (Finite Impulse Response model, $F(q) = C(q) = D(q) = 1$), OE-model (Output Error model, $C(q) = D(q)$), ARMAX-model ($F(q) = D(q) = A(q)$), and BJ-model (Box–Jenkins, all four polynomials different).

Approximating linear systems by ARX models. Suppose the true linear system is given by

$$y(t) = G_0(q)u(t) + H_0(q)e(t). \tag{8}$$

Suppose we build an ARX model (7) for high orders n_a and n_b. Then it is well known from Ljung and Wahlberg (1992) that, as n_a and n_b tend to infinity at the same time as the number of data N increases even faster, we have for the ARX-model estimate $\hat{A}_{n_a}(q)$ and $\hat{B}_{n_b}(q)$:

$$\frac{\hat{B}_{n_b}(q)}{\hat{A}_{n_a}(q)} \to G_0(q), \qquad \frac{1}{\hat{A}_{n_a}(q)} \to H_0(q) \quad \text{as } n_a, n_b \to \infty. \tag{9}$$

This is quite a useful result. ARX-models are easy to estimate. The estimates are calculated by linear least squares (LS) techniques, which are convex and numerically robust. Estimating a high order ARX model, possibly followed by some model order reduction, could thus be an alternative to the numerically more demanding general PEM criterion minimization (12) introduced in the next subsection. This has been extensively used, e.g. by Zhu (1989) and Zhu and Backx (1993). The only drawback with high order ARX-models is that they may suffer from high variance. That is the problem we will turn to in Section 4.

2.3. Fitting time-domain data

Suppose we have collected a data record in the time domain

$$\mathscr{Z} = \{u(1), y(1), \ldots, u(N), y(N)\}. \tag{10}$$

It is most natural to compare the model predicted values (5) with the actual outputs and form the criterion of fit

$$V_N(\theta) = \sum_{t=1}^{N} \big(y(t) - \hat{y}(t|\theta)\big)^2 \tag{11}$$

and define the parameter estimate

$$\hat{\theta}_N = \arg\min_{\theta} V_N(\theta). \tag{12}$$

In (12), and also in the sequel, if multiple solutions to the optimization problem exist, the equality has to be interpreted as a set inclusion. We call this the *Prediction Error Method, PEM*. It coincides with the *Maximum Likelihood, ML,* method if the noise source e is Gaussian. See, e.g. Ljung (1999) or Ljung (2002) for more details.

2.4. Bias and variance

The observations of the system output are certainly affected by noise and disturbances, which of course also will influence the estimated model (12). The disturbances are typically described as stochastic processes, which makes the estimate $\hat{\theta}_N$ a *random variable*. This has a certain probability distribution function and a mean and a variance. The difference between the mean and a true description of the system measures the *bias* of the model. If the mean coincides with the true parameters, the estimator is said to be *unbiased*. The total error in a model thus has two contributions: the bias and the variance.

A very attractive property of the PEM estimate (for Gaussian noise source) is that it is *asymptotically efficient* provided the model structure $\mathscr{M}$ contains a correct description of the true system. That means that as $N \to \infty$ the covariance matrix of $\hat{\theta}_N$ will approach the Cramér–Rao limit, so that no unbiased estimate can be better than the PEM estimate.

Generally speaking the model quality depends on the quality of the measured data and the flexibility of the chosen model structure (1). A more flexible structure typically has smaller bias, since it is easier to come closer to the true system. At the same time, it will have a higher variance: higher flexibility makes easier to be fooled by disturbances. So the trade-off between bias and variance to reach a small total error is a choice of balanced flexibility of the model structure.

3. Selection of model flexibility: AIC, BIC, CV

3.1. Adjusting the estimation criterion

With increasing flexibility, the fit to the estimation data in (12), i.e. $V_N(\hat{\theta}_N)$, will always improve, since the bad effect of the variance is not visible in that fit. To account for that it is necessary to add a penalty for model flexibility to assess the total quality of the model. A common technique for this is Akaike's criterion, see e.g. Akaike (1974) and Ljung (1999, pp. 505–507). Letting $\dim(\theta) = m$ and assuming the noise to be Gaussian, with unknown variance, one has

$$\hat{\theta}_N = \arg\min_{\theta} \left[\log(V_N(\theta)) + 2\frac{m}{N}\right], \quad \text{AIC} \tag{13}$$

where the minimization also takes place over a family of model structures with different number m of parameters. There is also a small-sample version, described in Hurvich and Tsai (1989) and known in the literature as corrected Akaike's criterion (AICc), defined by

$$\hat{\theta}_N = \arg\min_{\theta} \left[\log(V_N(\theta)) + 2\frac{m}{(N-m-1)}\right], \quad \text{AICc}. \tag{14}$$

Another variant places a larger penalty on the model flexibility:

$$\hat{\theta}_N = \arg\min_{\theta} \left[\log(V_N(\theta)) + \log(N)\frac{m}{N}\right], \quad \text{BIC, MDL}. \tag{15}$$

This is known as Bayesian information criterion (BIC), or Rissanen's Minimum Description Length (MDL) criterion, see e.g. Ljung (1999, pp. 505–507), Rissanen (1978) and Schwarz (1978).

3.2. Cross validation

Another important technique is known as *cross validation*, CV. This is certainly to be regarded among the most widely employed approaches to statistical model selection. The goal is to obtain an estimate of the prediction capability of future data of the model in correspondence with different choices of θ. Parameter selection is thus performed by optimizing the estimated prediction score. *Holdout validation* is the simplest form of CV: the available data are split in two parts, where one of them (*estimation set*) is used to estimate the model, and the other one (*validation set*) is used to assess the prediction capability. By ensuring independence of the model fit from the validation data, the estimate of the prediction performance is approximately unbiased. For models that do not require estimation of initial states, like FIR and ARX models, CV can be applied efficiently in more sophisticated ways by splitting the data into more portions, as described in Section 14.3.

4. Regularization of linear regression models

ARX-models, introduced in (7), belong to the class of well-known and much-used linear regression models. Before we look closer into properties of ARX-model estimates, it is useful to consider linear regression models in more general terms.

4.1. Linear regression models

A *linear regression model* has the form

$$y(t) = \varphi^T(t)\theta + e(t), \quad \theta \in \mathbb{R}^m. \tag{16}$$

Here y (the output) and φ (the regression vector) are observed variables, e is a noise disturbance and θ is the unknown parameter vector. In general $e(t)$ is assumed to be independent of $\varphi(t)$.

It is convenient to rewrite (16) in vector form, by stacking all the elements (rows) in $y(t)$ and $\varphi^T(t)$ to form the vectors (matrices) Y and Φ and obtain

$$Y = \Phi\theta + E. \tag{17}$$

The LS estimate of the parameter θ is

$$\hat{\theta} = \arg\min_{\theta} \|Y - \Phi\theta\|^2 \tag{18a}$$

$$= (\Phi^T\Phi)^{-1}\Phi^T Y \tag{18b}$$

where $\|\cdot\|$ represents the Euclidean norm. From (18a) to (18b), we have implicitly assumed that $\Phi^T\Phi$ is nonsingular. In many cases, like when $\Phi^T\Phi$ is singular or ill-conditioned, it makes sense to introduce a regularization term in (18a) by means of a regularization matrix P and consider the regularized least squares instead.

4.2. Regularized least squares

In order to *regularize* the estimate, we add a regularization term $\theta^T P^{-1}\theta$ in (18a) and obtain the following problem, often referred to as *regularized least squares* (ReLS):

$$\hat{\theta} = \arg\min_{\theta} \|Y - \Phi\theta\|^2 + \gamma\theta^T P^{-1}\theta \tag{19a}$$

$$= P\Phi^T(\Phi P\Phi^T + \gamma I_N)^{-1}Y; \quad \text{or} \tag{19b}$$

$$= (P\Phi^T\Phi + \gamma I_m)^{-1}P\Phi^T Y \tag{19c}$$

where γ is a positive scalar and I_m denotes the m-dimensional identity matrix.[2]

Remark 1. When P is singular, (19a) is not well-defined. In this case, consider the singular value decomposition of P: $P = [U_1\ U_2]\begin{bmatrix}\Lambda_P & 0\\ 0 & 0\end{bmatrix}\begin{bmatrix}U_1 & U_2\end{bmatrix}^T$ where Λ_P is a diagonal matrix with all diagonal elements being positive singular values of P and $U = \begin{bmatrix}U_1 & U_2\end{bmatrix}$ is an orthogonal matrix with U_1 having the same number of columns as Λ_P. Then (19a) should be interpreted as

$$\hat{\theta} = \arg\min_{\theta} \|Y - \Phi\theta\|^2 + \gamma\theta^T U_1\Lambda_P^{-1}U_1^T\theta \tag{20a}$$

$$\text{s.t. } U_2^T\theta = 0. \tag{20b}$$

It is easy to verify that (19b) or (19c) is still the optimal solution of (20). For convenience, we will use (19a) in the sequel and refer to (20) for its rigorous meaning for singular P.

The positive scalar γ is the so called regularization parameter which has to balance adherence to experimental data and the penalty term $\theta^T P^{-1}\theta$. This latter will improve the numerical properties of the estimator and decrease its variance, at the price of introducing some bias. To evaluate the model quality in a "classic" or "frequentist" setting, suppose that the data have been generated by (17) for a certain "true" vector θ_0 with noise E with variance $\mathscr{E}EE^T = \sigma^2 I_N$. Then, the mean square error (MSE) of the estimator $\hat{\theta}$ is

$$\mathscr{E}[(\hat{\theta} - \theta_0)(\hat{\theta} - \theta_0)^T] = \sigma^2\left(\frac{P\Phi^T\Phi}{\gamma} + I_m\right)^{-1} \times \left(\frac{P\Phi^T\Phi P}{\gamma^2} + \frac{\theta_0\theta_0^T}{\sigma^2}\right)\left(\frac{\Phi^T\Phi P}{\gamma} + I_m\right)^{-1}. \tag{21}$$

[2] Note that the step from (19b) to (19c) follows from the simple matrix equality $A(I_j + BA)^{-1} = (I_k + AB)^{-1}A$ which holds for every $k \times j$ matrix A and $j \times k$ matrix B.

A rational choice of P and γ is one that makes this MSE matrix small in some sense. How shall we think of good such choices? It is useful to first establish the following Lemma of algebraic nature.

Lemma 2. *Consider the matrix*

$$M(Q) = (QR + I)^{-1}(QRQ + Z)(RQ + I)^{-1} \tag{22}$$

where Q, R and Z are positive semidefinite matrices. Then for all Q

$$M(Z) \preceq M(Q) \tag{23}$$

where (23) means that $M(Q) - M(Z)$ is positive semidefinite.

The proof consists of straightforward calculations, see Chen et al. (2012).

Noting the expression of (21) and invoking Lemma 2, the question what P and γ give the best MSE of the regularized estimate has a clear answer: the equation $\sigma^2 P = \gamma\theta_0\theta_0^T$ needs to be satisfied. Thus, the following result holds.

Proposition 3 (*Optimal Regularizer for a Given θ_0*)**.** *Letting $\gamma = \sigma^2$, the regularization matrix*

$$P = \theta_0\theta_0^T \tag{24}$$

minimizes, in the sense of (23), the MSE matrix (21).

So, not surprisingly the best regularization depends on the unknown system.

Note that the MSE (21) is linear in $\theta_0\theta_0^T$. That means that if we compute the ReLS estimate with the same P for a collection of true systems θ_0, the average MSE over that collection will be given by (21) with $\theta_0\theta_0^T$ replaced by its average over the collection. In particular, if θ_0 is a random vector with $\mathscr{E}\theta_0\theta_0^T = \Pi$, we obtain the following result:

Proposition 4. *Consider (19a) with $\gamma = \sigma^2$. Then, the best average (expected) MSE for a random true system θ_0 with $\mathscr{E}\theta_0\theta_0^T = \Pi$ is obtained by the regularization matrix $P = \Pi$.*

With this we are very close to a Bayesian interpretation.

4.3. Bayesian interpretation

The following well known and simple result about conditioning jointly Gaussian random variable is a key element in Bayesian calculations. Let $x \sim \mathcal{N}(m, P)$ denote a Gaussian random variable with mean m and covariance matrix P, and consider

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\right). \tag{25a}$$

Then the conditional distribution of x_1 given x_2 is

$$x_1|x_2 \sim \mathcal{N}(\mu, \Sigma) \tag{25b}$$

$$\mu = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2) \tag{25c}$$

$$\Sigma = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}. \tag{25d}$$

In the current setup, we regard the vector θ as a random variable, say of Gaussian distribution with zero mean and covariance matrix Π, i.e. $\theta \sim \mathcal{N}(0, \Pi)$. In addition, let $e(t)$ in (16) be Gaussian, independent of θ, with zero mean and variance σ^2. Then, we have (17), with known Φ and $E \sim \mathcal{N}(0, \sigma^2 I_N)$. Hence, Y and θ will be jointly Gaussian variables:

$$\begin{bmatrix} \theta \\ Y \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Pi & \Pi\Phi^T \\ \Phi\Pi & \Phi\Pi\Phi^T + \sigma^2 I_N \end{bmatrix}\right). \tag{26}$$

The posterior distribution of θ given Y follows from (25)

$$\theta|Y \sim \mathcal{N}(\hat{\theta}, \Pi^*) \tag{27a}$$

$$\hat{\theta} = \Pi\Phi^T(\Phi\Pi\Phi^T + \sigma^2 I_N)^{-1}Y \tag{27b}$$

$$= (\Pi\Phi^T\Phi + \sigma^2 I_m)^{-1}\Pi\Phi^T Y \tag{27c}$$

$$\Pi^* = \Pi - \Pi\Phi^T(\Phi\Pi\Phi^T + \sigma^2 I_N)^{-1}\Phi\Pi. \tag{27d}$$

So this shows that the regularized LS estimate (19) is the mean of the posterior distribution (the MAP estimate) provided we choose $\gamma = \sigma^2$ and the regularization matrix P as the prior covariance matrix of θ, i.e., with $P = \Pi$. Note also that the Bayesian interpretation permits to compute uncertainty bounds around $\hat{\theta}$ using the posterior covariance Π^* given by (27d).

4.4. Tuning the regularization: marginal likelihood maximization

Can we estimate this matrix $P = \Pi$ in some way? Let P be parameterized in terms of the so-called *hyperparameters* $\eta \in \Gamma$, $P(\eta)$. Now, the bottom row of (26) states that Y is a Gaussian random vector with zero mean and covariance matrix

$$Z(\eta) = \Phi P(\eta)\Phi^T + \sigma^2 I_N. \tag{28}$$

Apart from constants, two times the negative logarithm of the probability density function of the Gaussian random vector Y is $Y^T Z(\eta)^{-1}Y + \log\det Z(\eta)$. That is also the negative log likelihood function for estimating η from Y, so the ML estimate of η will be

$$\hat{\eta} = \arg\min_{\eta\in\Gamma} Y^T Z(\eta)^{-1}Y + \log\det Z(\eta), \quad \text{MargLik.} \tag{29}$$

This maximization of the marginalized likelihood function is also known as *Empirical Bayes*. We have thus lifted the problem of estimating θ to a problem where we estimate parameters (in) P that describe the distribution of θ.

If the matrix Φ is not deterministic, but depends on E in such a way that row $\varphi^T(t)$ is independent of the element $e(t)$ in E, it is still true that (29) will be the ML estimate of η, although then Y is not necessarily Gaussian itself. See, e.g. Ljung (1999, Lemma 5.1) and Appendix in Pillonetto et al. (2011) for details.

Remark 5. The noise variance σ^2 is also typically not known and needs to be estimated. As suggested in Goodwin, Gevers, and Ninness (1992) and Ljung (1999), a simple and effective way is to estimate a low bias ARX (Pillonetto & De Nicolao, 2010) or FIR model (Chen et al., 2012) with least squares and use the sample variance as the estimate of σ^2. An alternative way is to treat σ^2 as an additional "hyper-parameter" contained in η, estimating it by solving (29), e.g. see Chen, Andersen, Ljung, Chiuso, and Pillonetto (2014) and MacKay (1992).

5. Regularization in system identification

We first consider the regularized FIR model identification problem, which is a useful intermediate step for the more advanced multi-input FIR model and ARX model identification problem.

5.1. FIR models

Let us now return to the impulse response estimation of $G(q, \theta)$ in (3) and assume it is finite (FIR) and described by:

$$y(t) = G(q, \theta)u(t) + e(t) = \sum_{k=1}^{m} g_k u(t-k) + e(t)$$

$$= \varphi_u^T(t)\theta_g + e(t) \tag{30}$$

where we have collected the m elements of $u(t-k)$ in $\varphi_u(t)$ and the m impulse response coefficients g_k in θ_g. That means that the estimation of FIR models is a linear regression problem. All that was said above about linear regressions, regularization and estimation of hyper-parameters can thus be applied to estimation of FIR models. In particular, suitable choices of P should reflect what is reasonable to assume about an impulse response: If the system is exponentially stable, the impulse response coefficients g_k should decay exponentially, and if the impulse response is smooth, neighboring values should have a positive correlation. That means that a suitable regularization matrix P^g for θ_g could be a matrix whose k, j element is

$$\text{DC}\quad P^g_{kj}(\eta) = \lambda\alpha^{(k+j)/2}\rho^{|j-k|};$$
$$\lambda \geq 0,\ 0 \leq \alpha < 1,\ |\rho| \leq 1;\ \eta = [\lambda, \alpha, \rho]. \tag{31}$$

Here α accounts for the exponential decay along the diagonal, while ρ describes the correlation across the diagonal (the correlation between neighboring impulse response coefficients). We call this matrix, or *kernel*, as it will be termed in Part II, DC for Diagonal/Correlated.

A special case is if we link $\rho = \sqrt{\alpha}$, leading to

$$\text{TC}\quad P^g_{kj}(\eta) = \lambda\alpha^{\max(k,j)};$$
$$\lambda \geq 0,\ 0 \leq \alpha < 1,\ \eta = [\lambda, \alpha] \tag{32}$$

which we call TC for Tuned/Correlated. The same kernel was introduced in Chen, Ohlsson, Goodwin, and Ljung (2011), Pillonetto, Chiuso, and De Nicolao (2010) and Pillonetto and De Nicolao (2011) with the name *First-order Stable Spline*.

A third useful kernel, which will be also explained in Part III, is called SS for Stable Spline:

$$\text{SS}\quad P^g_{kj}(\eta) = \lambda\left(\frac{\alpha^{k+j+\max(k,j)}}{2} - \frac{\alpha^{3\max(k,j)}}{6}\right)$$
$$\lambda \geq 0,\ 0 \leq \alpha < 1,\ \eta = [\lambda, \alpha]. \tag{33}$$

The hyperparameter η can then be tuned by (29) where $P^g(\eta)$ enters the definition of $Z(\eta)$ in (28). Efficient numerical implementation of this minimization problem is discussed in Carli, Chiuso, and Pillonetto (2012) and Chen and Ljung (2013). Once η is estimated, the impulse response can be computed by (19) with $\gamma = \sigma^2$. (The routine is implemented as `impulseest.m` in the 2012b version of Ljung (2013).)

This method of estimating impulse response, possibly followed by a model reduction of the high order FIR model, has been extensively tested in Monte Carlo simulations in Chen et al. (2012). They clearly show that the approach is a viable alternative to the classical ML/PEM methods, and may in some cases provide better models. An important reason for that is that the tricky question of model order determination is avoided.

5.2. Multi-input FIR-models

With several inputs, it is easy and natural to extend (30):

$$y(t) = \varphi^T_{u_1}(t)\theta^1_g + \cdots + \varphi^T_{u_k}(t)\theta^k_g + e(t) \tag{34a}$$
$$= \varphi_u(t)^T\theta_g + e(t) \tag{34b}$$

where $\varphi_{u_j}(t)$ contains the lagged inputs from input j and θ^j_g the impulse response coefficients from that input. The resulting linear regression (34b) simply stacks the contributions. If we assume that the responses from the different inputs have no mutual correlation, it is natural to partition the regularization matrix accordingly:

$$P(\eta) = \begin{bmatrix} P^{g_1}(\eta_1) & 0 & \cdots & 0 \\ 0 & P^{g_2}(\eta_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & P^{g_k}(\eta_k) \end{bmatrix} \tag{35}$$

with $P^{g_j}(\eta_j)$ as in any of (31)–(33), with different hyperparameters for each input.

5.3. General predictor models and ARX models

From the general linear predictor expression (5) we can write any predictor as two infinite impulse responses from y and u respectively. Note for ARX models (7) the expressions $1 - \frac{1}{H(q)} = 1 - A(q)$ and $\frac{G(q)}{H(q)} = B(q)$, so these infinite responses specialize to finite responses. In line with (9), these finite ARX-expressions become arbitrarily good approximators for general linear systems as the orders tend to infinity. We can write the ARX-model, (7) with one input as

$$\begin{aligned} y(t) &= -a_1y(t-1) - \cdots - a_{n_a}y(t-n_a) + b_1u(t-1) \\ &\quad + \cdots + b_{n_b}u(t-n_b) + e(t) \\ &= \varphi^T_y(t)\theta_a + \varphi^T_u(t)\theta_b + e(t) \end{aligned} \tag{36}$$

where $\theta_a = \begin{bmatrix} a_1 & \cdots & a_{n_a} \end{bmatrix}^T$, $\theta_b = \begin{bmatrix} b_1 & \cdots & b_{n_b} \end{bmatrix}^T$ and $\varphi_y(t)$, $\varphi_u(t)$ are made up from y and u in an obvious way. That means that also the ARX model is a linear regression model, to which the same ideas of regularization can be applied. Eq. (36) shows that the predictor consists of two impulse responses, from y and from u, and similar ideas on the parameterization of the regularization matrix can be used. The P-matrix in (19) can be partitioned along with θ_a, θ_b:

$$P(\eta_1, \eta_2) = \begin{bmatrix} P^a(\eta_1) & 0 \\ 0 & P^b(\eta_2) \end{bmatrix} \tag{37}$$

with $P^a(\eta)$, $P^b(\eta)$ as in any of (31)–(33).

Finally, the ARX model (36) is extended to multiple inputs in an obvious way. If there are several outputs, the easiest and most natural way is to treat each output channel as a separate linear regression as in (36) but with the other outputs appended in the same way as the inputs.

Remark 6. Note that linear systems satisfy the superposition property that impulse response of a linear system is the sum of impulse responses of its partial fraction expansion and also that the hyperparameter tuning problem (29) is non-convex, and for high-dimension η it may cause problems. It is therefore of interest to consider kernels that are formed linearly from multiple, known kernels P_k:

$$P(\eta) = \sum_{k=1}^{r} \eta_k P_k; \quad \eta_k \geq 0 \tag{38}$$

where P_k can be chosen as specific instances of the kernels DC, TC and SS, and also complemented by rank 1 kernels of the kind $\theta_0\theta_0^T$ (cf. (24)) for a collection of candidate models θ_0. This gives several advantages as described in Chen et al. (2014). For example, the tuning problem (29) for (38) is a difference of convex functions programming problem, whose locally optimal solutions can be found efficiently by using sequential convex optimization techniques, (Horst & Thoai, 1999; Tao & An, 1997). What is more, it favors sparse solutions, i.e. with many $\eta_k = 0$. This is very useful if P_k corresponds to different hypothesized structure and enables this kernel-based regularization method to tackle various structure detection problems in system identification, e.g., the sparse dynamic network identification and the segmentation of linear systems.

6. Regularized least squares and James–Stein estimators ⋆

Consider (19) under orthonormal design assumptions ($\Phi^T\Phi = I_m$), with regularization matrix proportional to the identity ($P = \lambda I_m$) and with σ^2 assumed known. Then, if γ is set to σ^2 and λ is estimated via marginal likelihood optimization, it is easy to show that ReLS reduces essentially to the famous James–Stein estimator (James & Stein, 1961; Stein, 1981). Hence, for $m > 2$, its performance, measured by the trace of the MSE matrix of $\hat{\theta}$, is uniformly better than that of the LS estimator for every θ_0. See also Efron and Morris (1973) for connections between James–Stein estimation and empirical Bayes approaches.

In the general case (non orthonormal Φ and/or generic kernel), a large variety of different estimators dominating LS can be found in the literature, see e.g. Berger (1982), Bhattacharya (1966), Bock (1975) and Shinozaki (1974), but ReLS (e.g. equipped with marginal likelihood for hyperparameters tuning) does not belong to this class. In fact, to get minimax properties (thus guaranteeing that the MSE never exceeds that of LS for every θ_0), ReLS structure needs to be suitably modified giving rise to generalized ReLS estimators (Strawderman, 1978). However, there is a price to pay: generalized ReLS is typically less effective against ill-conditioning, e.g. see Casella (1980) but also Maruyama and Strawderman (2005) for new minimax estimators with better numerical stability properties.

The Bayesian interpretation reported in Section 4.3 helps to understand why ReLS performance can prove superior to other minimax estimators when regularization is carefully tuned. In fact, ReLS concentrates the improvement in the most likely regions in the parameter space as specified by the chosen kernel (it will perform worse than LS only if the probability induced by the kernel predicts poorly the location of θ_0). In particular, ReLS equipped with kernels (31)–(33) outperforms LS in the region of exponentially decaying impulse responses. In this region, forcing the estimator to be minimax, one would loose most of the gain coming from smoothness and stability prior information, see also Section 3 in Berger (1982) for other insights.

Nevertheless, the development of effective minimax estimators for system identification appears an interesting issue. To our knowledge, this is an open problem when kernel parameters, such as (λ, α) in (32) and (33), have to be determined from data. Instead, when they can be fixed in advance, the estimator in Berger (1982) can be used: not only it is minimax but concentrates the improvement in ellipsoids defined by P in the parameter space. This approach is deeply connected with robust Bayesian estimation concepts, e.g. see Berger (1980, 1994).

7. Numerical illustrations

To illustrate the properties of the suggested algorithms, we consider two kinds of Monte Carlo experiments regarding the identification of randomly generated linear systems

$$y(t) = \left\{ \sum_{i=1}^{p} G_i(q) u_i(t) \right\} + H(q)e(t). \tag{39}$$

The experiment in Section 7.2 involves OE-models ($p = 1, H(q) = 1$) while ARMAX-models will be then considered in Section 7.3. Both of the experiments have been performed using MATLAB, equipped by the System Identification Toolbox (Ljung, 2013) as the numerical platform. [3] There, all the methods described in Sections 2 and 5 are implemented, and we refer to the manual of Ljung (2013) for details of the implementations.

[3] Data and related code are available at the website www.control.isy.liu.se/books/sysid/AutomaticaSurvey.

7.1. Performance measure: model generalization capability

First, we describe the performance measure adopted to compare different estimated models. At every Monte Carlo run, the system (39) is used to generate two kinds of data sets. The first type is the identification data set (also called training or estimation set)

$$\mathscr{Z} = \{u(1), y(1), \ldots, u(N), y(N)\}$$

while the second type is the test set

$$\mathscr{Z}^{new} = \{u^{new}(1), y^{new}(1), \ldots, u^{new}(M), y^{new}(M)\}.$$

The test set $\mathscr{Z}^{new}$ is generated by applying inputs (independent of those entering $\mathscr{Z}$) at $t = 0$, starting from null initial conditions.

The set $\mathscr{Z}^{new}$ serves to characterize the *generalization capability* of a model, i.e. its ability to predict future outputs, not contained in $\mathscr{Z}$. More specifically, let $\hat{y}_k^{new}(t|\hat{\theta})$ be the k-step-ahead predictor associated with an estimated model (characterized by $\hat{\theta}$). It yields the rule to predict (k-step-ahead) $y^{new}(t)$ from the knowledge of u^{new} up to time $t-1$ and y^{new} up to $t-k$. If $\bar{y}^{new}$ denotes the average output in $\mathscr{Z}^{new}$, the performance is measured by the fit

$$\mathscr{F}_k(\hat{\theta}) = 100 \left(1 - \sqrt{\frac{\sum_{t=1}^{M} \left(y^{new}(t) - \hat{y}_k^{new}(t|\hat{\theta}) \right)^2}{\sum_{t=1}^{M} (y^{new}(t) - \bar{y}^{new})^2}} \right) \tag{40}$$

which quantifies how much of the variance of y^{new} is captured by the k-step-ahead forecast. In the OE-model case, $\mathscr{F}_k$ is independent of k. Note that is essential that the k-step ahead predictions are computed with zero initial conditions, so as not to give any advantages to higher order models which could adjust the initial conditions to fit the test set. (This can be obtained using the MATLAB command `predict(model, data, k, 'ini', 'z')` where `model` and `data` are structures containing the estimated model and the test set $\mathscr{Z}^{new}$, respectively.)

The quality measure (40) will depend on the actual data in $\mathscr{Z}^{new}$, but as $M \to \infty$ it will be a true measure of the quality of the model, depending only on the true system $G_i(q), H(q)$, the model $G_i(q, \hat{\theta}), H(q, \hat{\theta})$ and the input spectrum. For example for an OE-model with zero mean white noise as input, one has

$$\mathscr{F}_k(\hat{\theta}) \to 100 \left(1 - \frac{\|G(q) - G(q, \hat{\theta})\|_2}{\|G(q)\|_2} \right) \tag{41}$$

with $\|S(q)\|_2$ being the ℓ_2 norm of the impulse response of a generic system $S(q)$.

The oracle. The test set $\mathscr{Z}^{new}$ and the fit $\mathscr{F}_k(\hat{\theta})$ are useful not only for model tests, but they can also be used as an *oracle*. In statistics the term oracle is often used as a means for correct information about model properties based on ideal and unrealistic sources. In that sense, $\mathscr{Z}^{new}$ is an oracle only if $M = \infty$ or if y has been generated from the system in a noise-free way. In the tests of OE models in Section 7.2 we shall use $\mathscr{Z}^{new}$ and $\mathscr{F}_1(\hat{\theta})$ with noise free data (and $M = 3000$) as a true oracle to select the best order of the OE model.

For the ARMAX tests in Section 7.3, which include a noise model, we use as an "approximate oracle" $\mathscr{Z}^{new}$ for $M = 3000$ and a noise source of the same character as in the estimation data. To decide what are the best ARMAX orders, an average k-step prediction performance is evaluated, and

$$\sum_{k=1}^{20} \mathscr{F}_k(\hat{\theta}) \tag{42}$$

is maximized.

7.2. Identification of discrete-time OE-models

We now consider two Monte Carlo studies of 1000 runs regarding identification of discrete-time OE-models

$$y(t) = G(q)u(t) + e(t), \quad G(q) = \frac{B(q)}{F(q)}.$$

At each run, a different rational transfer function G is generated as follows. First, a 30th order SISO continuous-time, strictly proper system was randomly generated (using the MATLAB command `rss.m`). The continuous-time system was then sampled at 3 times of its bandwidth. If all poles of the sampled system are within the circle with center at the origin and radius 0.95 on the complex plane, the system is used and saved.

During each run, the input in the estimation data set is generated as a realization from white Gaussian noise of unit variance filtered by a 2nd order rational transfer function obtained by the same type of generator defining G. The input delay is always equal to 1. Starting from zero initial conditions, 1000 input–output pairs are collected with the output corrupted by an additive white Gaussian noise. The SNR, i.e. ratio between the variance of the noiseless output and the noise, is randomly chosen in [1, 10] at every run. In the first experiment, the estimation data set $\mathscr{Z}$ contains the first 200 input–output pairs while all the 1000 pairs are used in the second case study.

Two types of test sets $\mathscr{Z}^{new}$ are generated at every run. The first one is the most challenging since it contains noiseless outputs obtained using a unit variance white Gaussian noise as input. The second one is obtained using a test input having the same statistics of the input in the estimation data.

The following 6 estimators are used:

- $Oe + Or1$. Classical PEM approach (11) equipped with an oracle. In particular, we consider candidate models where the order of the polynomials B and F is equal and can vary between 1 and 30. For every model order, estimation is performed solving (11). (The method is implemented in `oe.m` of the MATLAB System Identification Toolbox, (Ljung, 2013).) Then, the oracle chooses the estimate which maximizes the fit (40) (independent of k in this case) relative to the first test set (test input equal to white noise).
- $Oe + Or2$. The same as above except that the oracle chooses the model order maximizing the prediction performance on the second test set (test input with the same statistics of the input in the estimation data).
- $Oe + CV$. The same as above except that model order is selected using cross validation. In particular, identification data are split into two sets $\mathscr{Z}_a$ and $\mathscr{Z}_b$ containing, respectively, the first and the last $N/2$ input–output pairs in $\mathscr{Z}$. For different orders of the rational transfer function, the models are obtained by the PEM method using the estimation data $\mathscr{Z}_a$. Then the prediction errors are computed, with zero initial conditions, for the validation data $\mathscr{Z}_b$. (This can be obtained using `pe(model,data,'z')` where `model` contains the model obtained by `oe.m` and `data` contains $\mathscr{Z}_b$.) The model order minimizing the sum of the squared prediction errors is selected and the final impulse response estimate is obtained solving (11) for the complete data set $\mathscr{Z}$.
- $\{TC, SS, DC\}$. These are three ReLS estimators, see (19), equipped with the kernels DC (31), TC (32) and SS (33). The number of estimated impulse response coefficients is 200. At every run, the noise variance is estimated by fitting via LS a low-bias model for the impulse response, as e.g. described in Goodwin et al. (1992). Then, kernel hyperparameters (2 for SS and TC, 3 for DC) are obtained via marginal likelihood optimization, see (29).

7.2.1. Results

Fig. 1 reports the MATLAB boxplots of the 1000 fit measures returned by the estimators during the first experiment ($N = 200$, left panels) and the second experiment ($N = 1000$, right panels). Table 1 also displays the average fit values.

The top panels of Fig. 1 displays the fits relative to the first test set. The performance reference is $Oe + Or1$. It is apparent that the performance of $Oe + CV$ is unsatisfactory, far from that of $Oe + Or1$. In accordance with the analysis reported in Pillonetto and De Nicolao (2012), such an approach exhibits a poor control on model complexity. Hence, it often returns impulse response estimates affected by ill-conditioning. The performance of all the three regularized approaches is instead close to that of $Oe + Or1$, or also better. E.g., Table 1 reveals that TC and DC outperform the oracle when the data set size is 200.

The bottom panels of Fig. 1 display the fits relative to the second test set. The performance reference is now $Oe + Or2$. Prediction is now easier since the estimation and test data are more similar. As a consequence, the performance of $Oe + CV$ much increases but remains significantly inferior than that of the kernel-based estimators.

Finally, it is worth noticing that the performances of $Oe + Or1$ and $Oe + Or2$ appear quite different in each of the four scenarios illustrated by Fig. 1. This indicates that the model orders chosen by the oracle strongly depend on the prediction target. On the other hand, the generalization capability of the regularized estimators is always high, irrespective of the nature of the test set. Hence, one can argue that TC, SS and DC (which are estimators implementable in real applications) return impulse response estimates which are "the right synthesis" of those returned by the two oracle-based procedures (which are not implementable in practice since have access to noise free data contained in the test set). These outcomes have been recently confirmed also by an independent study (Olofsson, 2013).

7.3. Identification of discrete-time ARMAX-models

We now consider one Monte Carlo study of 1000 runs. At every run, data are generated by an ARMAX model of order 30 having p observable inputs u_i, i.e.

$$y(t) = \left\{\sum_{i=1}^{p} \frac{B_i(q)}{A(q)} u_i(t)\right\} + \frac{C(q)}{A(q)} e(t)$$

where p is the realization from a random variable uniformly distributed on $\{2, 3, 4, 5\}$. The polynomials A, B_i and C are randomly generated. (This has been obtained by using `drmodel.m`: the first call defines B_1 and A, the others the numerators of the remaining p rational transfer functions.) The system is used and saved if the following two requirements are satisfied. System and one-step ahead predictor poles have to stay inside the circle of radius 0.95 while the signal to noise ratio has to satisfy

$$1 \leq \frac{\sum_{i=1}^{p} \|G_i\|_2^2}{\|H\|_2^2} \leq 10$$

where $G_i(q) = \frac{B_i(q)}{A(q)}$, $H(q) = \frac{C(q)}{A(q)}$ with $\|G_i\|_2$, $\|H\|_2$ to denote the ℓ_2 norms of the system impulse responses, which are also constrained to be less than 10.

At every run, the input in the estimation data set is unit variance white Gaussian noise and $\mathscr{Z}$ contains 300 input–output pairs collected after getting rid of initial conditions effect. The test input is also white Gaussian noise and the performance measure is the fit (40) which now depends on the prediction horizon k.

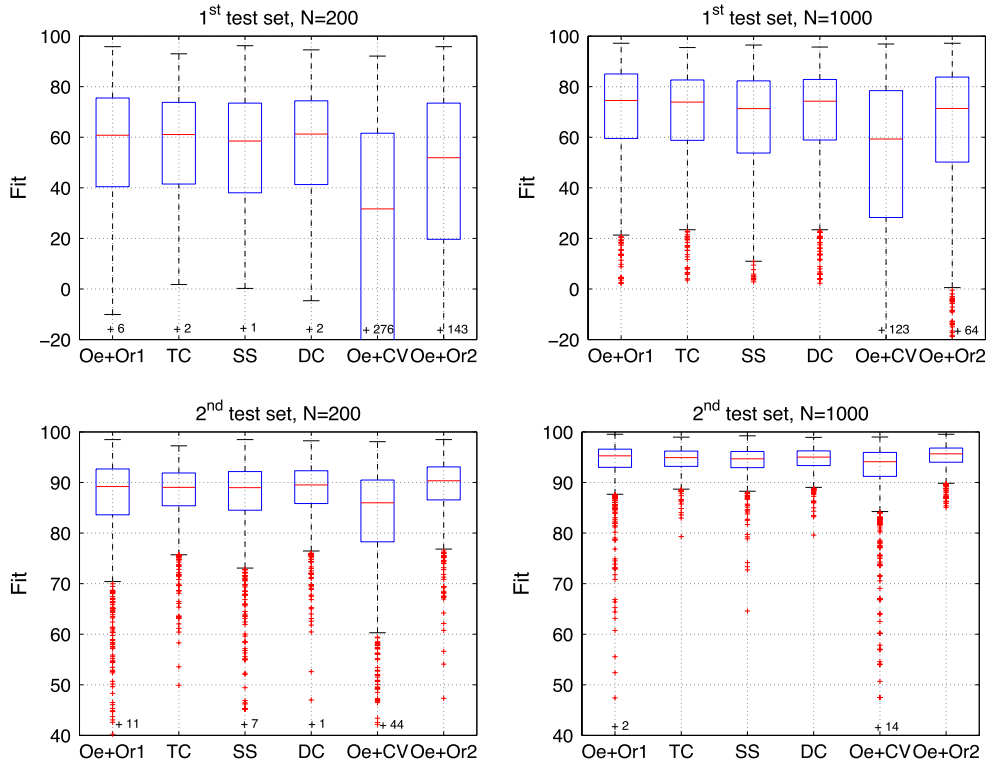

Fig. 1. Identification of discrete-time OE-models (Section 7.2). *Top* Boxplot of the 1000 prediction fits on future outputs generated using white noise as test input. Identification data consist of 200 (top left) or 1000 (top right) input–output pairs. *Bottom* Same results as in the top panel except that the statistics of the input in the estimation and test data set coincide. In all the four panels, the first and last boxplots report results from the estimators $Oe + Or1$ and $Oe + Or2$ which are not implementable in practice.

Table 1
Identification of discrete-time OE-models (Section 7.2). Average fit as a function of the identification data set size ($N = 200$ or $N = 1000$) and of the type of test set. The first and last columns report results from oracle-based estimators not implementable in practice.

	$Oe + Or1$	*TC*	*SS*	*DC*	$Oe + CV$	$Oe+ Or2$
1st test set, $N = 200$	56.2	56.6	54.8	56.5	−88.1	−14.9
1st test set, $N = 1000$	70.1	69.0	66.5	69.1	−43.2	14.8
2nd test set, $N = 200$	85.7	87.9	86.7	88.4	79.5	89.1
2nd test set, $N = 1000$	93.8	94.5	94.2	94.6	91.1	95.2

The following estimators are used:

- *PEM + Oracle*: this is the classical PEM approach (11) equipped with an ("approximate") oracle (42). The candidate model structures are ARMAX models defined by polynomials that all have the same degree. (The method is implemented in `pem.m` of the MATLAB System Identification Toolbox, (Ljung, 2013).) The maximum allowed order is 30. The oracle chooses the model order maximizing (42).
- *PEM + CV*: the same as above except that cross validation is used for model order selection. In particular, identification data are split into two sets $\mathscr{Z}_a$ and $\mathscr{Z}_b$ containing, respectively, the first and the last 150 input–output pairs in $\mathscr{Z}$. For different orders of the rational transfer function, the model is obtained solving (11) using the estimation data $\mathscr{Z}_a$. Then, the one-step-ahead prediction errors are computed with zero initial conditions for the validation data $\mathscr{Z}_b$. The model order minimizing the sum of the squared prediction errors is selected and the final impulse response estimate is obtained solving (11) using $\mathscr{Z}$.
- {*PEM + BIC*, *PEM + AICc*}: the same as above except that AIC-type criteria select the model order.
- {*TC*, *SS*, *DC*}: the unknown coefficients of the multi-input version of the ARX model (36) are estimated via ReLS. The length of each predictor impulse response is 50. The regularization matrices entering (the multi-input version of) (37) all consist of TC (32) or SS (33) or DC (31) kernels, sharing a different scale factor λ for every impulse response and a common variance decay rate α. The innovation variance is obtained using a low-bias ARX model (following the same approach described in the previous subsection). Then, hyperparameters are determined via marginal likelihood optimization. To deal with initial conditions effect, the first 50 input–output pairs in $\mathscr{Z}$ are used just as entries of the regression matrix.

The system inputs delay are assumed known and their values are provided to all the estimators described above.

7.3.1. Results

Fig. 2 displays the average of the fits $\mathscr{F}_k$ defined in (40) as a function of the prediction horizon k (left panel) and the MATLAB boxplots of the 1000 values of $\mathscr{F}_1$ (right panel).

One can see that the performance of *TC* and *SS* is similar and close to that of *PEM + Oracle* (even better for $k \leq 4$). For what regards *DC*, its average performance is always better than *PEM + Oracle*. These results are remarkable also recalling that *TC*, *SS* and *DC* are estimators that can be used in real applications while *PEM + Oracle* is an ideal tuning which has access to the test set.

Finally, one can see that the regularized approaches largely outperform PEM equipped with CV and the Akaike-like criteria.

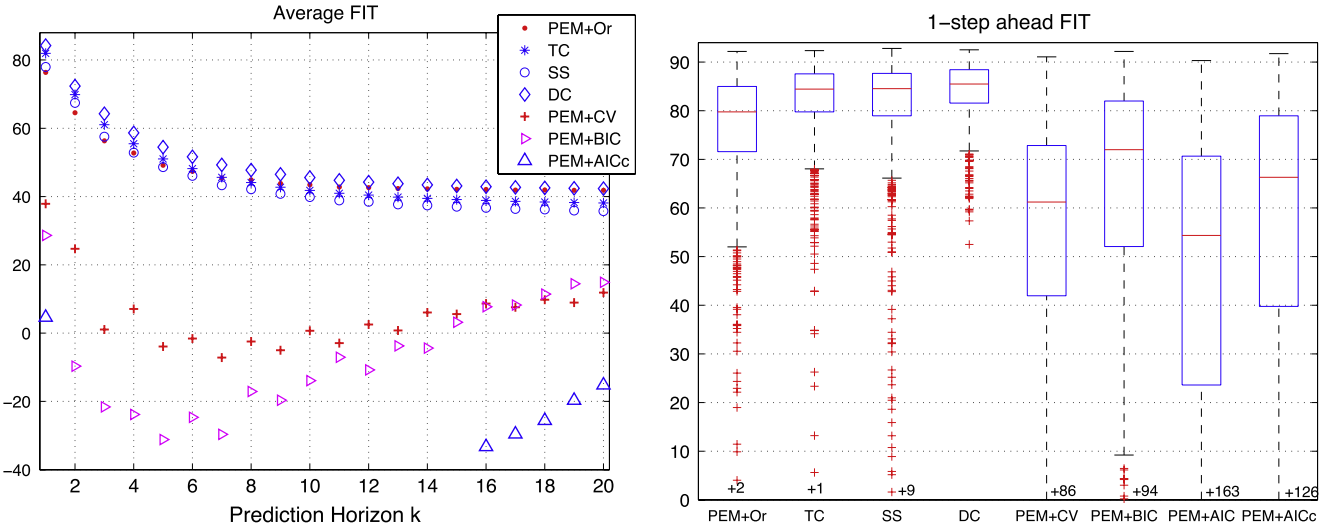

Fig. 2. Identification of ARMAX models (Section 7.3). *Left* Average of the k-step ahead fits $\mathscr{F}_k$ as defined in (40). *Right* Boxplots of the 1000 values of $\mathscr{F}_1$. Recall that *PEM + Oracle* uses additional information, having access to the test set to perform model order selection.

Remark 7. How can the DC be better than the oracle? Nothing can beat the oracle in the sense that PEM/OE equipped with any model order selection rule cannot beat the oracle. But the model order selection works with bias-variance trade-off among a finite set of given models. Regularization deals with that trade-off using a continuous set of regularization parameters, and may in this way come up with better performing trade-offs. Also, when studying the results in Fig. 2, it is important to keep in mind that these are experiments with relatively few (300) data, and quite complex systems (orders 30).

Remark 8. We have also tested a variant of the regularized algorithms given by *TC* with a single scale factor λ for all the regularization matrices. In this way, the dimension of the hyperparameter vector is independent of the number of system inputs, with the marginal likelihood to be optimized just over a two-dimensional space. The mean of the 20 fits values displayed in the left panel of Fig. 2 only slightly reduces, passing from 46.6 to 43.6.

8. An example with a real process

In robotics, a good model of the robot is one key of the success for high positioning accuracy/low tracking errors, which are performance specifications and the driving elements for any servo design. In Torfs, Vuerinckx, Swevers, and Schoukens (1998) a process with a vibrating flexible robot arm is described. The experimental data from this process have also been analyzed in Pintelon and Schoukens (2012b, Section 11.4.4). The input is the driving couple and the output is the acceleration of the tip of the robot arm. In total, 40960 data points were collected at a sampling rate of 500 Hz. A portion of the input–output data is shown in the left panel of Fig. 3. The right panel of the same figure displays the empirical frequency function estimate obtained by these data. (This is implemented in `etfe.m` of the MATLAB System Identification Toolbox (Ljung, 2013).)

We have built models both using standard PEM and regularized FIR-model techniques. Since the true system is unknown we cannot evaluate the models by model fit to the actual system. Instead we use cross validation ideas, and measure how well the estimated models can reproduce the output on validation portions of the data that were not used for estimation. We selected estimation data to be the portion 1:7000 and the validation data to be the portion 10,000:40,960. We estimated nth order state space models without disturbance model for $n = 1, \ldots, 36$ using the prediction error method PEM (via the command `pem(data,n,'dist','no')`) and calculated their fit to validation data as in (40). The fits are shown as a function of n in Fig. 4. The best fit is 78.7% and obtained for order $n = 18$. Then, the regularized FIR model (30) with order $m = 3000$ is estimated using (19), with the DC kernel (31), tuned by the marginalized likelihood method (29), and the unknown input data are set to zero. (This method is available in MATLAB's System Identification Toolbox (Ljung, 2013) as the command `impulseest(data,3000,0,opt)` where, in the option `opt`, we set `opt.RegulKernel='dc'; opt.Advanced.AROrder=0`.) The fit for a regularized FIR model of order 3000 is 81.8% and clearly better than any of the PEM-estimated state space models. For illustration, this fit is also shown in Fig. 4 as a horizontal line.

One can say that a FIR model of order 3000 is quite large, but it is interesting to note that it can be reduced to low order state-space models by $\mathscr{L}_2$ model order reduction. For example, the fit of a reduced state-space model of order $n = 15$ is 79.5%, which is better than the best PEM-estimated state space model.

One may also say that estimating a FIR model of order 3000 is not practical. One way to deal with this issue is to decimate the original data by a factor of 10 and repeat the tests. They show that with the DC kernel, a regularized FIR model of order 300 gives a fit of 87.0% while the best PEM-estimated state-space model has a fit of 82.2%.

Part II. Function estimation by regularized kernel methods

In this part, we study function estimation by regularized kernel methods.

9. Function estimation problem and RKHS

Methods for estimating (learning) a function g in a functional relationship $y = g(x)$ from observed samples of y and x are the basic building blocks for black-box estimation techniques. Given a finite set of pairs $(x_i, y_i) \in \mathscr{X} \times \mathbb{R}$, where $\mathscr{X}$ is a non-empty set, the goal is synthesizing a function g having a good generalization capability in the sense that, for a new pair (x, y), the prediction $g(x)$ is close to y (e.g. in the MSE sense). The classical parametric approach uses a model $g_\theta : \mathscr{X} \to \mathbb{R}$ depending on a vector of parameters $\theta \in \mathbb{R}^m$. A very simple example is a finite-dimensional polynomial model, e.g. $g_\theta(x) = \theta_1 + \theta_2 x + \theta_3 x^2$.

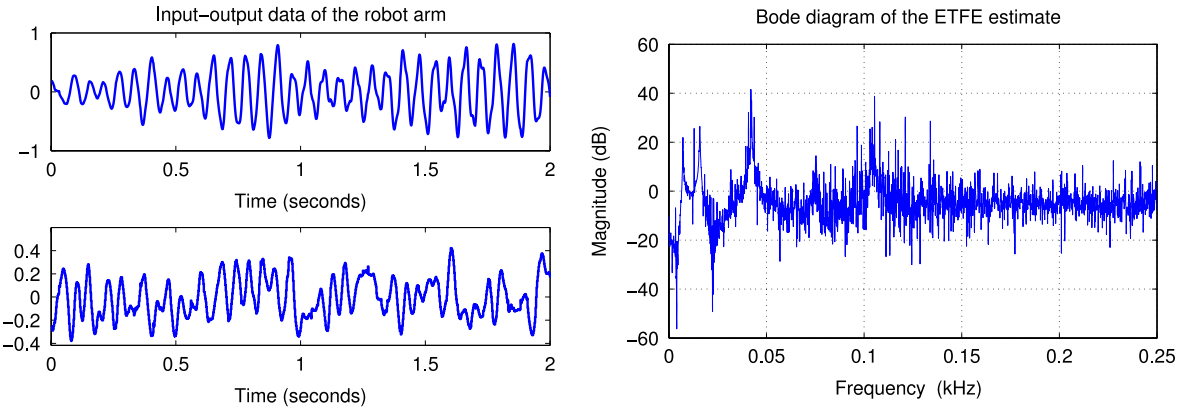

Fig. 3. Identification of the flexible robot arm. *Left* The first 1000 pairs of input–output data: the input is the driving couple (bottom) and the output is the tip of the robot arm (top). *Right* Magnitude of the empirical transfer function estimate (ETFE) constructed using all 40,960 data.

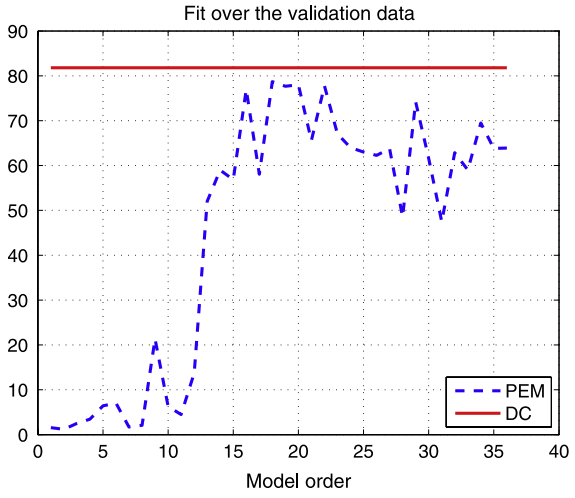

Fig. 4. Identification of the flexible robot arm. Values of $\mathscr{F}_1$ for different model order $n = 1, \ldots, 36$. The solid line is the fit for the regularized DC FIR model.

As mentioned in Part I, a well known parametric regression method is the classical LS one: the vector θ is obtained by minimizing a functional of the form

$$\sum_{i=1}^{N}(y_i - g_\theta(x_i))^2 \tag{43}$$

that dates back to Gauss. It is convenient to adopt a model which is linear in the parameter vector θ, i.e. $g_\theta(x) = \sum_{i=1}^{m}\theta_i\phi_i(x)$, where the basis functions ϕ_i are fixed in advance. In this way, global minimization with respect to θ can be obtained just solving a linear system of equations.

As already seen, within the parametric approach a major issue is the choice of the model order m ($\dim(\theta)$), which is typically carried out by using model validation techniques, see Section 3.1 and Ljung (1999), Söderström and Stoica (1989). Increasing the order improves the LS penalty eventually leading to data interpolation. However, overparameterized models, as a rule, perform poorly when used to predict the output from new input data. A further problem with overparameterized models is that the problem of minimizing (43) may become ill-posed in the sense of Hadamard (1922). In particular, the solution may be highly sensitive to small perturbations of the data y_i. That is the same observation as expressed at the end of Section 2.4 that flexible models give higher variance.

9.1. Regularization in reproducing kernel Hilbert spaces

Is there a way to reconcile flexibility of the model class with well-posedness of the solution? The question has been extensively investigated in the literature of inverse problems, opening the way to regularization techniques (Tikhonov & Arsenin, 1977) that have been widely adopted also in the machine learning literature. In the specific case of function estimation, the regularization approach is in fact an alternative paradigm to traditional parametric estimation. Instead of constraining the unknown function to a specific parametric structure, g is searched over a possibly infinite-dimensional functional space $\mathscr{H}$. The key ingredient to avoid overfitting and ill-posedness is the introduction of a regularizer J in the objective functional:

$$\min_{g\in\mathscr{H}}\left(\sum_{i=1}^{N}(y_i - g(x_i))^2 + \gamma J(g)\right). \tag{44}$$

The regularization term $J(g)$ is designed so as to penalize undesired behaviors. For instance, the regularizer given by the energy of first-order derivative, i.e. $J(g) = \int (g^{(1)}(x))^2 dx$, penalizes the presence of high-frequency components in the function g. The positive parameter γ, which was already introduced in (19), controls the relative importance of the error term $\sum_{i=1}^{N}(y_i - g(x_i))^2$, and the regularizer $J(g)$. A mathematically rigorous and elegant analysis of regularization methods is possible when $\mathscr{H}$ is a Hilbert space, namely a space endowed with an inner product $\langle\cdot,\cdot\rangle_{\mathscr{H}}$, which is complete (meaning that all the Cauchy sequences converge) with respect to the induced norm $\|g\|_{\mathscr{H}} = \sqrt{\langle g, g\rangle_{\mathscr{H}}}$. In such a case, a typical regularizer is the squared norm in the Hilbert space, i.e. $J(g) = \|g\|^2_{\mathscr{H}}$ which leads to:

$$\min_{g\in\mathscr{H}}\left(\sum_{i=1}^{N}(y_i - g(x_i))^2 + \gamma\|g\|^2_{\mathscr{H}}\right). \tag{45}$$

As for the Hilbert space $\mathscr{H}$, a basic requirement is that every function in the space be point-wise well defined everywhere on $\mathscr{X}$. In addition, we will assume that pointwise evaluations are continuous linear functionals on $\mathscr{H}$, i.e.

$$\forall x \in \mathscr{X}, \quad \exists C_x < \infty : \ |g(x)| \le C_x\|g\|_{\mathscr{H}}, \quad \forall g \in \mathscr{H}. \tag{46}$$

Notice that the above condition is stronger than requiring $g(x) < \infty\ \forall x$ due to the fact that C_x can depend on x but not on g.

Definition 9 (*RKHS*)**.** A reproducing kernel Hilbert space (RKHS) over a non-empty set $\mathscr{X}$ is a Hilbert space of functions $g : \mathscr{X} \to \mathbb{R}$ such that (46) holds.

When a RKHS is adopted as hypothesis space, problem (45) is well-posed: there exists a unique solution that is little sensitive to perturbations of the data (Tikhonov & Arsenin, 1977). The theory of RKHS has been mainly developed by Aronszajn (Aronszajn, 1950). As suggested by the name itself, the concept of RKHS is strongly linked with that of positive semidefinite kernel.

Definition 10 (*Positive Semidefinite Kernel*)**.** Let $\mathscr{X}$ denote a non-empty set. A symmetric function $K : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ is called *positive semidefinite kernel* if, for any finite natural number p, it holds

$$\sum_{i=1}^{p}\sum_{j=1}^{p} a_i a_j K(x_i, x_j) \geq 0, \quad \forall (x_k, a_k) \in (\mathscr{X}, \mathbb{R}),\ k = 1, \ldots, p.$$

Given a kernel K, the *kernel section* $K_x \in \mathscr{H}$ centered at x is $K_x(a) = K(x, a), \forall a \in \mathscr{X}$. The following theorem provides the connection between RKHS and positive semidefinite kernels.

Theorem 1 (*Moore–Aronszajn*)**.** *To every RKHS $\mathscr{H}$ there corresponds a unique positive semidefinite kernel K, called the reproducing kernel, such that the reproducing property holds:*

$$g(x) = \langle g, K_x \rangle_{\mathscr{H}}, \quad \forall (x, g) \in (\mathscr{X}, \mathscr{H}). \tag{47}$$

Conversely, given a positive semidefinite kernel K, there exists a unique RKHS of real valued functions defined over $\mathscr{X}$ whose reproducing kernel is K.

Remark 11. The Moore–Aronszajn theorem gives a one-to-one correspondence between RKHS of functions and positive semidefinite kernels (Aronszajn, 1950). It follows that the Hilbert space $\mathscr{H}$ is completely characterized by its reproducing kernel. This also means that the kernel choice specifies both the hypothesis space $\mathscr{H}$ and the regularizer $J(g) = \|g\|^2_{\mathscr{H}}$ entering the estimator (45).

In particular, from the proof of the theorem, e.g. reported on p. 35 of Cucker and Smale (2001), it follows that every RKHS is built from the kernel as follows. First, consider all the functions of the type $g(x) = \sum_{i=1}^{p} a_i K_{x_i}(x)$ for every choice of p, a_i and x_i. This defines a subspace that is then equipped with a suitable inner product inducing the norm $\|g\|^2_{\mathscr{H}} = \sum_{i=1}^{p}\sum_{j=1}^{p} a_i a_j K(x_i, x_j)$ (see Cucker and Smale (2001) for details). Finally, adding all the limits of Cauchy sequences to this subspace provides the RKHS associated with K. Thus, every function in the RKHS is a linear combination of a possibly infinite number of kernel sections. Assume for instance $K(x_1, x_2) = \exp\left(-\|x_1 - x_2\|^2\right)$. Then, all the functions in the corresponding RKHS are sums, or limits of sums, of functions proportional to Gaussians. It can be shown that every function of the space $\mathscr{H}$ inherits properties such as smoothness and integrability of the kernel. This fact has an important consequence on modeling: instead of specifying a whole set of basis functions, it suffices to choose a single positive semidefinite kernel function that encodes the desired properties of the function to be synthesized.

9.2. The representer theorem

Hereby, let I_N be the $N \times N$ identity matrix. In addition, the (column) vector Y contains the available output measurements y_i while $\mathbf{K} \in \mathbb{R}^{N\times N}$ is a positive semidefinite matrix (called *kernel matrix*, or Gram matrix) such that $\mathbf{K}_{ij} = K(x_i, x_j)$.

The importance of RKHS in the context of regularization methods stems from the following central result (Kimeldorf & Wahba, 1971b) showing that the solution of the variational problem (45) admits a finite-dimensional representation.

Theorem 2 (*Representer Theorem*)**.** *If $\mathscr{H}$ is a RKHS, the minimizer of (45) is*

$$\hat{g}(x) = \sum_{i=1}^{N} \hat{c}_i K_{x_i}(x) \tag{48}$$

where $\hat{c} = [\hat{c}_1, \ldots, \hat{c}_N]^T$ is given by

$$\hat{c} = (\mathbf{K} + \gamma I_N)^{-1} Y. \tag{49}$$

Similarly to the traditional linear parametric approach, the optimal function is a linear combination of basis functions. However, a fundamental difference is that the number of basis functions is now equal to the number of data pairs, and is thus not fixed a-priori. In fact, the basis functions appearing in the expression of the minimizer $\hat{g}$ are just the kernel sections K_{x_i} centered on the input data. The estimator (48)–(49) is also called in the literature *regularization network* (Poggio & Girosi, 1990) or *least squares support vector machine* (Suykens et al., 2002). Its asymptotic behavior, as N goes to ∞, has been studied in many recent works, see e.g. Smale and Zhou (2007) and Wu, Ying, and Zhou (2006), also in the context of NARX identification (De Nicolao & Trecate, 1999).

Remark 12. More general versions of the representer theorem exist (Schölkopf et al., 2001). In particular, when the quadratic terms $\sum_{i=1}^{N}(y_i - g(x_i))^2$ are replaced by general convex functions, e.g. the Huber (Huber, 1981) or the Vapnik loss which leads to support vector regression (Vapnik, 1998), (48) still holds under mild assumptions. The adopted loss then determines how to obtain the expansion coefficients: generally, in place of (49), $\hat{c}$ requires the solution of a (possibly non differentiable) convex optimization problem.

9.3. The bias space ⋆

Sometimes, it can be useful to enrich the RKHS $\mathscr{H}$ with a low-dimensional parametric part, the so called *bias space*. This is typically defined by linear combinations of functions $\{\phi_k\}_{k=1}^{m}$. E.g., if the unknown g exhibits a linear trend, one may let $m = 2$ and $\phi_1(x) = 1, \phi_2(x) = x$, and assume that the unknown function is sum of two functions, one in $\mathscr{H}$ and the other one in the bias space. In other words, the hypothesis space is $\mathscr{H} + \text{span}\{\phi_1, \ldots, \phi_m\}$. Then, the estimated function is $\hat{g}(x) + \sum_{k=1}^{m} \hat{\theta}_k \phi_k(x)$, where $\hat{g}$ and $\hat{\theta}$ solve

$$\min_{\substack{g\in\mathscr{H},\\ \theta\in\mathbb{R}^m}} \left(\sum_{i=1}^{N} (y_i - g(x_i) - \sum_{k=1}^{m} \theta_k \phi_k(x_i))^2 + \gamma \|g\|^2_{\mathscr{H}} \right). \tag{50}$$

Note that the expansion coefficients gathered in θ are not subject to any penalty term but a low value for m avoids overfitting. An extension of the representer theorem holds (Schölkopf et al., 2001): the function estimate is

$$\sum_{i=1}^{N} \hat{c}_i K_{x_i}(x) + \sum_{k=1}^{m} \hat{\theta}_k \phi_k(x) \tag{51}$$

where the $\hat{c}_i$ and $\hat{\theta}_k$ are obtained optimizing (50) with g replaced by $\sum_{i=1}^{N} c_i K_{x_i}$. In particular, assume that $\Phi \in \mathbb{R}^{N\times m}$ is full column rank, with $N \geq m$ and $\Phi_{ij} = \phi_j(x_i)$. Then, it follows from Theorem 1.3.1 in Chapter 1 of Wahba (1990) that $\hat{\theta} = \left(\Phi^T A^{-1} \Phi\right)^{-1} \Phi^T A^{-1} Y$ and $\hat{c} = A^{-1}\left(Y - \Phi\hat{\theta}\right)$, with $A = \mathbf{K} + \gamma I_N$.

10. Kernels

Since the reproducing kernel completely characterizes the hypothesis space $\mathscr{H}$, its choice has a crucial impact on the ability of predicting future output data (generalization performance). A large variety of positive semidefinite kernel functions have been introduced in the literature over the years. In the following, we limit the attention to few illustrative examples, referring the interested reader to more comprehensive treatments of the

subject, e.g. Hofmann, Schölkopf, and Smola (2008), Schölkopf and Smola (2001) and Shawe-Taylor and Cristianini (2004).

10.1. Linear kernels and regularized linear regression ⋆

Let $\mathscr{X} = \mathbb{R}^m$ with every input location x thought of as an m-dimensional column vector. Let also $P \in \mathbb{R}^{m\times m}$ denote a symmetric and positive semidefinite matrix. Then, we can define a linear kernel as follows

$$K(x_1, x_2) = x_1^T P x_2.$$

The space $\mathscr{H}$ induced by such K is simply a space of linear functions $g : \mathbb{R}^m \to \mathbb{R}$. In fact, from Remark 11 we know that each function is a linear combination of kernel sections. In this particular case, it is not difficult to verify that for every $g(x)$ there exists $a \in \mathbb{R}^m$ such that $g(x) = K_a(x) = x^T P a$. If P is assumed of full rank, $\mathscr{H}$ contains all the functions

$$g(x) = x^T\theta, \quad \theta \in \mathbb{R}^m$$

where $\theta := Pa$. Furthermore, $\|g\|^2_{\mathscr{H}}$ is given by

$$\langle K_a(\cdot), K_a(\cdot)\rangle_{\mathscr{H}} = K(a, a) = a^T P a = \theta^T P^{-1}\theta.$$

Then, the regularization problem (45) can be reformulated (in terms of θ) as

$$\min_{\theta\in\mathbb{R}^m} \left(\|Y - \Phi\theta\|^2 + \gamma\theta^T P^{-1}\theta\right)$$

where the ith row of Φ is x_i^T. This is now exactly the ReLS formulation derived as (19) in Part I. Notice that the regularization matrix P, defining the kernel K, induces the penalty term $\theta^T P^{-1}\theta$. In addition, if x_i includes past values of the system input, the entries of θ are the unknown impulse response coefficients. This establishes a correspondence between regularized FIR estimation and RKHS regularization via linear kernels. Another correspondence will be stated in Part III (Section 11.3).

10.2. Kernels given by a finite number of basis functions ⋆

Let the map $\phi : \mathscr{X} \to \mathbb{R}^m$ be $\phi(x) = \begin{pmatrix}\phi_1(x) & \cdots & \phi_m(x)\end{pmatrix}$ and let $K(x_1, x_2) = \phi(x_1)\phi(x_2)^T$. It is easy to verify that K is a positive semidefinite kernel, and the associated RKHS coincides with the m-dimensional space spanned by the basis functions ϕ_i. The associated kernel matrix is given by $\mathbf{K} = \Phi\Phi^T$, where $\Phi \in \mathbb{R}^{N\times m}$ with $\Phi_{ij} = \phi_j(x_i)$. Then, the solution of (45) is (48) where $\hat{c}$ satisfies $\hat{\theta} = \Phi^T\hat{c}$ with

$$\hat{\theta} = \arg\min_{\theta\in\mathbb{R}^m} \left(\|Y - \Phi\theta\|^2 + \gamma\|\theta\|^2\right) \quad \text{s.t. } \theta = \Phi^T c. \tag{52}$$

In particular, one has[4]

$$\hat{\theta} = \Phi^T\left(\Phi\Phi^T + \gamma I_N\right)^{-1} Y = \left(\Phi^T\Phi + \gamma I_m\right)^{-1}\Phi^T Y$$

that again coincides with the classical ReLS formula (19) introduced in Part I with $P = I_m$.

10.3. Radial basis kernels

The class of continuous positive semidefinite kernels

$$K(x_1, x_2) = h(x_1 - x_2), \tag{53}$$

includes the class of radial basis kernels (RBF) of the form $K(x_1, x_2) = g(\|x_1 - x_2\|)$, such as the popular Gaussian kernel

$$K(x_1, x_2) = \exp\left(-\rho\|x_1 - x_2\|^2\right), \quad \rho > 0. \tag{54}$$

Differently from the kernels described in Sections 10.1 and 10.2, the RKHS associated with any non-constant RBF kernel is infinite-dimensional (it cannot be spanned by a finite number of basis functions). An explicit characterization of the associated RKHS is described in Steinwart (2002) and Steinwart, Hush, and Scovel (2006).

10.4. Spline kernels

To simplify the exposition, let $\mathscr{X} = [0, 1]$; let also $g^{(j)}$ be the jth derivative of g, with $g^{(0)} := g$. Intuitively, in many circumstances an effective regularizer is obtained by penalizing the energy of the pth derivative of g, i.e.

$$J(g) = \int_0^1 \left(g^{(p)}(x)\right)^2 dx.$$

Now, an interesting question is whether this penalty term can be cast in the RKHS theory. The answer is positive. In fact, consider the Sobolev space of functions g whose first $p - 1$ derivatives are absolutely continuous and satisfy $g^{(j)}(0) = 0$ for $j = 0, \ldots, p - 1$ (Adams & Fournier, 2003). Then, this is a particular RKHS $\mathscr{H}$ if the norm is defined by

$$\|g\|^2_{\mathscr{H}} = \int_0^1 \left(g^{(p)}(x)\right)^2 dx.$$

The corresponding kernel is the so called spline kernel

$$K(s, t) = \int_0^1 G_p(s, u)G_p(t, u)du \tag{55}$$

where

$$G_p(r, u) = \frac{(r - u)_+^{p-1}}{(p - 1)!}, \qquad (u)_+ = \begin{cases} u & \text{if } u \geq 0 \\ 0 & \text{otherwise.}\end{cases} \tag{56}$$

Note that the Laplace transform of $G_p(\cdot, 0)$ is $1/s^p$. When $p = 1$, one obtains the linear spline kernel given by

$$K(s, t) = \min\{s, t\} \tag{57}$$

whereas $p = 2$ leads to the cubic spline kernel:

$$K(s, t) = \frac{st\min\{s, t\}}{2} - \frac{(\min\{s, t\})^3}{6}. \tag{58}$$

When the kernel (55) is adopted, it follows from the representer theorem that the estimate $\hat{g}$ is a smoothing spline, whose derivatives are continuous exactly up to order $2p - 2$ (Wahba, 1990). As an example, from (58) one can see that for $p = 2$ the kernel sections K_{x_i}, whose linear combinations provide $\hat{g}$, are the well known cubic smoothing splines consisting of piecewise third-order polynomials, see also Fig. 5.

Spline functions enjoy notable numerical properties originally investigated in the interpolation scenario. In particular, piecewise polynomials avoid Runge's phenomenon (Runge, 1901) (presence of large oscillations in the reconstructed function) which e.g. arises when high-order polynomials are employed. Fit convergence rates are discussed e.g. in Ahlberg and Nilson (1963) and Atkinson (1968).

Recall that the spline kernel induces an infinite-dimensional RKHS of functions which all satisfy the constraints $g^{(j)}(0) = 0$ for $j = 0, \ldots, p - 1$. Then, to cope with nonzero initial conditions, the spline kernel is typically enriched with a low-dimensional parametric part (the bias space already discussed in Section 9.3) given by the space of polynomials up to order $p - 1$. The enriched

[4] Notice that the optimal solution θ of the unconstrained problem (52) automatically satisfies the constraint $\theta = \Phi^T c$, and therefore the latter can be dropped without loss of generality.

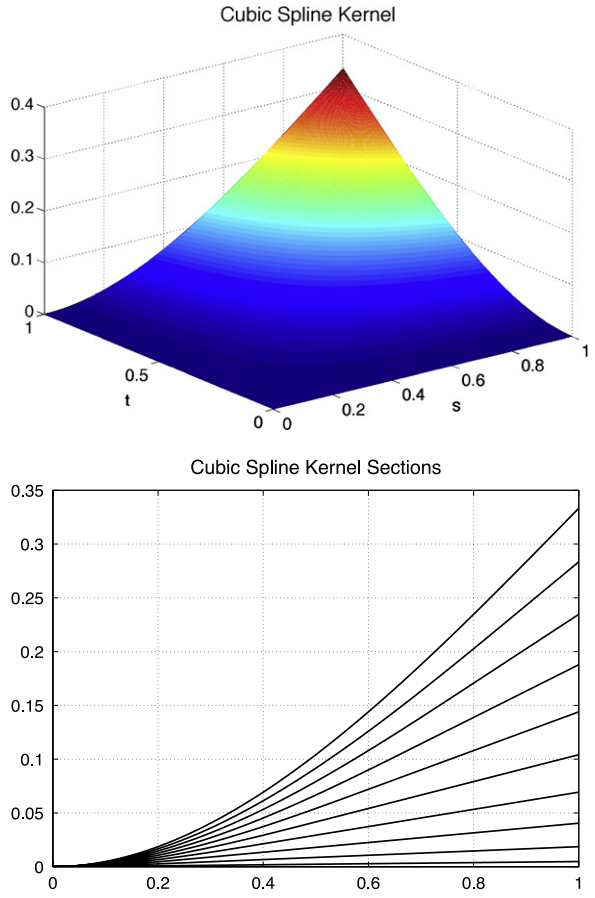

Fig. 5. Cubic spline kernel (top) and kernel sections $K_{x_i}(x)$ for $x_i = 0.1, 0.2, \ldots, 1$ (bottom).

space is $\mathscr{H} \oplus \text{span}\{1, x, \ldots, x^{p-1}\}$ and the spline estimator solves

$$\min_{\substack{g\in\mathscr{H},\\ \theta\in\mathbb{R}^p}} \left(\sum_{i=1}^{N} \left(y_i - g(x_i) - \sum_{k=1}^{p} \theta_k x_i^{k-1} \right)^2 + \gamma \int_0^1 \left(g^{(p)}(x) \right)^2 dx \right) \tag{59}$$

whose explicit solution is given by (51) setting $\phi_k(x) = x^{k-1}$ and $\Phi_{ij} = x_i^{j-1}$.

10.5. Numerical example using the cubic spline estimator

To illustrate the effect of the regularization parameter γ on the performance of ReLS, we consider a simple numerical example. The problem is the reconstruction of $e^{\sin(8x)}$, $x \in [0, 1]$, from 100 noisy samples obtained from uniform sampling of the independent variable on the unit interval. Data are corrupted by additive white Gaussian noise with standard deviation 0.3, see Fig. 6. We adopt the estimator (59) with $p = 2$. The results associated with three different values of γ are displayed in the three panels of Fig. 6. The estimate in the left panel is affected by oversmoothing: the value of γ is too large overweighting the norm of g in (59). This introduces a large bias in the estimator: the model is too rigid, unable to follow the data. The opposite situation is visible in the middle panel where a too low value for γ is used. In turn, this overweights the loss function in (59), leading to a high variance estimator: the model is overly flexible and overfits the measurements. Finally, the estimate in the right panel of Fig. 6 is obtained using an oracle which has access to the true function and minimizes the MSE. It selects that value of γ, denoted by γ_{opt}, that establishes the optimal trade-off between bias and variance, returning the regularized estimate closest to truth according to a quadratic criterion. The example demonstrates that the choice of γ can be seen as the counterpart of model order selection in the classical parametric paradigm.

10.6. Concluding remarks of the section

The regularized estimator (45) has been introduced, pointing out the importance of the kernel that defines the hypothesis space and the regularization penalty.

Regularized estimation has been widely used in applications fields related to statistics and machine learning. In recent years, it has been exploited also for nonlinear system identification and prediction, mainly resorting to Gaussian or polynomial kernels (Schölkopf & Smola, 2001). Diverse applications call for different definitions of the function domain $\mathscr{X}$. For instance, in the algorithms for NARX identification and time-series prediction described in Girard, Rasmussen, Quinonero-Candela, and Murray-Smith (2003), Leithead, Solak, and Leith (2003) and Pillonetto, Chiuso, and Quang (2011), the elements of $\mathscr{X}$ are vectors containing system input samples. Regularized approaches for estimation of partially linear models can be found in Espinoza, Suykens, and De Moor (2005), Li, Li, Su, and Chun (2006) and Xu and Chen (2009). There, the linear behavior is captured by a parametric model (the bias space) while the nonlinear distortion is estimated by kernel-based techniques. Other regularized approaches for nonlinear and state-space models identification can be found in Frigola, Lindsten, Schon, and Rasmussen (2013), Frigola and Rasmussen (2013) and Hall, Rasmussen, and Maciejowski (2012). A connection between Volterra and Wiener nonlinear system representation and the RKHS induced by the polynomial kernel, together with an efficient identification scheme, can be found in Franz and Schölkopf (2006). Other recent applications regard Wiener, Hammerstein and Wiener–Hammerstein system identification (Falck et al., 2012; Falck, Pelckmans, Suykens, & De Moor, 2009; Goethals, Pelckmans, Falck, Suykens, & De Moor, 2010; Goethals, Pelckmans, Suykens, & De Moor, 2005; Lindsten, Schön, & Jordan, 2013). For example, in Lindsten et al. (2013) the elements of $\mathscr{X}$ are the outputs coming from the linear block of the Wiener structure. A Gaussian kernel is then used to recover the static nonlinearity present in the second block. Conditions ensuring the statistical consistency of this identification scheme have been derived in Pillonetto (2013) using RKHS theory.

Part III. Continuous-time system identification as a function estimation problem

In this part, we study continuous-time system identification as a function estimation problem.

11. Impulse response estimation problem

There is a particular linear system identification problem that naturally lends itself to being solved using the function estimation techniques previously described. This happens when a linear time-invariant system is excited by an impulsive input and noisy samples of its output are collected. Then, the identification of the impulse response is perfectly equivalent to learning a univariate function of time. In the following, this simple scenario will be referred to as the static case. It is apparent that assuming an impulsive input is too restrictive in system identification where, as seen in Part I, it is necessary to consider dynamic scenarios in which the system is excited by arbitrary inputs (Ljung, 1999; Söderström & Stoica, 1989).

Hereafter, we will consider continuous-time single-input–single-output (SISO) dynamic systems. We also assume that the system is linear, time-invariant, and causal, denoting with u the

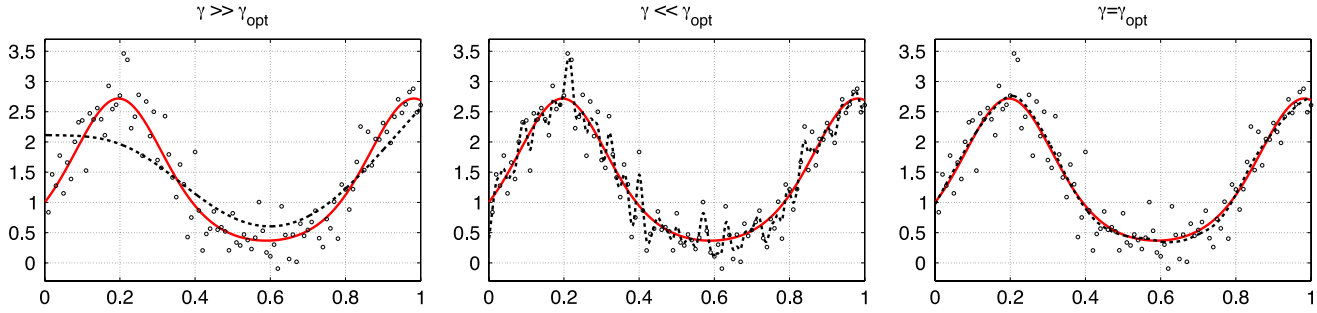

Fig. 6. Cubic spline estimates using three different values of γ: truth (solid line), noisy data ($\circ$) and estimate (dashed line).

input and with $y(t_i)$ the output measured at the time instant t_i. Assuming an output error model, one has

$$y(t_i) = \int_0^{+\infty} u(t_i - s)g(s)ds + e_i, \quad i = 1, \ldots, N \tag{60}$$

where e_i is white noise and the unknown function g is now the system impulse response. Let

$$y_i = y(t_i), \qquad Y = (y_1, y_2, \ldots, y_N)^T.$$

Then, the identification problem consists of reconstructing g starting from a known input u and Y. Compared to the static scenario described in the previous sections, the following two peculiarities can be noticed:

- the domain of the function g is one-dimensional and given by the positive real axis, i.e. $\mathscr{X} = \mathbb{R}^+$ and $g : \mathbb{R}^+ \mapsto \mathbb{R}$;
- compared to the static case, the nature of the data is more complex since, in place of $g(x_i)$, the measurement model involves the functional L_i given by

$$L_i[g] = \int_0^{+\infty} u(t_i - s)g(s)ds.$$

As mentioned in the preamble, inverting this convolution integral is an inverse problem associated with potentially ill-conditioned numerical solutions, e.g. see Twomey (1977). Ill-conditioning is particularly severe when u is a low-pass/smooth signal and worsens when the output signal is sampled more frequently, i.e. when more information comes from the system. In the case of uniform sampling, these phenomena can be given a spectral characterization using the Szegö theorem (Ekstrom, 1973). So, the problem area should benefit from the classical regularization works of Tikhonov and Phillips (Phillips, 1962; Tikhonov & Arsenin, 1977) as well as from the connection between regularization techniques for inverse problems and the function estimation paradigm of Part II, see also De Vito, Rosasco, Caponnetto, De Giovannini, and Odone (2005). In the next subsection, along the line developed in Pillonetto and De Nicolao (2010), we extend the RKHS framework to SISO linear system identification.

11.1. Regularized identification in RKHS and the representer theorem for system identification

Following the framework described in Section 9, the impulse response g is seen as an element of the RKHS $\mathscr{H}$ associated with a "causal" kernel $K : \mathbb{R}^+ \times \mathbb{R}^+ \mapsto \mathbb{R}$. For the time being, K will be a generic kernel (the choice of the most suitable kernel for system identification will be discussed later in Section 13). The use of the regularizer $\|g\|^2_{\mathscr{H}}$ to avoid ill-posedness and ill-conditioning leads to the estimator

$$\min_{g \in \mathscr{H}} \left(\sum_{i=1}^{N} (y_i - L_i[g])^2 + \gamma \|g\|^2_{\mathscr{H}} \right) \tag{61}$$

that coincides with (45) except that $L_i[g]$ replaces $g(x_i)$.

For future developments, given the system input u, it is useful to define the output kernel O as follows

$$O(t, \tau) = \int_0^{+\infty} u(t - x) \left(\int_0^{+\infty} u(\tau - a)K(x, a)da \right) dx$$

and the *output kernel matrix* $\mathbf{O} \in \mathbb{R}^{N \times N}$ as a positive semidefinite matrix with (i, j) entry given by

$$\mathbf{O}_{ij} = L_i[L_j[K]] = O(t_i, t_j). \tag{62}$$

From the Theorem 2 (representer theorem), we have seen that a remarkable feature of problem (45) is that it admits a finite-dimensional representation. One may wonder whether this holds also for the solution of (61). The answer is positive if the linear functionals L_i are continuous on $\mathscr{H}$, i.e. if

$$\forall i \quad \exists C_i < \infty : \ |L_i[g]| \leq C_i \|g\|_{\mathscr{H}}, \quad \forall g \in \mathscr{H}.$$

From RKHS theory it is known that the linear functional L_i is continuous iff $L_i[K_x]$ is a function in $\mathscr{H}$, see Aronszajn (1950) for details.

Theorem 3 (*Representer Theorem for System Identification*)**.** *If $\mathscr{H}$ is a RKHS induced by K and each $L_i : \mathscr{H} \to \mathbb{R}$ is a continuous linear functional, the minimizer of (61) is*

$$\hat{g}(x) = \sum_{i=1}^{N} \hat{c}_i L_i[K_x] \tag{63}$$

where the $\hat{c}_i$ are the components of

$$\hat{c} = (\mathbf{O} + \gamma I_N)^{-1} Y. \tag{64}$$

The result above is obtained following the same arguments of the proof of Theorem 1.3.1 in Wahba (1990), see also Yuan and Tony Cai (2010) where asymptotic properties of the estimator (63) are also discussed.

Thus, also the solution of the variational problem (61) admits a finite-dimensional representation. The difference w.r.t. the static case is that the basis functions are no more the kernel sections, but their convolution with the input u, i.e. $L_i[K_x] = \int_0^{+\infty} u(t_i - a)K(x, a)da$. As in the static case, more general versions of Theorem 3 can be found, e.g. see Dinuzzo and Schölkopf (2012).

11.2. The bias space ⋆

As in the static case, it can be useful to enrich the estimator (61) with a parametric component, e.g. a low-order rational transfer function. An approach consists of using an enlarged space, already introduced in Section 9.3, given by $\mathscr{H} + \text{span}\{\phi_1, \ldots, \phi_m\}$. Thus, the impulse response is assumed to be $g + \sum_{k=1}^{m} \theta_k \phi_k$ with $g \in \mathscr{H}$ and the regularization problem (61) is modified as follows

$$\min_{\substack{g \in \mathscr{H}, \\ \theta \in \mathbb{R}^m}} \sum_{i=1}^{N} \left(y_i - L_i \left[g + \sum_{k=1}^{m} \theta_k \phi_k \right] \right)^2 + \gamma \|g\|^2_{\mathscr{H}}. \tag{65}$$

Again, assume $\Phi \in \mathbb{R}^{N\times m}$, with $N \geq m$, of full rank and defined by $\Phi_{ij} = L_i[\phi_j]$. Then, using Theorem 1.3.1 in Wahba (1990), the estimate is

$$\sum_{i=1}^{N} \hat{c}_i L_i[K_x] + \sum_{k=1}^{m} \hat{\theta}_k \phi_k(x) \tag{66}$$

where $\hat{\theta} = \left(\Phi^T A^{-1}\Phi\right)^{-1}\Phi^T A^{-1} Y$ and $\hat{c} = A^{-1}\left(Y - \Phi\hat{\theta}\right)$, with $A = \mathbf{O} + \gamma I_N$.

11.3. Regularized FIR estimation as regularization in RKHS⋆

Even if stated in a continuous-time fashion, all the theory so far exposed holds on general function domains. Hence, it also covers the discrete-time setting. For instance, IIR systems can be handled considering $\mathscr{X} = \mathbb{N}$. Then, all the results above hold just replacing continuous-time convolutions with their discrete-time versions, i.e. $L_i[g] = \sum_{j=1}^{+\infty} u(t_i - j)g(j)$. It is also instructive to discuss in some depth the connection of the RKHS framework with the FIR case treated in Part I. This will give also some insight about the nature of the RKHS norm entering (61).

First, recall that in Part I the estimate of the vector θ containing the FIR coefficients was

$$\hat{\theta} = \arg\min_{\theta} \|Y - \Phi\theta\|^2 + \gamma\theta^T P^{-1}\theta. \tag{67}$$

Let g_θ denote the impulse response such that $g_\theta(i) = \theta_i$. We now show that there exists a suitably defined RKHS (dependent on P) such that the function estimation problem (61) is equivalent to (67) in the sense that $\hat{g}(i) = \hat{\theta}_i$, $i = 1, \ldots, m$, where $\hat{\theta}_i$ is the ith entry of $\hat{\theta}$.

To state the connection, it suffices introducing the following correspondences:

$$\mathscr{X} = \{1, 2, \ldots, m\} \tag{68a}$$

$$K : \mathscr{X} \times \mathscr{X} \to \mathbb{R} \quad \text{s.t. } K(i, j) = P_{ij} \tag{68b}$$

$$L_i[g_\theta] = \Phi(i, :)\theta \tag{68c}$$

where $\Phi(i, :)$ denotes the ith row of Φ. In fact, since P is a symmetric and positive semidefinite matrix, K is a positive semidefinite kernel. Therefore, Theorem 1 says that there is a unique RKHS $\mathscr{H}$ of real valued functions over $\{1, 2, \ldots, m\}$ having K as reproducing kernel. Notice that there exist only m kernel sections given by the m columns of P. Therefore, recalling Remark 11, each $g_\theta \in \mathscr{H}$ has the representation

$$g_\theta(x) = \sum_{j=1}^{m} a_j K(j, x) \tag{69}$$

with $\|g_\theta\|^2_{\mathscr{H}} = a^T P a$, where $a = [a_1, \ldots, a_m]^T$. Using the definition of g_θ and (69), one obtains $\theta_i = g_\theta(i) = \sum_{j=1}^{m} a_j K(j, i) = P(i, :)a$. It follows that $\theta = Pa$ and $\|g_\theta\|^2_{\mathscr{H}} = \theta^T P^{-1}\theta$. Hence, under the assumptions (68), (67) coincides with (61). From elementary algebra, we also obtain that the solution $\hat{\theta}$ admits the following two equivalent expressions

$$\hat{\theta} = (\Phi^T\Phi + \gamma P^{-1})^{-1}\Phi^T Y \tag{70a}$$

$$= P\Phi^T(\Phi P\Phi^T + \gamma I_N)^{-1} Y. \tag{70b}$$

It is interesting to notice that the last expression (70b) can be derived from the representer theorem. In fact, each basis function $L_i[K_x]$ corresponds to the ith column of $P\Phi^T$. In addition, the expansion coefficients are $\hat{c} = (\Phi P\Phi^T + \gamma I_N)^{-1} Y$. This last equivalence is in perfect agreement with (64): in fact, the last correspondence in (68), in combination with (62), leads to the output kernel matrix

$$\mathbf{O} = \Phi P\Phi^T. \tag{71}$$

Finally, the correspondence can be stated also in the case of singular P. In fact, under assumption (68), even if P is not full rank, the representer theorem still guarantees that (70b) is the solution of (61) and this coincides with the solution of (20) discussed in Remark 1.

12. Connection with Bayesian estimation of continuous-time Gaussian stochastic processes

As in the FIR case, the kernel-based regularization approach described in the previous section can be given a probabilistic interpretation in a Bayesian framework. This connection was initially studied in Kimeldorf and Wahba (1971a) in the context of spline regression, see also Girosi, Jones, and Poggio (1995), Rasmussen and Williams (2006) and Wahba (1990). The main point is to see the function estimation problem as the Bayesian estimation of a continuous-time Gaussian stochastic process on $\mathbb{R}^+$, from noisy observations.

To set up our Bayesian framework, the N entries of Y are modeled as in (60), i.e. $y_i = L_i[g] + e_i$, where

- the e_i are zero-mean Gaussian of variance σ^2, mutually independent and independent of g;
- the system impulse response g is a zero-mean Gaussian stochastic process, on $\mathbb{R}^+$, with autocovariance λK and independent of e_i.

Thus, the prior knowledge on g is given within the probabilistic description of the random process. To make an example, smoothness can be enforced by selecting a covariance K such that irregular profiles, e.g. nondifferentiable functions, are probabilistically unlikely. Notice also that giving a probabilistic description of an unknown signal is not new in the systems and control literature, a paradigmatic example being given by the theory of Wiener filtering, prediction and smoothing (Wiener, 1949).

Given (column) random vectors u and v, we define $Cov[u, v] := \mathscr{E}[(u - \mathscr{E}[u])(v - \mathscr{E}[v])^T]$. Since linear transformations of Gaussian processes preserve Gaussianity, the vector $z = [L_1[g] \quad \ldots \quad L_N[g]]$ (the noiseless output evaluated at time instants t_i) is a multivariate zero-mean normal vector. Furthermore, since $Cov(z_i, z_j) = \lambda L_i[L_j[K]]$, the covariance matrix of z is $\lambda\mathbf{O}$, where $\mathbf{O}$ is the output kernel matrix defined in (62). In view of (60) and the independence of g and e_i, it follows that $g(x)$ and Y are jointly Gaussian for any $x \in \mathbb{R}^+$. Hence, the posterior $\mathbf{p}(g(x)|Y)$ is Gaussian as well, as recalled in (25). In our case we obtain (Papoulis, 1984)

$$Cov(g(x), z_i) = \lambda L_i[K_x]$$

$$Cov(y_i, y_j) = Cov(z_i, z_j) + \sigma^2\delta_{ij} = \lambda\mathbf{O}_{ij} + \sigma^2\delta_{ij}$$

where δ_{ij} is the Kronecker Delta. Using (25) with $\gamma = \frac{\sigma^2}{\lambda}$:

$$\mathscr{E}[g(x)|Y] = [L_1[K_x] \ldots L_N[K_x]]\,(\mathbf{O} + \gamma I_N)^{-1} Y = \sum_{i=1}^{N} \hat{c}_i L_i[K_x]$$

where $\hat{c}_i$ is the ith entry of vector $\hat{c}$ defined in (64). Hence, the minimum variance estimate coincides with the solution of (61) when the proper $\mathscr{H}$ is chosen. This is summarized in the following proposition.

Proposition 13. *Consider (60) where*

- *the e_i are independent and Gaussian, of variance σ^2;*
- *the system impulse response g is a zero-mean Gaussian stochastic process, independent of the noise, with autocovariance λK.*

Let $\mathscr{H}$ be the RKHS induced by K. Then, given Y, the minimum variance estimate of $g(x)$ is $\hat{g}(x)$ where

$$\hat{g} = \arg\min_{g\in\mathscr{H}} \left(\sum_{i=1}^{N}(y_i - L_i[g])^2 + \gamma\|g\|^2_{\mathscr{H}}\right), \quad \gamma = \frac{\sigma^2}{\lambda}.$$

13. Kernels for continuous-time system identification

13.1. The Gaussian and the cubic spline kernels are not suited to impulse response estimation

Consider a very simple causal impulse response given by $g(x) = e^{-x}$, $x \in \mathbb{R}^+$. We consider the problem of estimating g when the system input is a unit impulse at 0 and 100 noisy output samples are uniformly collected in the interval [0, 0.5]. The measurement noise is additive and Gaussian with variance equal to 0.01.

We consider a Monte Carlo study of 300 runs. At each run a different set of noisy measurements is generated, an example being reported in the top left panel of Fig. 7. Notice that, in view of the impulsive nature of u, the function g can be reconstructed exploiting the static scenario described in Part II. In particular, the impulse response is estimated by solving Problem (45) adopting a quadratic loss function. Concerning the choice of the kernel, we will first test the performance of the cubic spline kernel defined in (58), equipped with the bias space accounting for non zero initial conditions, and then the Gaussian kernel reported in (54). At any run j, the optimal regularization parameter γ, as well as the Gaussian kernel width ρ, are those maximizing the fit measure

$$100\% \left(1 - \sqrt{\frac{\int_0^1 (\hat{g}_j(x) - g(x))^2 dx}{\int_0^1 (g(x))^2 dx}} \right)$$

where $\hat{g}_j$ is the impulse response estimate obtained at the jth run. Fig. 7 (top right) shows the 300 estimates of g obtained by the cubic spline kernel. One can see that $g(0)$ tends to be underestimated and that the variance of the estimator is large. In fact, many estimates exhibit oscillations and $\hat{g}_j$ tends to diverge as x increases. The average fit is 79%.

The bottom left panel shows the same kind of results but obtained using the Gaussian kernel. One can see that it outperforms the cubic spline kernel but the variance of the estimator is still large. Oscillations are visible in many estimates $\hat{g}_j$ and the average fit is 89%.

This system identification problem is rather simple because g is very basic and its noisy samples are directly available. However, even if such a favorable setting is seldom encountered in practice, this example demonstrates that a straight application of standard machine learning techniques, based on kernels including only smoothness information, is doomed to produce poor estimates in the system identification field.

13.2. The notion of stable kernel and sufficient conditions for kernel stability

Here, and in the sequel, all the introduced kernels are causal, i.e. different from zero only on $\mathbb{R}^+ \times \mathbb{R}^+$.

The poor performances of the Gaussian and spline kernels in impulse response identification stem from the lack of constraints on system stability. Let us discuss this issue under the deterministic RKHS framework. The necessary and sufficient condition for a system to be BIBO stable is that $g \in \mathscr{L}^1$, where $\mathscr{L}^1$ is the space of functions on $\mathbb{R}^+$ such that

$$\int_{\mathbb{R}^+} |g(x)| dx < \infty.$$

Therefore, the impulse response should be searched for in a RKHS which is a subspace of $\mathscr{L}^1$.

Definition 14. Let $\mathscr{H}$ be the RKHS induced by a kernel K. Then, K is said to be stable if $\mathscr{H} \subset \mathscr{L}^1$.

It turns out that integrability is a sufficient condition for a kernel to be stable (Carmeli, Vito, & Toigo, 2006).

Proposition 15. *Let $\mathscr{H}$ be the RKHS induced by K. Then,*

$$\int_{\mathbb{R}^+} \int_{\mathbb{R}^+} |K(x_1, x_2)| dx_1 dx_2 < \infty \Longrightarrow \mathscr{H} \subset \mathscr{L}^1. \tag{72}$$

In addition, considering only nonnegative-valued kernels K^+, i.e. $K^+(x_1, x_2) \geq 0$, $\forall x_1, x_2 \in \mathbb{R}^+$, the condition becomes also necessary:

$$\int_{\mathbb{R}^+} \int_{\mathbb{R}^+} K^+(x_1, x_2) dx_1 dx_2 < \infty \iff \mathscr{H} \subset \mathscr{L}^1. \tag{73}$$

Hence, when the impulse response is searched within a RKHS, an integrable kernel enforces BIBO stability.

Now, consider the Gaussian kernel (54) which is nonnegative. For every $\rho > 0$, it is easy to see that it does not satisfy the necessary and sufficient stability condition (73)

$$\int_{\mathbb{R}^+} \int_{\mathbb{R}^+} \exp\left(-\rho(x_1 - x_2)^2\right) dx_1 dx_2 = +\infty.$$

This implies that the Gaussian kernel is not stable, see also Minh (2006). Hence, it is not suited for system identification. The same holds for the spline kernels (57) and (58) over $\mathbb{R}^+ \times \mathbb{R}^+$.

Remark 16. Notice that every kernel can be easily made stable just by truncation, i.e. setting $K(x_1, x_2) = 0$ for $x_1, x_2 > T$. However, stable kernels, e.g. such that $K(x, x) > 0$ for all $x \in \mathbb{R}^+$, encode the information that the variability of g is asymptotically decreasing. For instance, in Fig. 7 the Gaussian kernel performance would not be satisfactory also introducing a truncation as oscillations would still be present.

Remark 17. The limitations of the Gaussian and spline kernels for impulse response estimation can also be understood resorting to the Bayesian interpretation of regularization, where the kernels are seen as the covariances of Gaussian processes. Considering e.g. the radial basis class, the covariance admits the representation $h(|x_1 - x_2|)$, so that it can only describe stationary stochastic processes. This means that the variance of g is constant over time whereas the variability of a stable impulse response is expected to be asymptotically decreasing. Spline kernels are even worse because the signal variance is asymptotically increasing. This explains the undue oscillations affecting the estimates in the top right and bottom left panel of Fig. 7.

13.3. The necessary and sufficient condition for kernel stability ⋆

For $1 \leq p \leq \infty$, let $\mathscr{L}^p$ the classical Lebesgue space of p-power integrable functions on $\mathbb{R}^+$. In particular, $\mathscr{L}^\infty$ denotes the space of the essentially bounded functions with respect to the Lebesgue measure (Rudin, 1987). The following definition derives from the more general Definition 4.1 in Carmeli et al. (2006).

Definition 18. Let $1 \leq p \leq \infty$ and $q = \frac{p}{p-1}$ with the convention $\frac{p}{p-1} = \infty$ if $p = 1$ and $\frac{p}{p-1} = 1$ if $p = \infty$. Then, the kernel K is said to be q-bounded if

(1) the kernel section $K_x \in \mathscr{L}^p$ for almost all x, i.e. for every $x \in \mathbb{R}^+$ except on a set of null Lebesgue measure,
(2) the function $\int_{\mathbb{R}^+} K(x, a) f(a) da \in \mathscr{L}^p$ for all $f \in \mathscr{L}^q$.

It turns out that q-boundedness is the key condition for a kernel to induce a RKHS contained in $\mathscr{L}^p$. In turn, this implies that the concepts of stable kernel and ∞-bounded kernel are equivalent, see also Proposition 4.2 in Carmeli et al. (2006).

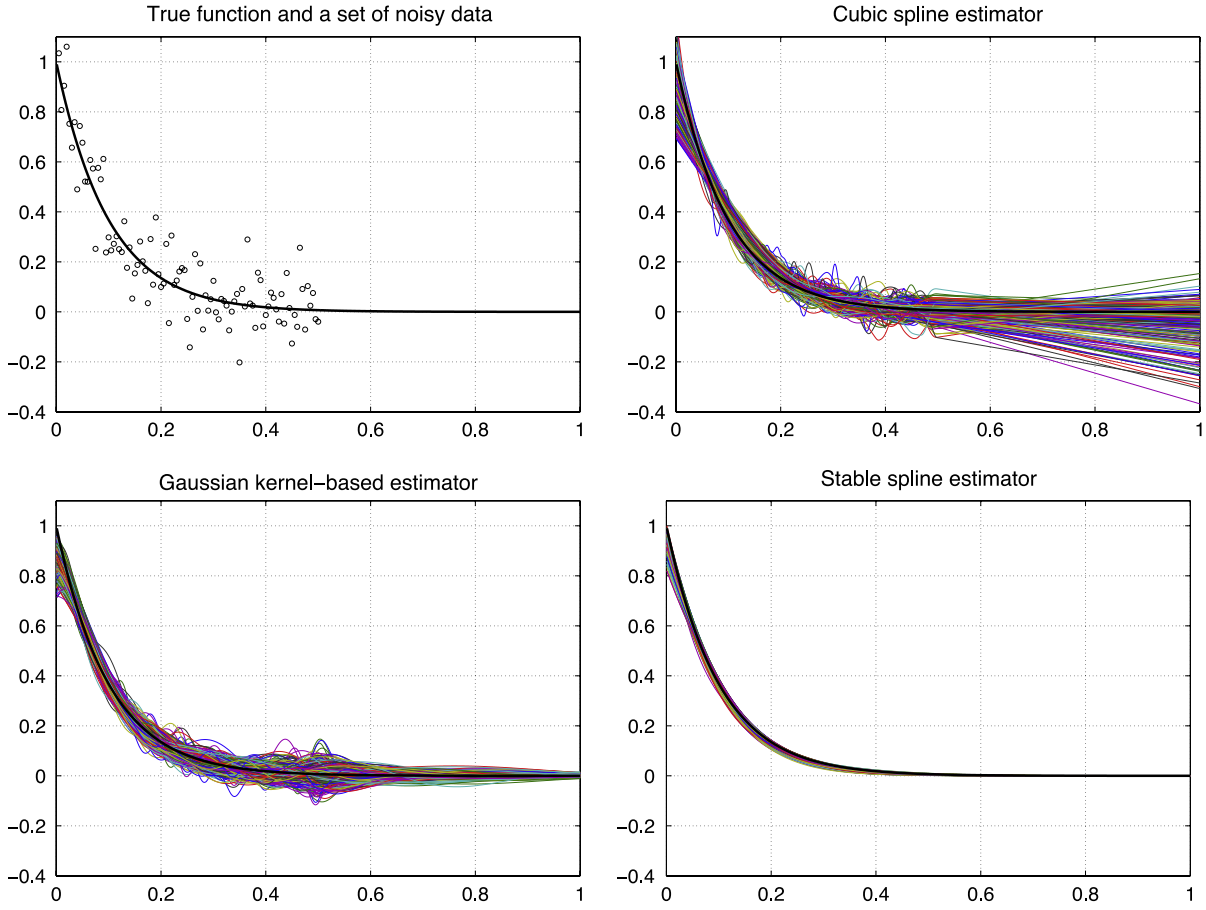

Fig. 7. The left top panel shows the true impulse response (thick line) and one out of the 300 noisy data sets (∘). The other three panels display the true impulse response (thick line) and the 300 impulse response estimates obtained using the cubic spline, the Gaussian and the stable spline kernel. The optimal regularization parameters, minimizing the reconstruction error, are adopted at any run.

Theorem 4. *Let $\mathscr{H}$ be the RKHS induced by K. Then, one has*

$$\mathscr{H} \subset \mathscr{L}^p \iff K \text{ is } q\text{-bounded}.$$

In particular, setting $q = \infty$, one obtains

$$\mathscr{H} \subset \mathscr{L}^1 \iff \int_{\mathbb{R}^+} \left| \int_{\mathbb{R}^+} K(x,a) f(a) da \right| dx < +\infty, \quad \forall f \in \mathscr{L}^\infty. \tag{74}$$

Some final remarks are in order. It can be shown that the integrability of a kernel K implies its ∞-boundedness, see Corollary 4.1 in Carmeli et al. (2006), and this leads to the stability sufficient condition (72). Also, if a nonnegative-valued and ∞-bounded kernel is given, using $f = 1 \in \mathscr{L}^\infty$ in (74), one obtains that the kernel is also integrable. Hence, when nonnegative-valued kernels are considered, ∞-boundedness and integrability are equivalent concepts and this leads to (73).

13.4. The optimal continuous-time kernel

In this subsection, we generalize the concept of best regularization matrix derived in Section 4.2 for the FIR case ($\mathscr{X} = \{1, 2, \ldots, m\}$) to the continuous-time setting ($\mathscr{X} = \mathbb{R}^+$). Assume that the data have been generated by (60) for a certain "true" impulse response g_0. We use $\bar{g}_0$ and $\hat{\bar{g}}$ to denote any finite-dimensional vector obtained by sampling on the same arbitrary input locations g_0 and its estimate $\hat{g}$ from (61). We can then design K minimizing the MSE given by

$$MSE_{\hat{g}} = \mathscr{E}\left[(\hat{\bar{g}} - \bar{g}_0)(\hat{\bar{g}} - \bar{g}_0)^T\right]. \tag{75}$$

Then, the next result (whose proof is reported in the Appendix) shows that there exists one choice leading to the optimal kernel, which is the natural generalization of that described in discrete-time by Proposition 3.

Proposition 19 (*Optimal Kernel for a Given g_0*)**.** *Consider (61) with $\gamma = \sigma^2$ and $\hat{K}$ the kernel defined by*

$$\hat{K}(x_i, x_j) = g_0(x_i) g_0(x_j), \quad (x_i, x_j) \in \mathscr{X} \times \mathscr{X}. \tag{76}$$

Then, for every kernel K, the matrix in (75) obeys the following matrix inequality

$$MSE_{\hat{g}}(\hat{K}) \preceq MSE_{\hat{g}}(K), \tag{77}$$

i.e. the matrix $MSE_{\hat{g}}(K) - MSE_{\hat{g}}(\hat{K})$ is positive semidefinite for any K.

Hence, (76) defines the RKHS to be used in (61) to obtain the best achievable performance of ReLS.

As done in the discrete-time setting, we can ask a related question, still from a frequentist perspective: Over a certain set of randomized true impulse responses, modeled as stochastic processes with $\mathscr{E} g_0(x_i) g_0(x_j) = K(x_i, x_j)$, what is the regularizer which, coupled with a quadratic loss, leads to the best *average* MSE? The answer is in the following proposition (the proof is

similar to that of Proposition 19 contained in the Appendix and is therefore omitted).

Proposition 20. *Consider (61) with $\gamma = \sigma^2$. Then, the best average (expected) MSE for a random true system g with $\mathscr{E} g_0(x_i)g_0(x_j) = K(x_i, x_j)$ is obtained using the RKHS induced by K as hypothesis space and $J(g) = \|g\|^2_{\mathscr{H}}$ as regularizer.*

Note that Proposition 20 is an exact counterpart of Proposition 4. It has also analogies with Proposition 13 but here the impulse response can be a non Gaussian stochastic process.

13.5. Stable spline kernels and their relationship with TC and SS

We now discuss a class of exponentially stable kernels.

We start noticing that, according to Proposition 19, the optimal kernel is such that its diagonal $K(t, t)$, as well as its off diagonals, decay exponentially to zero. The key idea developed in Pillonetto and De Nicolao (2010) to build a stable kernel which has these two features is an exponential change of coordinates to remap $\mathbb{R}^+$ into $[0, 1]$, then using a spline kernel for functions defined there. This leads to the class of so called *stable spline kernels* which, by construction, inherit all the approximation capabilities of the spline curves (Atkinson, 1968), but, different from them, are intrinsically stable. This class has been also derived using Bayesian and maximum entropy arguments: in some sense it represents the least committing priors when smoothness and stability is the sole information on g (Pillonetto & De Nicolao, 2011).

Now, let K be the general spline kernel on $[0, 1]^2$ reported in (55). Then, it follows from the discussion above that the stable spline kernel S of order p is defined by

$$S(x_1, x_2) = K(e^{-\beta x_1}, e^{-\beta x_2}), \quad (x_1, x_2) \in \mathbb{R}^+ \times \mathbb{R}^+. \tag{78}$$

Notice that S depends on $\beta \in \mathbb{R}^+$ that regulates the change of coordinates and can be thought of as a kernel parameter related to the dominant pole of the system.

Using Proposition 15, one can verify that S is stable for every $\beta > 0$ and order p, whose choice controls the degree of regularity of g. Typical values are $p = 1$ or $p = 2$. Setting $p = 1$ one obtains

$$S(x_1, x_2) = e^{-\beta \max(x_1, x_2)}. \tag{79}$$

Letting $\alpha = \exp(-\beta)$, one can see that the sampled version of this kernel becomes $\alpha^{\max(i,j)}$, for $i, j \in \mathbb{N}$, which coincides with the TC kernel (32).

The choice $p = 2$ instead leads to the second-order stable spline kernel (see also Fig. 8), i.e.

$$S(x_1, x_2) = \frac{e^{-\beta(x_1+x_2+\max(x_1,x_2))}}{2} - \frac{e^{-3\beta \max(x_1,x_2)}}{6}. \tag{80}$$

Letting again $\alpha = \exp(-\beta)$, its sampled version now leads to the SS kernel (33). Hence, remarkably, both the TC and SS kernels introduced in Part I are deeply connected with the smoothing splines.

Let us now investigate the nature of the RKHS induced by the stable spline kernels, focusing just on the case $p = 1$. The RKHS norm associated with (79) is Pillonetto et al. (2010)

$$\|g\|^2_{\mathscr{H}} = \int_{\mathbb{R}^+} \left(g^{(1)}(x)\right)^2 \frac{e^{\beta x}}{\beta} dx.$$

This equation gives insights on the nature of the hypothesis space: compared to the classical Sobolev space induced by the linear spline kernel, the norm, besides the energy of the first-order derivative of g, includes also a weight proportional to $e^{\beta x}$. Thus, S induces a space of continuous and stable functions which decay to zero at least exponentially.

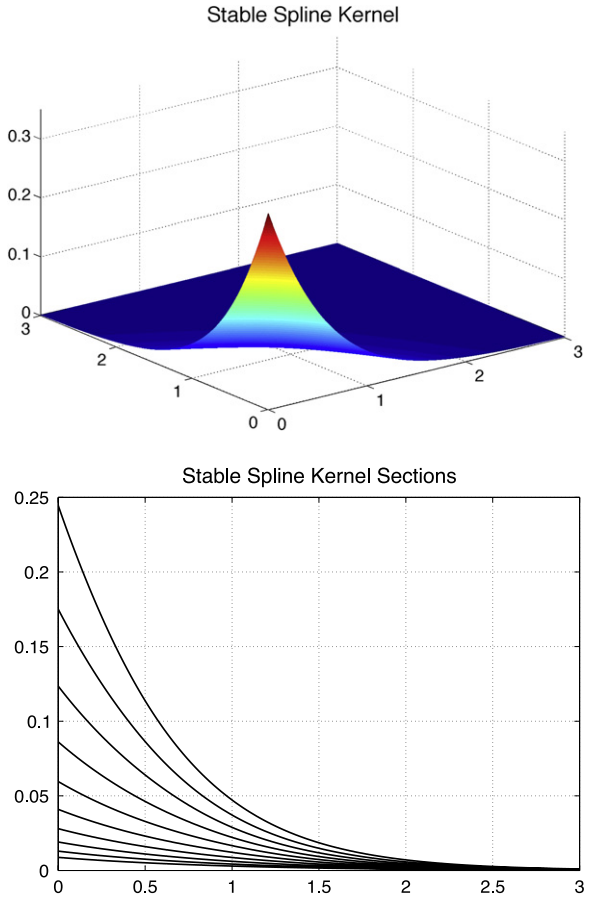

Fig. 8. Stable spline kernel of order $p = 2$, $\beta = 1$ (top) and kernel sections $K_{x_i}(x)$ for $x_i = 0.2, 0.4, \ldots, 2$ (bottom).

Let us now reconsider the simulation study of Section 13.1. The bottom right panel of Fig. 7 plots the 300 impulse response estimates obtained using the stable spline kernel of order 2 and the optimal values of γ and β computed at any run. The benefit of including the stability constraint in the nonparametric estimator is apparent: all the estimates are close to the true impulse response. The average fit is 96.6%.

14. Tuning of the design parameters

Tuning of the kernel and regularization parameters contained in the vector $\eta \in \Gamma$ is the counterpart of model order selection in the classical parametric paradigm. Therefore, it has a major impact on the identification performance. Some effective tuning methods are described in the following.

14.1. Marginal likelihood optimization

The technique here introduced was already briefly discussed in Section 4.4 and adopted in the numerical experiments of Section 7 in Part I. It is rooted into the probabilistic interpretation of problem (19) and its more general version (61). The impulse response is seen as a zero-mean Gaussian stochastic process (or random vector) of covariance λK and the vector η contains the scale factor λ, the parameters entering the kernel K and possibly also the noise variance σ^2.

The hyperparameter vector η is estimated by optimizing the so called marginal likelihood $\mathbf{p}(Y|\eta)$, i.e. the joint density $\mathbf{p}(Y|g, \eta)\mathbf{p}(g|\eta)$ where the dependence on the unknown impulse

response is integrated out. In particular, define $Z(\eta) = \lambda\mathbf{O}(\eta) + \sigma^2 I_N$ where the output kernel matrix $\mathbf{O}$ was defined in (62). Then, exploiting the Gaussianity of g and the measurements noise, one has

$$\hat{\eta} = \arg\min_{\eta\in\Gamma} Y^T Z(\eta)^{-1} Y + \log\det(Z(\eta)), \quad \text{MargLik}. \tag{81}$$

Note that, when the discrete-time version of $\mathbf{O}$ given by (71) is used, (81) coincides with (29) introduced in Part I, with $Z(\eta)$ given by (28).

By relying upon marginalization, this tuning method relates to the concept of Bayesian evidence and embodies the Occam's razor principle, i.e. unnecessarily complex models are automatically penalized. See Cox (1946) and MacKay (1992) for nice discussions and also Section 6.6 of MacKay (1992) for a comparison between this approach and cross validation. Some theoretical results which corroborate its robustness, by finding connections with MSE minimization, are described in Aravkin, Burke, Chiuso, and Pillonetto (2012) and Carli, Chen, Chiuso, Ljung, and Pillonetto (2012).

Once η is determined, following the Empirical Bayes paradigm (Berger, 1985; Maritz & Lwin, 1989), the impulse response can be computed by (61) setting $\gamma = \sigma^2/\lambda$ and using the representer theorem. Full Bayes approaches have been also developed exploiting stochastic simulation techniques, e.g. Markov chain Monte Carlo (Andrieu, Doucet, & Holenstein, 2010; Gilks, Richardson, & Spiegelhalter, 1996; Ninness & Henriksen, 2010). In this context, also η is seen as a random vector and the posterior of g and η is recovered in sampled form. Hence, the final estimate of g, as well as its Bayes intervals, account also for the uncertainty of the hyperparameters, e.g. see Magni, Bellazzi, and De Nicolao (1998) and Pillonetto and Bell (2007).

The pointwise evaluation of the marginal likelihood using (81) takes $O(N^3)$ operations. In the FIR or ARX case treated in Part I and Section 11.3, when $m \ll N$ it is useful to resort to the equivalent expression

$$\hat{\eta} = \arg\min_{\eta\in\Gamma} (N-m)\log(\sigma^2) + \log(\det(\sigma^2 I_m + \lambda P\Phi^T\Phi)) + \frac{Y^T Y}{\sigma^2} - Y^T\Phi\left(\frac{\sigma^4}{\lambda}P^{-1} + \sigma^2\Phi^T\Phi\right)^{-1}\Phi^T Y$$

which reduces the computational load to $O(Nm^2)$, see Chen and Ljung (2013) for other implementation details. Many efficient approximations of the marginal likelihood for the general case have been also developed, see Carli, Chiuso et al. (2012), Lázaro-Gredilla, Quiñonero-Candela, Rasmussen, and Figueiras-Vidal (2010), Quiñonero-Candela and Rasmussen (2005) and references therein.

14.2. C_p statistics

In all the remaining part of the section, the impulse response is no more interpreted as a stochastic process, but is a deterministic function. Considering (61), the vector η now contains the regularization parameter γ and the parameters entering the kernel K.

For future developments, exploiting the structure of the impulse response estimate $\hat{g}$ in (63), it is useful to note the following linear relationship between the output estimates $\hat{y}_i(\eta) := L_i[\hat{g}]$ collected in the vector $\hat{Y}(\eta)$ and the output measurements in Y:

$$\hat{Y}(\eta) = H(\eta)Y \tag{82}$$

where $H(\eta) = \mathbf{O}(\eta)\,(\mathbf{O}(\eta) + \gamma I_N)^{-1}$ with $\mathbf{O}$ given by (62) in continuous-time or by (71) in discrete-time. The trace of the matrix $H(\eta)$ in (82) corresponds to the so called *degrees of freedom* associated with the estimator (61):

$$d_f(\eta) = \operatorname{tr}(H(\eta)). \tag{83}$$

It can be verified that $0 \le d_f(\eta) \le N$ and that, provided $\mathbf{O}$ is full rank, $d_f(\eta)$ varies from N to 0 as the regularization parameter γ (which, as said, is a component of η) goes from 0 to $+\infty$. The fact that $d_f(\eta)$ is a real number is in agreement with the nature of the regularization method: the flexibility of the model (its degree of freedom) can be changed with continuity through the tuning of the regularization parameter, as also illustrated in Fig. 6.

Now, recall that $L_i[g]$ is the noiseless system output at instant t_i. Then, the error

$$\mathscr{E}\left[\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i(\eta) - L_i[g])^2\right] \tag{84}$$

(where expectation is taken only w.r.t. the measurement noise) provides an indication of model capability to predict outputs from the system fed with an input similar to that entering the identification data. It comes that minimization of an estimate of (84) is another viable way to perform selection of the design parameter vector η. In particular, when the noise variance is known,[5] an unbiased estimator of (84) is related to the C_p statistic (Mallows, 1973) that leads to the following estimator for η:

$$\hat{\eta} = \arg\min_{\eta\in\Gamma} \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i(\eta))^2 + \frac{2d_f(\eta)}{N}\sigma^2, \quad \text{CP}. \tag{85}$$

It is interesting to note that, for σ^2 known, the above expression coincides with the AIC criterion discussed in Part I (Section 3.1) except that the dimension m of the vector θ which parameterized the model structure, i.e. $\dim(\theta)$, is now replaced by the degrees of freedom parameterized by η, i.e. $d_f(\eta)$ as defined in (83).

14.3. Cross validation, PRESS and GCV

In Section 3.2, we have already introduced one of the basic variants of cross-validation, namely hold-out validation. Multi-stage versions of cross-validation divide the available data set into several complementary subsets. Each stage is a holdout validation step where some of the subsets are used to train the model and the remaining ones are used to evaluate predictive performances. The procedure is repeated multiple times by rotating on the choice of estimation and validation data sets. Eventually, a overall CV score is computed by averaging the scores obtained at each stage, whereby CV is an estimate of the predictive capability of the model.

One of the most common variant is *k-fold* CV where the data are partitioned in k disjoint subsets (folds) of about the same size, and then k estimation rounds are performed. The partition can be obtained by various means, ranging from a completely random sampling to a carefully designed *stratified* sampling aimed at keeping the different folds balanced in some way. At each round, one of the folds plays the role of validation set while the remaining $k-1$ are used for estimation. Validation performances are then averaged out over the rounds to produce the CV score. Performing *k-fold* CV with a large number of folds tends to reduce the bias in the estimate of the prediction score, at the expense of a higher variability. On the other hand, a low number of folds makes the estimate more biased but also more stable. Often in practice, a

[5] If σ^2 is unknown, it can be preliminarily estimated and then plugged into the C_p expression (85).

heuristic choice of the number of folds, such as $k = 5$, 10, already leads to satisfactory performances.

Leave-one-out validation is an extreme case of k-fold cross-validation discussed in Section 3.2 where $k = N$ (the validation set at each round contains only one point). The leave-one-out score with a quadratic loss is known as PRESS (predicted residual sums of squares) (Allen, 1974; Wang & Cluett, 1996). For large data sets, PRESS evaluation for a given η may seem expensive, requiring the computation of a number of function estimates equal to the number of data points. However, in the case of the linear estimator (82), PRESS evaluation reduces to a single model estimation (Wahba, 1990, Theorem 4.2.1) and the estimate of η is

$$\hat{\eta} = \arg\min_{\eta\in\Gamma} \frac{1}{N}\sum_{i=1}^{N}\left(\frac{y_i - \hat{y}_i(\eta)}{1 - h_{ii}(\eta)}\right)^2, \quad \text{PRESS} \tag{86}$$

where h_{ii} is the ith diagonal element of the matrix H.

By replacing each h_{ii} in (86) with their average, one obtains an approximation of PRESS known as Generalized Cross Validation (GCV) (Craven & Wahba, 1979; Golub, Heath, & Wahba, 1979). Its formulation involves the degrees of freedom $d_f(\eta)$ defined in (83): the hyperparameter estimate is

$$\hat{\eta} = \arg\min_{\eta\in\Gamma} \frac{1}{N}\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i(\eta))^2}{(1 - d_f(\eta)/N)^2}, \quad \text{GCV}.$$

In common with PRESS, a practical advantage of GCV over the statistics described in the previous subsection is that it does not require estimating the variance σ^2. GCV has also other interesting properties, such as invariance to rotations of Y, that sometimes makes it a more desirable estimator than PRESS, see Golub et al. (1979) and Wahba (1990).

15. Computational issues

We review some methods for the numerical computation of $\hat{g}$ in (61) relying on the relationship between machine learning and convex optimization (Bennett & Parrado-Hernandez, 2006; Bottou, Chapelle, DeCoste & Weston, 2007; Rockafellar, 1970).

Regardless of the nature of the function domain $\mathscr{X}$, the representer theorem implies that $\hat{g}$ is the sum of N basis functions with the optimal weights solving the system of linear equations (64). If the data set size N is large, plain application of a solver with cost $O(N^3)$ can be highly inefficient. Therefore, many alternative schemes have been developed.

Firstly, observe that in the case of regularized FIR estimation, that, as discussed in Section 11.3, is a special case of (61), one may well have $m \ll N$. So, it can be advantageous to use the formulation (70a) which provides $\hat{g}$ at the cost of $O(Nm^2)$ operations. More generally, for any $\mathscr{X}$, the direct solution of the N-dimensional linear system can be avoided by using approximate representations of the kernel function (Bach & Jordan, 2005; Kulis, Sustik, & Dhillon, 2006), based e.g. on the Nyström method or greedy strategies (Smola & Schölkopf, 2000; Williams & Seeger, 2000; Zhang & Kwok, 2010). Another effective approximation relies on an application of the Mercer theorem (e.g. see Section 2 in Cucker and Smale (2001)) ensuring that the kernel K admits an expansion in terms of eigenfunctions ψ_j and corresponding eigenvalues ζ_j, taken in descending order without loss of generality. A pth order approximation of $K(x, a)$ is $\sum_{j=1}^{p}\zeta_j\psi_j(x)\psi_j(a)$: this is a low-order kernel associated with a subspace $\mathscr{S}_p$ of H spanned by $\{\psi_j\}_{j=1}^{p}$, see Section 3 in Cucker and Smale (2001). Interestingly, after computing the various $L_i[\psi_j]$, a solution of (61) with H replaced by $\mathscr{S}_p$ can be obtained with $O(Np^2)$ operations. The solution obtained in this way may provide accurate approximations of $\hat{g}$ also when $p \ll N$, see Ferrari-Trecate, Williams, and Opper (1999), Pillonetto and Bell (2007), Zhu and Rohwer (1996) and Zhu, Williams, Rohwer, and Morciniec (1998). The efficacy of such approximation method for system identification has been shown in Carli, Chiuso et al. (2012), exploiting the closed form expansions of the stable spline kernels (79) and (80) reported in Pillonetto et al. (2010). A family of approximations based on the concept of pseudo input locations are also reviewed in Lázaro-Gredilla et al. (2010), Quiñonero-Candela and Rasmussen (2005) and Snelson and Ghahramani (2006).

If N is very large and it is not possible to store the entire kernel matrix in memory, another alternative is given by so called decomposition methods (List & Simon, 2004, 2007). The idea underlying them is simple: a subset of the coefficients c_i, called working set, is selected, and the associated low-dimensional subproblem is solved. Thus, only the corresponding entries of the output kernel matrix need to be loaded into the memory. An extreme case of decomposition method is coordinate descent, where the working set contains only one coefficient, e.g. see Dinuzzo (2011). In this case, the algorithm updates a single c_i at each iteration by solving a sub-problem of dimension one. This approach is becoming popular in machine learning and statistics, e.g. state-of-the-art solvers for large-scale supervised learning, such as glmnet (Friedman, Hastie, & Tibshirani, 2010) for generalized linear models, are based on these techniques.

16. Continuous-time example: estimation of cerebral hemodynamics

The quantitative assessment of cerebral hemodynamics is crucial to understand brain function in both normal and pathological states. For this purpose, an important technique is bolus-tracking magnetic resonance imaging (MRI), which relies upon the principles of tracer kinetics for nondiffusible tracers (Zierler, 1962, 1965). Interestingly, in this scenario quantification of cerebral hemodynamics corresponds to solving a time-invariant linear system identification problem (Calamante, Thomas, Pell, Wiersma, & Turner, 1999). In fact, the input is the measured arterial function while the output is the tracer concentration within a given tissue volume of interest. The system impulse response g (proportional to the so called *tissue residue function*) carries fundamental information on the system under study, e.g. the cerebral blood flow is given by the maximum of g.

We consider the same simulation described in Zanderigo, Bertoldo, Pillonetto, and Cobelli (2009). The known system input is a typical arterial function given by $u(t) = (t-10)^3 e^{-\frac{2t}{3}}$ for $t \geq 10$ and null elsewhere, while the impulse response is the (causal part of the) Lorentzian[6] in the top panel of Fig. 9 (thick line). The noiseless output is reported in the bottom panel of Fig. 9 (thick line). We consider a Monte Carlo study where, at every run, the impulse response has to be estimated from a distinct set of 80 noisy output samples. The measurements affected by pseudorandom noise at the first run are shown in the bottom panel of Fig. 9 (○). Data are realistic, and representative of an extremely ill-conditioned problem, generated as detailed in subsection II.A of Zanderigo et al. (2009), using parameters typical of a normal subject and a signal to noise ratio equal to 20.

[6] We have also used three other benchmark impulse responses present in Zanderigo et al. (2009) which well represent typical tissue residue functions. Results (not shown) are similar to those here described.

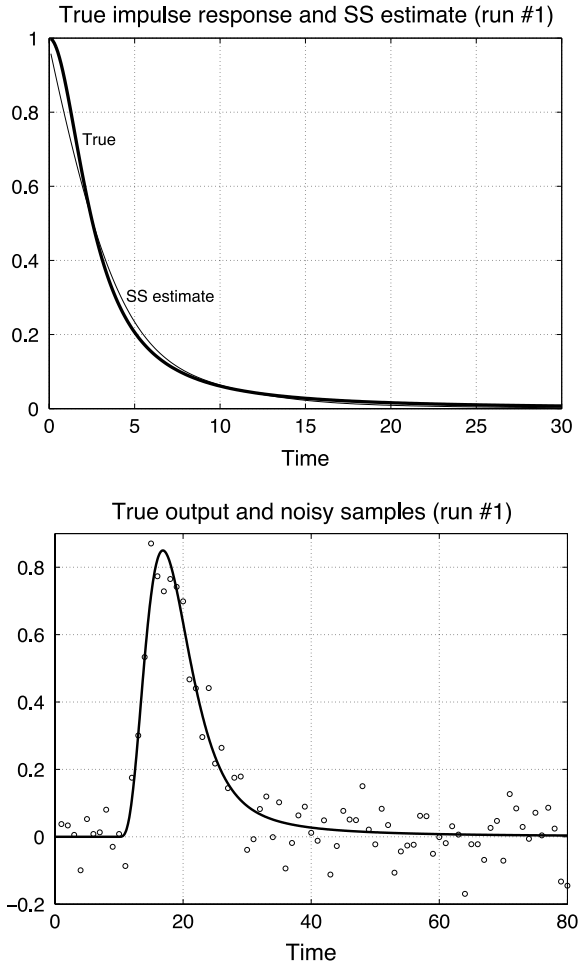

Fig. 9. Assessment of cerebral hemodynamics using magnetic resonance imaging (Section 16). The top panel shows the true system impulse response (thick line) and the SS estimate obtained at the first Monte Carlo run (thin line). The noiseless output (solid line) and the measurements (○) are displayed in the bottom panel.

The following five estimators are adopted:

- *Lag + Or*. The unknown g is modeled as the sum of Laguerre basis functions defined, in the Laplace domain, by

$$\frac{(s-p)^{k-1}}{(s+p)^k}, \quad p > 0,\ k = 1, 2, \ldots.$$

  At every run, the expansion coefficients are estimated by LS setting p and the number of basis functions to the values minimizing the MSE, i.e. leading to the estimate closest to truth according to a quadratic criterion. In particular, the value of p is searched over the grid $[0.01, 0.05, 0.1, 0.15, \ldots, 0.6]$ while the maximum allowed number of basis functions is 30. This is an ideal tuning, not implementable in practice, that provides the upper bound on the performance of a LS estimator based on Laguerre expansions.
- $\{Lag + AICc, Lag + BIC\}$. The same as above except that the number of Laguerre functions is chosen by AICc or BIC assuming σ^2 unknown.
- $\{SS, SS_1\}$. They are defined by (61), with the first-order (SS_1) or the second-order (SS) stable spline kernel, see (79) and (80). The noise variance and the hyperparameters (λ, β) are obtained via marginal likelihood optimization.

The top panel of Fig. 9 shows that the estimate returned by *SS* at the first run (continuous line) is close to the true impulse response. Indeed, the Monte Carlo study shows that *SS* performs better than the other 4 estimators. In particular, the boxplots of the fits obtained by all the estimators are displayed in Fig. 10. One can see that the fit returned by *SS* is often even better than that by *Lag + Or* that uses additional information not available to *SS*. Notice also that the regularized estimators outperform the classical approaches exploiting Laguerre functions and AICc or BIC.

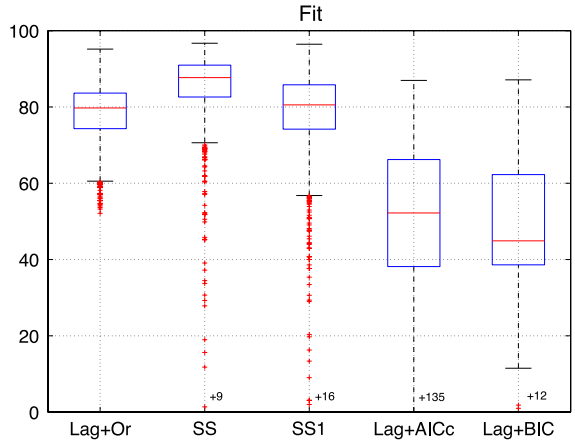

Fig. 10. Assessment of cerebral hemodynamics using magnetic resonance imaging (Section 16). Boxplots of the fits achieved by the five estimators after the 1000 runs. Recall that the estimator *Lag + Or* is not implementable in practice.

17. Other regularization techniques for system identification

This survey is focused on regularization techniques for system identification based on quadratic penalty terms that exploit suitable positive semidefinite kernels embodying information about the system to be identified. However, other types of regularizers have been and are being investigated in the context of system identification. With reference to the discrete-time case, these alternative regularization schemes amount to solving

$$\arg\min_g \sum_{t=1}^{N} \left(y(t) - \sum_{k=1}^{m} g(k)u(t-k) \right)^2 + \gamma J(g) \tag{87}$$

where J is different from the standard quadratic penalty.

An example is given by the use of the ℓ_1 norm penalizer, an approach popularized in the machine learning literature by so called LASSO methods for joint estimation and variable selection (Tibshirani, 1994). In our context, this amounts to solving (87) with $J(g) = \sum_{k=1}^{m} |g(k)|$. The main difference w.r.t. the quadratic penalty is that this regularizer promotes sparsity, i.e. it has the capability to force to zero several coefficients of the estimated impulse response. This estimator and other variants based on ℓ_1 regularizer have been investigated in the system identification literature (Rojas & Hjalmarsson, 2011; Rojas, Wahlberg, & Hjalmarsson, 2013; Toth, Hjalmarsson, & Rojas, 2012; Welsh, Rojas, Hjalmarsson, & Wahlberg, 2012).

Another example is given by methods that construct penalty terms based on the so-called nuclear norm. The nuclear norm (or trace norm) of a matrix A can be defined as the sum of its singular values:

$$\|A\|_* = \sum_i \sigma_i(A).$$

Since the nuclear norm coincides with the convex envelope of the rank function on the spectral ball, it has been often used as a convex surrogate of the rank function, following the popular paper (Fazel, Hindi, & Boyd, 2001). It is well known that the minimum realization

order (also known as McMillan degree) of a discrete time LTI system coincides with the rank of the Hankel operator constructed from the impulse response coefficients:

$$H(g) = \begin{pmatrix} g(1) & g(2) & g(3) & \cdots \\ g(2) & g(3) & g(4) & \cdots \\ g(3) & g(4) & g(5) & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

Such observation can be readily exploited by adopting a high order FIR model and solving (87) with $J(g) = \|H(g)\|_*$ which will fit the data while encouraging a low McMillan degree, see e.g. Grossmann, Jones, and Morari (2009) and Mohan and Fazel (2010). The idea has been extended to the estimation of VAR (Signoretto & Suykens, 2012) and Box–Jenkins models (Hjalmarsson, Welsh, & Rojas, 2012), applying a nuclear norm penalty to high-order ARX models. An atomic norm regularizer (Chandrasekaran, Recht, Parrilo, & Willsky, 2012) which mimics the Hankel nuclear norm, and is approximated by a suitable ℓ_1 penalty, is described in Shah, Bhaskar, Tang, and Recht (2012). A comparison between quadratic and nuclear norm regularization for linear system identification can be found in Chiuso, Chen, Ljung, and Pillonetto (2013). Nuclear norm regularization has been also adopted for subspace identification and estimation of Hammerstein systems (Falck, Suykens, Schoukens, & De Moor, 2010; Fazel, Kei, Sun, & Tseng, 2013; Liu & Vandenberghe, 2009).

18. Conclusions

In this survey we have taken a broad approach to one basic problem in system identification, namely to estimate the impulse response of a linear system. The state-of-the art techniques for this were reviewed in Part I, together with classical statistical regularization techniques equipped with recent tuning tools. In Parts II and III, the problem was also set in a general function estimation framework employing Reproducing Kernel Hilbert Space theory, displaying the connections with machine learning and the links between reproducing kernels and Bayesian estimation. We have shown how this abstract and broad perspective is quite useful in guiding the selection of useful kernels and tuning rules.

These broad encounters have clearly been beneficial and fruitful for classical system identification. The numerical examples, especially for standard system identification models in Section 7, show that conventional techniques may be outperformed by carefully tuned regularized estimates. One way to explain this is that the difficult choice of model complexity (model order) can be circumvented by careful regularization. In a way, the bias-variance trade-off that is behind any reasonable choice of model complexity, gets a whole new dimension and richness in the choice of (continuous) regularization parameters compared to the choice of (discrete) model order.

Acknowledgments

We thank Johan Schoukens and Jan Swevers for letting us use the robot arm data in Section 8.

Appendix. Proof of Proposition 19

Below, $z_0(t)$ denotes the true noiseless system output with $\hat{z}$ its estimate obtained convolving the solution of (61) with the system input. In addition, $\bar{z}_0$ and $\hat{\bar{z}}$ indicate any finite-dimensional vector obtained by sampling, respectively, z_0 and $\hat{z}$ on the same temporal instants, including the t_i where the measurements are available. We start noticing that the desired kernel must also lead to the optimal estimator of every linear transformation of g_0. Hence, it has also to minimize (in matrix sense) $MSE(K) = \mathscr{E}\left[(\hat{\bar{z}} - \bar{z}_0)(\hat{\bar{z}} - \bar{z}_0)^T\right]$. Notice that the measurement vector Y is related to $\bar{z}_0$ by $Y = \Phi\bar{z}_0 + E$ where each row of Φ suitably selects the right component of $\bar{z}_0$. Hence, it follows from Theorem 1 in Chen et al. (2012) that the optimal kernel must satisfy[7] $L_t L_\tau[K(\cdot,\cdot)] = z(t)z(\tau)$ where, given a function f, $L_t[f]$ returns the convolution between f and the system input evaluated at t. In turn, one has

$$\hat{K}(x_1, x_2) = (g_0(x_1) + \eta(x_1))(g_0(x_2) + \eta(x_2)) \tag{88}$$

for a certain η belonging to the null space of L_t, $\forall t$.

In what follows, let now $\bar{z}_0$ be the restriction of $z_0(t)$ just on the time instants $\{t_i\}_{i=1}^N$ where the measurements are collected. In this way, the output kernel matrix in (62) with $K = \hat{K}$ is $\mathbf{O} = \bar{z}_0\bar{z}_0^T$ and it comes from the representer theorem that the optimal estimator of $\bar{z}_0$ is $\hat{\bar{z}} = \mathbf{O}(\mathbf{O} + \gamma I_N)^{-1} Y$. In addition, in view of (88) and the above expression, the estimator of $g_0(x)$ minimizing the MSE can be written as

$$\hat{g}(x) = (g_0(x) + \eta(x))\bar{z}_0^T\mathbf{O}^\dagger\hat{\bar{z}} \tag{89}$$

where $\mathbf{O}^\dagger$ is the pseudoinverse of $\mathbf{O}$ and we used the equalities $\bar{z}_0^T = \bar{z}_0^T\mathbf{O}^\dagger\mathbf{O}$ and $L_t\hat{K}(x,\cdot) = (g(x) + \eta(x))z(t)$.

Thus, (89) shows that $\hat{g}(x)$ is a linear transformation of $\hat{\bar{z}}$ which however still depends on the unknown η. We now prove that the MSE is minimized setting $\eta(x) = 0$, $\forall x$. In fact, let $A = g_0(x)\bar{z}_0^T\mathbf{O}^\dagger$. The optimal estimator for $A\bar{z}_0$ must be $A\hat{\bar{z}}$ and one has

$$A\bar{z}_0 = g_0(x)\bar{z}_0^T\mathbf{O}^\dagger\bar{z}_0 = g_0(x), \qquad A\hat{\bar{z}} = g_0(x)\bar{z}_0^T\mathbf{O}^\dagger\hat{\bar{z}}.$$

Extension of the above formulas to the multivariate case is straightforward and this completes the proof.

References

Adams, R., & Fournier, J. (2003). *Sobolev spaces*. Academic Press.
Ahlberg, J., & Nilson, E. (1963). Convergence properties of the spline fit. *Journal of Society for Industrial and Applied Mathematics*, *11*, 95–104.
Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723.
Allen, D. M. (1974). The relationship between variable selection and data augmentation and a method for prediction. *Technometrics*, *16*, 125–127.
Andrieu, C., Doucet, A., & Holenstein, R. (2010). Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *72*(3), 269–342.
Aravkin, A., Burke, J., Chiuso, A., & Pillonetto, G. (2012). On the MSE properties of empirical Bayes methods for sparse estimation. In *Proceedings of the 16th IFAC symposium on system identification*, SysId 2012.
Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society*, *68*, 337–404.
Atkinson, K. (1968). On the order of convergence of natural cubic spline interpolation. *SIAM Journal on Numerical Analysis*, *5*(1), 89–101.
Bach, F., & Jordan, M. (2005). Predictive low-rank decomposition for kernel methods. In *Proceedings of the 22nd international conference on machine learning*, ICML'05. (pp. 33–40). New York, NY, USA: ACM.
Bennett, K., & Parrado-Hernandez, E. (2006). The interplay of optimization and machine learning research. *Journal of Machine Learning Research*, *7*, 1265–1281.
Berger, J. (1980). A robust generalized Bayes estimator and confidence region for a multivariate normal mean. *The Annals of Statistics*, *8*, 716–761.
Berger, J. (1982). Selecting a minimax estimator of a multivariate normal mean. *The Annals of Statistics*, *10*, 81–92.
Berger, J. (1985). *Springer series in statistics*, *Statistical decision theory and Bayesian analysis* (2nd ed.). Springer.
Berger, J. (1994). An overview of robust Bayesian analysis. *Test*, *3*, 5–124.
Bertero, M. (1989). Linear inverse and ill-posed problems. *Advances in Electronics and Electron Physics*, *75*, 1–120.
Bhattacharya, P. K. (1966). Estimating the mean of a multivariate normal population with general quadratic loss function. *The Annals of Mathematical Statistics*, *37*, 1819–1824.

[7] To simplify the exposition, we discard situations where the optimal kernel derived in Chen et al. (2012) is not unique.

Bock, M. (1975). Minimax estimators of the mean of a multivariate normal distribution. *The Annals of Statistics*, *3*, 209–218.

Bottou, L., Chapelle, O., DeCoste, D., & Weston, J. (Eds.) (2007). *Large scale kernel machines*. Cambridge, MA, USA: MIT Press.

Calamante, F., Thomas, D., Pell, G., Wiersma, J., & Turner, R. (1999). Measuring cerebral blood flow using magnetic resonance imaging techniques. *Journal of Cerebral Blood Flow and Metabolism*, *19*, 701–735.

Carli, F., Chen, T., Chiuso, A., Ljung, L., & Pillonetto, G. (2012). On the estimation of hyperparameters for Bayesian system identification with exponential kernels. In *Proceedings of the IEEE conference on decision and control*, CDC 2012.

Carli, F., Chiuso, A., & Pillonetto, G. (2012). Efficient algorithms for large scale linear system identification using stable spline estimators. In *Proceedings of the 16th IFAC symposium on system identification*, SysId 2012.

Carmeli, C., Vito, E. D., & Toigo, A. (2006). Vector valued reproducing kernel Hilbert spaces of integrable functions and Mercer theorem. *Analysis and Applications*, *4*, 377–408.

Casella, G. (1980). Minimax ridge regression estimation. *The Annals of Statistics*, *8*, 1036–1056.

Chandrasekaran, V., Recht, B., Parrilo, P., & Willsky, A. (2012). The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, *12*(6), 805–849.

Chen, T., Andersen, M. S., Ljung, L., Chiuso, A., & Pillonetto, G. (2014). System identification via sparse multiple kernel-based regularization using sequential convex optimization techniques. *IEEE Transactions on Automatic Control*.

Chen, T., & Ljung, L. (2013). Implementation of algorithms for tuning parameters in regularized least squares problems in system identification. *Automatica*, *49*(7), 2213–2220.

Chen, T., Ohlsson, H., Goodwin, G., & Ljung, L. (2011). Kernel selection in linear system identification—part II: a classical perspective. In *Proceedings of CDC-ECC*.

Chen, T., Ohlsson, H., & Ljung, L. (2012). On the estimation of transfer functions, regularizations and Gaussian processes—revisited. *Automatica*, *48*(8), 1525–1535.

Chiuso, A., Chen, T., Ljung, L., & Pillonetto, G. (2013). Regularization strategies for nonparametric system identification. In *Proceedings of the 52nd annual conference on decision and control*, CDC.

Cox, R. (1946). Probability, frequency, and reasonable expectation. *American Journal of Physics*, *14*(1), 1–13.

Craven, P., & Wahba, G. (1979). Smoothing noisy data with spline functions. *Numerische Mathematik*, *31*, 377–403.

Cucker, F., & Smale, S. (2001). On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, *39*, 1–49.

De Nicolao, G., Sparacino, G., & Cobelli, C. (1997). Nonparametric input estimation in physiological systems: problems, methods and case studies. *Automatica*, *33*, 851–870.

De Nicolao, G., & Trecate, G. (1999). Consistent identification of NARX models via regularization networks. *IEEE Transactions on Automatic Control*, *44*(11), 2045–2049.

De Vito, E., Rosasco, L., Caponnetto, A., De Giovannini, U., & Odone, F. (2005). Learning from examples as an inverse problem. *Journal of Machine Learning Research*, *6*, 883–904.

Dinuzzo, F. (2011). Analysis of fixed-point and coordinate descent algorithms for regularized kernel methods. *IEEE Transactions on Neural Networks*, *22*(10), 1576–1587.

Dinuzzo, F., & Schölkopf, B. (2012). The representer theorem for Hilbert spaces: a necessary and sufficient condition. In P. Bartlett, F. Pereira, C. Burges, L. Bottou, & K. Weinberger (Eds.), *Advances in neural information processing systems, vol. 25* (pp. 189–196).

Efron, B., & Morris, C. (1973). Stein's estimation rule and its competitors—an empirical Bayes approach. *Journal of the American Statistical Association*, *68*(341), 117–130.

Ekstrom, M. (1973). A spectral characterization of the ill-conditioning in numerical deconvolution. *IEEE Transactions on Audio and Electroacoustics*, *21*(4), 344–348.

Espinoza, M., Suykens, J. A. K., & De Moor, B. (2005). Kernel based partially linear models and nonlinear identification. *IEEE Transactions on Automatic Control*, *50*(10), 1602–1606.

Evgeniou, T., Pontil, M., & Poggio, T. (2000). Regularization networks and support vector machines. *Advances in Computational Mathematics*, *13*, 1–150.

Falck, T., Dreesen, P., Brabanter, K. D., Pelckmans, K., De Moor, B., & Suykens, J. (2012). Least-squares support vector machines for the identification of Wiener–Hammerstein systems. *Control Engineering Practice*, *20*, 1165–1174.

Falck, T., Pelckmans, K., Suykens, J., & De Moor, B. (2009). Identification of Wiener–Hammerstein systems using LS-SVMs. In *Proceedings of the 15th IFAC symposium on system identification*, SYSID 2009, pp. 820–825.

Falck, T., Suykens, J., Schoukens, J., & De Moor, B. (2010). Nuclear norm regularization for overparametrized Hammerstein systems. In *Proceedings of the 49th annual conference on decision and control*, CDC, pp. 7202–7207.

Fazel, M., Hindi, H., & Boyd, S. (2001). A rank minimization heuristic with application to minimum order system approximation. In *Proceedings of the 2001 American control conference, 2001, vol. 6* (pp. 4734–4739).

Fazel, M., Kei, P. T., Sun, D., & Tseng, P. (2013). Hankel matrix rank minimization with applications to system identification and realization. *SIAM Journal on Matrix Analysis and Applications*, *34*(3), 946–977.

Ferrari-Trecate, G., Williams, C., & Opper, M. (1999). Finite-dimensional approximation of Gaussian processes. In *Proceedings of the 1998 conference on advances in neural information processing systems* (pp. 218–224). Cambridge, MA, USA: MIT Press.

Franz, M., & Schölkopf, B. (2006). A unifying view of Wiener and Volterra theory and polynomial kernel regression. *Neural Computation*, *18*, 3097–3118.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1–22.

Frigola, R., Lindsten, F., Schon, T., & Rasmussen, C. (2013). Bayesian inference and learning in Gaussian process state-space models with particle MCMC. *Advances in Neural Information Processing Systems (NIPS)*.

Frigola, R., & Rasmussen, C. (2013). Integrated pre-processing for Bayesian nonlinear system identification with Gaussian processes. In *Proceedings of the 52nd annual conference on decision and control*, CDC.

Gilks, W., Richardson, S., & Spiegelhalter, D. (1996). *Markov chain Monte Carlo in practice*. London: Chapman and Hall.

Girard, A., Rasmussen, C.E., Quinonero-Candela, J., & Murray-Smith, R. (2003). Bayesian regression and Gaussian process priors with uncertain inputs—application to multiple-step ahead time series forecasting. In *Proc. of NIPS*.

Girosi, F., Jones, M., & Poggio, T. (1995). Regularization theory and neural networks architectures. *Neural Computation*, *7*(2), 219–269.

Goethals, I., Pelckmans, K., Falck, T., Suykens, J., & De Moor, B. (2010). NARX identification of Hammerstein systems using least-squares support vector machines. In *Lecture notes in control and information sciences: vol. 404. Block-oriented nonlinear system identification* (pp. 241–258). London: Springer.

Goethals, I., Pelckmans, K., Suykens, J., & De Moor, B. (2005). Identification of MIMO Hammerstein models using least squares support vector machines. *Automatica*, *41*(7), 1263–1272.

Golub, G., Heath, M., & Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, *21*(2), 215–223.

Goodwin, G., Gevers, M., & Ninness, B. (1992). Quantifying the error in estimated transfer functions with application to model order selection. *IEEE Transactions on Automatic Control*, *37*(7), 913–928.

Grossmann, C., Jones, C., & Morari, M. (2009). System identification via nuclear norm regularization for simulated moving bed processes from incomplete data sets. In *Proceedings of the 48th IEEE conference on decision and control*, CDC, pp. 4692–4697.

Hadamard, J. (1922). *Lectures on Cauchy's problem*. New Haven, CT: Yale University Press.

Hall, J., Rasmussen, C., & Maciejowski, J. (2012). Modelling and control of nonlinear systems using Gaussian processes with partial model information. In *Proceedings of the 51st annual conference on decision and control*, CDC.

Hanke, M. (1995). The minimal error conjugate gradient method is a regularization method. *Proceedings of the Americal Mathematical Society*, *123*, 3487–3497.

Hansen, P. (1987). The truncated SVD as a method for regularization. *BIT*, *27*, 534–553.

Hengland, M. (2007). Approximate maximum a posteriori with Gaussian process priors. *Constructive Approximation*, *26*, 205–224.

Hjalmarsson, H., Welsh, J., & Rojas, C. (2012). Identification of Box–Jenkins models using structured ARX models and nuclear norm relaxation. In *16th IFAC symposium on system identification*, IFAC, pp. 322–327.

Hofmann, T., Schölkopf, B., & Smola, A. J. (2008). Kernel methods in machine learning. *Annals of Statistics*, *36*(3), 1171–1220.

Horst, R., & Thoai, N. V. (1999). DC programming: overview. *Journal of Optimization Theory and Applications*, *103*(1), 1–43.

Huber, P. J. (1981). *Robust statistics*. New York, NY, USA: John Wiley and Sons.

Hunt, B. (1970). The inverse problem of radiography. *Mathematical Biosciences*, *8*, 161–179.

Hurvich, C., & Tsai, C. (1989). Regression and time series model selection in small samples. *Biometrika*, *76*, 297–307.

James, W., & Stein, C. (1961). Estimation with quadratic loss. In *Proceedings of the 4th Berkeley symposium on mathematical statistics and probability, vol. I* (pp. 361–379). University of California Press.

Kimeldorf, G., & Wahba, G. (1971a). A correspondence between Bayesan estimation of stochastic processes and smoothing by splines. *Annals of Mathematics Statistics*, *41*(2), 495–502.

Kimeldorf, G., & Wahba, G. (1971b). Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, *33*(1), 82–95.

Kulis, B., Sustik, M., & Dhillon, I. (2006). Learning low-rank kernel matrices. In *Proceedings of the 23rd international conference on machine learning*, ICML'06. (pp. 505–512). New York, NY, USA: ACM.

Lázaro-Gredilla, M., Quiñonero-Candela, J., Rasmussen, C., & Figueiras-Vidal, A. (2010). Sparse spectrum Gaussian process regression. *Journal of Machine Learning Research*, *99*, 1865–1881.

Leithead, W.E., Solak, E., & Leith, D.J. (2003). Direct identification of nonlinear structure using Gaussian process prior models. In *Proceedings of European control conference*, ECC 2003.

Li, Y., Li, L., Su, H., & Chun, J. (2006). Least squares support vector machine based partially linear model identification. In *Lecture notes in computer science: vol. 4113. Intelligent computing* (pp. 775–781). Berlin Heidelberg: Springer.

Lindsten, F., Schön, T., & Jordan, M. (2013). Bayesian semiparametric Wiener system identification. *Automatica*, *49*, 2053–2063.

List, N., & Simon, H.U. (2004). A general convergence theorem for the decomposition method. In *Proceedings of the 17th annual conference on computational learning theory*, pp. 363–377.

List, N., & Simon, H. (2007). General polynomial time decomposition algorithms. *Journal of Machine Learning Research*, *8*, 303–321.
Liu, Z., & Vandenberghe, L. (2009). Interior-point method for nuclear norm approximation with application to system identification. *SIAM Journal on Matrix Analysis and Applications*, *31*(3), 1235–1256.
Ljung, L. (1999). *System identification—theory for the user* (2nd ed.). Upper Saddle River, N.J: Prentice-Hall.
Ljung, L. (2002). Prediction error estimation methods. *Circuits, Systems, and Signal Processing*, *21*(1), 11–21.
Ljung, L. (2013). *System identification toolbox V8.3 for Matlab*. 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.
MacKay, D. (1992). Bayesian interpolation. *Neural Computation*, *4*, 415–447.
Magni, P., Bellazzi, R., & De Nicolao, G. (1998). Bayesian function learning using MCMC methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(12), 1319–1331.
Mallows, C. (1973). Some comments on C_p. *Technometrics*, *15*, 661–675.
Maritz, J. S., & Lwin, T. (1989). *Empirical Bayes method*. Chapman and Hall.
Maruyama, Y., & Strawderman, W. (2005). A new class of generalized Bayes minimax ridge regression estimators. *The Annals of Statistics*, 1753–1770.
Minh, H. (2006). Reproducing Kernel Hilbert spaces in learning theory. Ph.D. thesis, Brown University.
Mohan, K., & Fazel, M. (2010). Reweighted nuclear norm minimization with application to system identification. In *American control conference*, ACC, pp. 2953–2959.
Nemirovskii, A. (1986). The regularization properties of the adjoint gradient method in ill-posed problems. *USSR Computational Mathematics and Mathematical Physics*, *26*(2), 7–16.
Ninness, B., & Henriksen, S. (2010). Bayesian system identification via MCMC techniques. *Automatica*, *46*(1), 40–51.
Olofsson, E. (2013). Extended benchmark results for impulse response estimation. In *Proceedings of the 52nd IEEE conference on decision and control*, CDC 2013.
Papoulis, A. (1984). *Probability, random variables and stochastic processes*. Mc Graw-Hill.
Phillips, D. (1962). A technique for the numerical solution of certain integral equations of the first kind. *Journal of the Association for Computing Machinery*, *9*, 84–97.
Pillonetto, G. (2013). Consistent identification of Wiener systems: a machine learning viewpoint. *Automatica*, *49*(9), 2704–2712.
Pillonetto, G., & Bell, B. (2007). Bayes and empirical Bayes semi-blind deconvolution using eigenfunctions of a prior covariance. *Automatica*, *43*(10), 1698–1712.
Pillonetto, G., Chiuso, A., & De Nicolao, G. (2011). Prediction error identification of linear systems: a nonparametric Gaussian regression approach. *Automatica*, *47*(2), 291–305.
Pillonetto, G., Chiuso, A., & De Nicolao, G. (2010). Regularized estimation of sums of exponentials in spaces generated by stable spline kernels. In *Proceedings of the IEEE American cont. conf.*, Baltimora, USA.
Pillonetto, G., Chiuso, A., & Quang, M. H. (2011). A new kernel-based approach for nonlinear system identification. *IEEE Transactions on Automatic Control*, *56*(12), 2825–2840.
Pillonetto, G., & De Nicolao, G. (2010). A new kernel-based approach for linear system identification. *Automatica*, *46*(1), 81–93.
Pillonetto, G., & De Nicolao, G. (2011). Kernel selection in linear system identification—part I: a Gaussian process perspective. In *Proceedings of CDC-ECC*.
Pillonetto, G., & De Nicolao, G. (2012). Pitfalls of the parametric approaches exploiting cross-validation or model order selection. In *Proceedings of the 16th IFAC symposium on system identification*, SysId 2012.
Pintelon, R., & Schoukens, J. (2012a). *System identification—a frequency domain approach* (2nd ed.). New York: IEEE Press.
Pintelon, R., & Schoukens, J. (2012b). *System identification: a frequency domain approach* (2nd ed.). John Wiley & Sons.
Poggio, T., & Girosi, F. (1990). Networks for approximation and learning. *Proceedings of the IEEE*, *78*, 1481–1497.
Quiñonero-Candela, J., & Rasmussen, C. (2005). A unifying view of sparse approximate Gaussian process regression. *Journal of Machine Learning Research*, *6*, 1939–1959.
Rasmussen, C., & Williams, C. (2006). *Gaussian processes for machine learning*. The MIT Press.
Rissanen, J. (1978). Modelling by shortest data description. *Automatica*, *14*, 465–471.
Rockafellar, R. (1970). *Princeton landmarks in mathematics, Convex analysis*. Princeton University Press.
Rojas, C., & Hjalmarsson, H. (2011). Sparse estimation based on a validation criterion. In *Proceedings of the 50th IEEE conference on decision and control and European control conference*, CDC-ECC11.
Rojas, C., Wahlberg, B., & Hjalmarsson, H. (2013). A sparse estimation technique for general model structures. In *Proceedings of the European control conference*, ECC13.
Rudin, W. (1987). *Real and complex analysis*. Singapore: McGraw-Hill.
Runge, C. (1901). Uber empirische funktionen und die interpolation zwischen aquidistanten ordinaten. *Zeitschrift für Mathematik und Physik*, *46*, 224–243.
Saitoh, S. (1988). *Pitman research notes in mathematics series: vol. 189. Theory of reproducing kernels and its applications*. Harlow: Longman Scientific and Technical.
Schölkopf, B., Herbrich, R., & Smola, A. J. (2001). A generalized representer theorem. *Neural Networks and Computational Learning Theory*, *81*, 416–426.
Schölkopf, B., & Smola, A. J. (2001). *Learning with kernels: support vector machines, regularization, optimization, and beyond, (Adaptive computation and machine learning)*. MIT Press.
Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.
Söderström, T., & Stoica, P. (1989). *System identification*. Prentice-Hall.
Shah, P., Bhaskar, B., Tang, G., & Recht, B. (2012). Linear system identification via atomic norm regularization. In *Proceedings of the 51st annual conference on decision and control*, CDC, pp. 6265–6270.
Shawe-Taylor, J., & Cristianini, N. (2004). *Kernel methods for pattern analysis*. Cambridge University Press.
Shinozaki, N. (1974). A note on estimating the mean vector of a multivariate normal distribution with general quadratic loss function. *Keio Engineering Reports*, *27*, 105–112.
Signoretto, M., & Suykens, J. (2012). Convex estimation of cointegrated VAR models by a nuclear norm penalty. In *Proceedings of the 16th IFAC symposium on system identification*, Sysid.
Smale, S., & Zhou, D. (2007). Learning theory estimates via integral operators and their approximations. *Constructive Approximation*, *26*, 153–172.
Smola, A., & Schölkopf, B. (2000). Sparse greedy matrix approximation for machine learning. In *Proceedings of the seventeenth international conference on machine learning*, ICML '00. (pp. 911–918). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.
Snelson, E., & Ghahramani, Z. (2006). Sparse Gaussian processes using pseudo-inputs. In *Advances in neural information processing systems, vol. 18* (pp. 1257–1264). MIT Press.
Stein, C. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 1135–1151.
Steinwart, I. (2002). On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, *2*, 67–93.
Steinwart, I., Hush, D., & Scovel, C. (2006). An explicit description of the reproducing Kernel Hilbert space of Gaussian RBF kernels. *IEEE Transactions on Information Theory*, *52*, 4635–4643.
Strawderman, W. (1978). Minimax adaptive generalized ridge regression estimators. *Journal of the American Statistical Association*, *73*, 623–627.
Suykens, J., Gestel, T. V., Brabanter, J. D., De Moor, B., & Vandewalle, J. (2002). *Least squares support vector machines*. Singapore: World Scientific.
Tao, P. D., & An, L. T. H. (1997). Convex analysis approach to D. C. programming: theory, algorithms and applications. *ACTA Mathematica Vietnamica*, *22*, 289–355.
Tarantola, A. (2005). *Inverse problem theory and methods for model parameter estimation*. Philadelphia: SIAM.
Tibshirani, R. (1994). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, *58*, 267–288.
Tikhonov, A., & Arsenin, V. (1977). *Solutions of ill-posed problems*. Washington, D.C.: Winston/Wiley.
Torfs, D. E., Vuerinckx, R., Swevers, J., & Schoukens, J. (1998). Comparison of two feedforward design methods aiming at accurate trajectory tracking of the end point of a flexible robot arm. *IEEE Transactions on Control Systems Technology*, *6*(1), 2–14.
Toth, R., Hjalmarsson, H., & Rojas, C. (2012). Sparse estimation of rational dynamic models. In *Proceedings of the 16th IFAC symposium on system identification*, SYSID 2012.
Twomey, S. (1977). *Introduction to the mathematics of inversion in remote sensing and indirect measurements*. New York: Elsevier.
Vapnik, V. (1998). *Statistical learning theory*. New York, NY, USA: Wiley.
Wahba, G. (1990). *Spline models for observational data*. Philadelphia: SIAM.
Wang, L., & Cluett, W. (1996). Use of PRESS residuals in dynamic system identification. *Automatica*, *32*, 781–784.
Welsh, J., Rojas, C., Hjalmarsson, H., & Wahlberg, B. (2012). Sparse estimation for basis function selection in wideband system identification. In *Proceedings of the 16th IFAC symposium on system identification*, SYSID 2012.
Wiener, N. (1949). *Extrapolation, interpolation, and smoothing of stationary time series*. Cambridge, MA: MIT Press.
Williams, C. I., & Seeger, M. (2000). Using the Nyström method to speed up kernel machines. In *Proceedings of the 2000 conference on advances in neural information processing systems* (pp. 682–688). Cambridge, MA, USA: MIT Press.
Wu, Q., Ying, Y., & Zhou, D. (2006). Learning rates of least-square regularized regression. *Foundations of Computational Mathematics*, *6*, 171–192.
Xu, Y., & Chen, D. (2009). Partially-linear least-squares regularized regression for system identification. *IEEE Transactions on Automatic Control*, *54*(11), 2637–2641.
Yao, Y., Rosasco, L., & Caponnetto, A. (2007). On early stopping in gradient descent learning. *Constructive Approximation*, *26*(2), 289–315.
Yuan, M., & Tony Cai, T. (2010). A reproducing kernel Hilbert space approach to functional linear regression. *Annals of Statistics*, *38*, 3412–3444.
Zanderigo, F., Bertoldo, A., Pillonetto, G., & Cobelli, C. (2009). Nonlinear stochastic regularization to characterize tissue residue function in bolus-tracking mri: assessment and comparison with svd, block-circulant svd, and tikhonov. *IEEE Transactions on Biomedical Engineering*, *56*(5), 1287–1297.

Zhang, K., & Kwok, J. (2010). Clustered Nyström method for large scale manifold learning and dimension reduction. *IEEE Transactions on Neural Networks*, *21*(10), 1576–1587.
Zhu, Y. (1989). Black-box identification of MIMO transfer functions: asymptotic properties of prediction error models. *International Journal of Adaptive Control and Signal Processing*, *3*(4), 357–373.
Zhu, Y., & Backx, T. (1993). *Identification of multivariable industrial processes for diagnosis and control*. London: Springer Verlag.
Zhu, H., & Rohwer, R. (1996). Bayesian regression filters and the issue of priors. *Neural Computing and Applications*, *4*, 130–142.
Zhu, H., Williams, C., Rohwer, R., & Morciniec, M. (1998). Gaussian regression and optimal finite dimensional linear models. In *Neural networks and machine learning*. Berlin: Springer-Verlag.
Zierler, K. (1962). Theoretical basis of indicator-dilution methods for measuring flow and volume. *Circulation Research*, *10*, 393–407.
Zierler, K. (1965). Equations for measuring blood flow by external monitoring of radioisotopes. *Circulation Research*, *16*, 309–321.

Gianluigi Pillonetto was born on January 21, 1975 in Montebelluna (TV), Italy. He received the Doctoral degree in Computer Science Engineering cum laude from the University of Padova in 1998 and the Ph.D. degree in Bioengineering from the Polytechnic of Milan in 2002. In 2000 and 2002 he was visiting scholar and visiting scientist, respectively, at the Applied Physics Laboratory, University of Washington, Seattle. From 2002 to 2005 he was Research Associate at the Department of Information Engineering, University of Padova. Since 2005 he has been Assistant Professor of Control and Dynamic Systems at the Department of Information Engineering, University of Padova. His research interests are in the field of system identification, estimation and machine learning.

Francesco Dinuzzo is a Research Scientist at IBM Research, Dublin (Ireland) since December 2013. From April 2011 until December 2013 he was a Research Scientist at the Max Planck Institute for Intelligent Systems, Tübingen (Germany). He received a Master degree in Computer Science Engineering in 2007 and the Ph.D. degree in Mathematics and Statistics in 2011, both from the University of Pavia (Italy). He also received a Master degree in Science and Technology in 2007 from Istituto Universitario di Studi Superiori (Italy). He has held visiting positions at Massachusetts Institute of Technology (Center for Biological and Computational Learning), National Taiwan University (Machine Learning Group), ETH Zurich (Machine Learning Laboratory), and Institute of Statistical Mathematics, Tokyo. His research interests include machine learning, optimization, non-linear control, system identification, regularization methods, and distributed computation.

Tianshi Chen was born in China in November 1978. He received his M.E. degree from Harbin Institute of Technology in 2005, and Ph.D. from Chinese University of Hong Kong in December 2008. From April 2009 to March 2011, he was a postdoctoral researcher at the Automatic Control group of Linköping University, 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.

Giuseppe De Nicolao was born in Padova, Italy, in 1962. In 1986 he received the Laurea (master degree) cum laude in Electronic Engineering (“Laurea”) from the Polytechnic of Milan, Italy. From 1987 to 1988 he was with the Biomathematics and Biostatistics Unit of the Institute of Pharmacological Researches “Mario Negri”, Milano. In 1988 he joined the Italian National Research Council (C.N.R.) as a researcher scientist of the Center of System Theory in Milano. From 1992 to 2000 he was Associate Professor and, since 2000, Full Professor of Model Identification and Data Analysis in the Department of Computer Science and Systems Engineering of the University of Pavia (Italy). In 1991, he has been a visiting fellow at the Department of Systems Engineering of the Australian National University, Canberra. From 1999 to 2001, he has been Associate Editor of the IEEE Transactions on Automatic Control and, since 2007, Associate Editor of Automatica. His research interests include Bayesian learning, neural networks, model predictive control, optimal and robust filtering and control, deconvolution techniques, modeling, identification and control of biomedical systems, statistical process control and fault diagnosis for semiconductor manufacturing.

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 Linköping, Sweden. 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. He is 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, an Honorary Professor of the Chinese Academy of Mathematics and Systems Science, 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, and in 2003 he received the Hendrik W. Bode Lecture Prize from the IEEE Control Systems Society, and he was the 2007 recipient of the IEEE Control Systems Award.