Automatica 47 (2011) 291–305

Contents lists available at ScienceDirect

Automatica

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

Prediction error identification of linear systems: A nonparametric Gaussian regression approach☆

Gianluigi Pillonetto [a,*], Alessandro Chiuso [b], Giuseppe De Nicolao [c]

[a] Department of Information Engineering, University of Padova, Padova, Italy

[b] Dipartimento di Tecnica e Gestione dei Sistemi Industriali, University of Padova, Vicenza, Italy

[c] Dipartimento di Informatica e Sistemistica, University of Pavia, Pavia, Italy

ARTICLE INFO

Article history:
Available online 28 December 2010

Keywords:
Linear system identification
Predictor estimation
Kernel-based methods
Bayesian estimation
Regularization
Gaussian processes
Subspace methods

ABSTRACT

A novel Bayesian paradigm for the identification of output error models has recently been proposed in which, in place of postulating finite-dimensional models of the system transfer function, the system impulse response is searched for within an infinite-dimensional space. In this paper, such a nonparametric approach is applied to the design of optimal predictors and discrete-time models based on prediction error minimization by interpreting the predictor impulse responses as realizations of Gaussian processes. The proposed scheme describes the predictor impulse responses as the convolution of an infinite-dimensional response with a low-dimensional parametric response that captures possible high-frequency dynamics. Overparameterization is avoided because the model involves only a few hyperparameters that are tuned via marginal likelihood maximization. Numerical experiments, with data generated by ARMAX and infinite-dimensional models, show the definite advantages of the new approach over standard parametric prediction error techniques and subspace methods both in terms of predictive capability on new data and accuracy in reconstruction of system impulse responses.

1. Introduction

The most widespread approach to optimal prediction of discrete-time systems relies on prediction error methods (PEMs), for which a large corpus of theoretical results is available. For a detailed treatment, the interested reader may refer to Ljung (1999) and Soderstrom and Stoica (1989), and the references therein. The statistical properties of prediction error (and maximum likelihood) methods are well understood under the assumption that the model class is fixed, and they show that these kinds of procedure are in some sense optimal, at least for large samples. Within this parametric paradigm, a key point is the selection of the most adequate model structure, which is usually carried out by resorting to complexity measures such as AIC and BIC (Akaike, 1974; Schwarz, 1978). In particular, AIC-type criteria are minimax-rate optimal for regression function estimation, while BIC-type ones are consistent (Hannan, 1980; Hurvich & Tsai, 1989); notably, for standard order estimation criteria, these properties are mutually exclusive (Yang, 2005). Theoretical properties of estimators obtained after model selection, also called *post model selection estimators* (PMSEs), are generally hard to study, as discussed for instance in Leeb and Pötscher (2005) and the references therein. Not surprisingly, sample properties of PMSEs, such as impulse response estimators or predictors, when tested on experimental data (see e.g. Section 6), may depart sharply from those predicted by "standard" (i.e. without model selection) statistical theory, which suggests that PEMs should be asymptotically efficient for Gaussian innovations.

In this paper, we follow an alternative route to prediction and identification of linear systems by adopting a Bayesian point of view. Bayesian approaches to identification are by no means new: there is an extensive literature whose origins can be traced back to the 1980s; see e.g. Doan, Litterman, and Sims (1984) and Kitagawa and Gersh (1984, 1985), and also the more recent book (Kitagawa & Gersch, 1996). Nonparametric approaches have also been used in the identification of nonlinear system models, for example in Young, McKenna, and Bruun (2001), where state-dependent parameters in nonlinear transfer function models are

☆ This research has been partially supported by the PRIN Projects "New Methods and Algorithms for Identification and Adaptive Control of Technological Systems" and "Artificial Pancreas: Physiological Models, Control Algorithms and Clinical Test", by the Progetto di Ateneo CPDA090135/09 funded by the University of Padova and by the European Community's Seventh Framework Programme under agreement no. FP7-ICT-223866-FeedNetBack. The material in this paper was not presented at any conference. This paper was recommended for publication in revised form by Associate Editor Tongwen Chen under the direction of Editor Ian R. Petersen.

* Corresponding author. Tel.: +39 0498277607; fax: +39 049 8277699.
E-mail addresses: giapi@dei.unipd.it (G. Pillonetto), chiuso@dei.unipd.it (A. Chiuso), giuseppe.denicolao@unipv.it (G. De Nicolao).

0005-1098/$ – see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.automatica.2010.11.004

estimated using a special nonparametric approach based on recursive fixed interval smoothing (Young & Pedregal, 1999). Our approach is also based on nonparametric estimation, but it is applied to the identification of linear systems described by impulse responses (Pillonetto & De Nicolao, 2010). Rather than postulating finite-dimensional structures for the system transfer function, e.g. ARX, ARMAX or Laguerre (Goodwin, Braslavsky, & Seron, 2002; Milanese & Vicino, 1991), the system impulse response is searched for within an infinite-dimensional space. In order to circumvent the intrinsical ill-posed nature of the problem, regularization methods, admitting a Bayesian interpretation, are employed (Barry, 1986; Bertero, 1989; Pillonetto & Bell, 2007; Tikhonov & Arsenin, 1977). This is similar in spirit to the work in Kitagawa and Gersh (1985), McVinish, Braslavsky, and Mengersen (2006), and Young et al. (2001). The real difference is made by the recent introduction of a prior distribution on the impulse response such that the realizations are almost surely (a.s.) BIBO stable (Pillonetto & De Nicolao, 2010). This method has been shown to compare very favorably with established parametric approaches in the identification of output error models (Pillonetto & De Nicolao, 2010). Along this line, it is of interest to extend this nonparametric paradigm to the design of optimal predictors and the identification of discrete-time models based on prediction error minimization (hereafter, without loss of generality, the analysis will be restricted to SISO systems).

In the nonparametric approach to predictor estimation, the main point is to see the predictor as a system with two inputs (past outputs and inputs of the predicted system) and one output (output predictions). Therefore, predictor design amounts to estimating two impulse responses. In the proposed method, they are modeled as realizations of a Gaussian process (Pillonetto, De Nicolao, Chierici, & Cobelli, 2009; Rasmussen & Williams, 2006; Smola & Schölkopf, 2003). The resulting problem is harder than that treated in Pillonetto and De Nicolao (2010) since not only are the unknown functions assumed to belong to infinite-dimensional spaces but dependence on past outputs involves an operator which itself depends on noisy measurements. In addition, we further refine the kernel (hereafter, the terms kernel and autocovariance will be used indifferently[1]) for system identification introduced in Pillonetto and De Nicolao (2010). In fact, in the present paper, impulse responses are the convolution of an infinite-dimensional nonparametric component and a low-order finite-dimensional one. The latter is used to capture high-frequency oscillations, e.g. poles with negative real part.

The overall scheme for predictor estimation and system identification via predictor error minimization relies on an empirical Bayesian paradigm (Maritz & Lwin, 1989). First, the dimension and the components of the unknown hyperparameter vector, characterizing the prior, are estimated using marginal likelihood optimization in a low-dimensional space. In the second and final step, the hyperparameters are set to their estimates, and minimum variance estimates of the impulse responses are computed. In particular, we show that the optimal estimates of the predictor impulse responses are the solution of a Tikhonov-type variational problem, defined in a reproducing kernel Hilbert space (RKHS) (Aronszajn, 1950; Cucker & Smale, 2001), whose solution is well defined and has a regularization network structure (Poggio & Girosi, 1990).

Numerical experiments, with data generated by ARMAX models of different orders and also by infinite-dimensional systems, show that the proposed approach yields substantial improvement over existing methods both in terms of predictive capability on new data and accuracy in the reconstruction of system impulse responses.

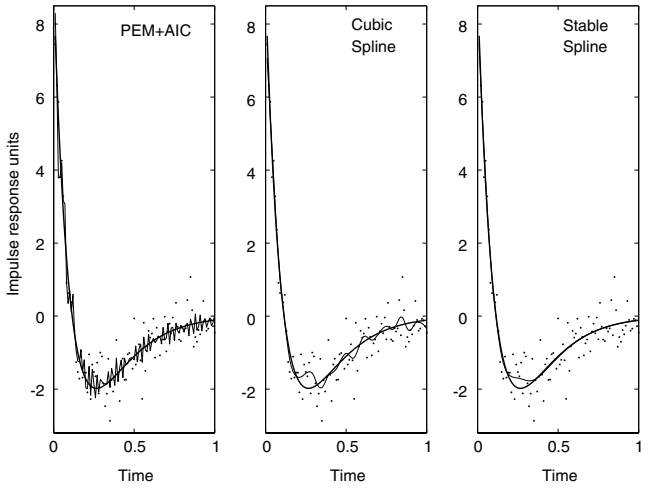

Fig. 1. Output-error impulse response identification: case study. The true impulse response (thick line) and noisy output samples (dots) are plotted in all panels. The solid lines represent the estimates obtained by PEM + AIC (left) and by nonparametric regularization exploiting the cubic spline kernel (middle) and the stable spline kernel (right).

The paper is organized as follows. In Section 2, the differences between the parametric and the kernel-based nonparametric approach to system identification are illustrated via a case study. With respect to the nonparametric viewpoint, the so-called stable spline kernel, originally introduced in Pillonetto and De Nicolao (2010), is also reviewed. Section 3 reports the statement of the predictor estimation problem. In Section 4, the Gaussian prior on predictor impulse responses is defined by refining the stable spline kernel reported in Section 2. In Section 5, a numerical algorithm which returns the unknown components of the prior and the estimates of predictor and system impulse responses is worked out. In Sections 6 and 7, simulated data are used to demonstrate the effectiveness of the proposed approach. Our conclusions, given in Section 8, end the paper. All the proofs are gathered in Appendix B. In the paper, vectors are column vectors and, given a vector or a sequence w, w_i denotes the i-th element of w.

2. The nonparametric approach to linear system identification

2.1. A simple case study

In order to introduce the nonparametric approach, let us consider a simple output-error system identification problem. In particular, we are given a continuous-time system fed with an input u which is a Dirac delta. Thus, the measurement model is

$$y_k = h(t_k) + e_k, \quad k = 1, 2, \ldots, n, \tag{1}$$

where h is the continuous-time unknown impulse response, $\{t_k\}$ are the positive sampling times, $y^+ = [y_1 \quad \ldots \quad y_n]^T$ is the measurements vector, and $e = [e_1 \quad \ldots \quad e_n]^T$ is made of samples of white Gaussian noise with variance σ^2.

A particular instance of this problem is displayed in Fig. 1, where the unknown impulse response (thick line) has to be estimated from 100 noisy output measurements (dots) collected on the unit interval (for details about the generation of this example, see Appendix A).

2.2. The parametric approach to system identification

The classical approach to impulse response estimation relies on a finite-dimensional parameterization of h:

[1] This terminology stems from the fact that, as we shall recall later on, there is a perfect correspondence between Tikhonov-type regularization problems in reproducing kernel Hilbert spaces (RHKSs) with kernel $K(\cdot,\cdot)$ and Bayesian estimation where a Gaussian process prior with covariance function $K(\cdot,\cdot)$ is assigned.

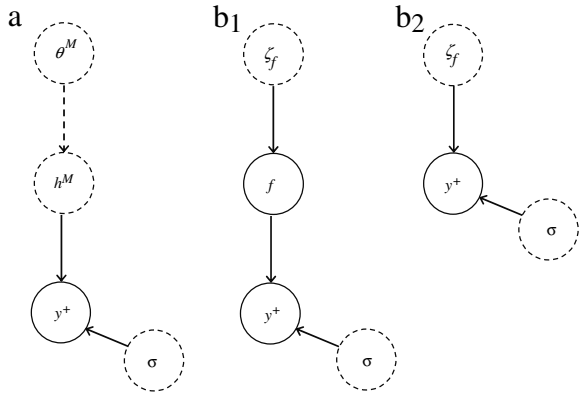

Fig. 2. Bayesian network describing the classical parametric modeling approach to system identification (case a) and the nonparametric one adopted in this paper (case b_1). The latter case with f integrated out (the joint distribution of f and y^+ is marginalized with respect to f) is also reported (case b_2). In the network, dotted lines denote deterministic variables and relationships, while solid lines denote stochastic variables and relationships.

$$y_k = h^M(t_k; \theta^M) + e_k, \quad k = 1, 2, \ldots, n, \tag{2}$$

where θ^M is the d^M-dimensional vector which gathers all the unknown parameters of the model once a certain structure M is postulated. This is graphically depicted by the Bayesian network in Fig. 2 (left side). Nodes and arrows are either dotted or solid depending on being representative of either deterministic or stochastic quantities/relationships. Thus, one can see that θ^M is a deterministic parameter vector whose knowledge fully defines h^M.

For a given model structure M, the maximum likelihood estimate of θ^M is

$$\hat{\theta}^M = \arg\max_{\theta \in \mathbb{R}^{d^M}} \mathbf{p}_{\theta^M}(y^+), \tag{3}$$

where $\mathbf{p}_{\theta^M}$ denotes the probability density function of y^+. In real applications, the model structure M, and hence the dimension d^M of θ^M, is unknown, and must be inferred from data. One of the most commonly used approaches for the selection of model complexity hinges on the Akaike criterion (AIC) which, under the stated assumptions, is formulated as

$$\hat{M} = \arg\min_{M \in \mathscr{M}} \quad 2d^M + n \ln[RSS(\hat{\theta}^M)], \tag{4}$$

where $\mathscr{M}$ is a set of competitive model structures and RSS is the residual sum of squares.

According to (3), (4), we proceed to solve the case study assuming that the z-transform $H(z)$ of $h(t)$ is the ratio of two polynomials, both with maximum allowed order equal to 10. The order of the two polynomials chosen by the AIC is 8 for both the numerator and the denominator of $H(z)$. In the left panel of Fig. 1, the solid line indicates the estimate of the impulse response obtained by using the classical PEM approach, as implemented in the `oe.m` function of the MATLAB System Identification Toolbox (Ljung, 2007). The estimate obtained is not satisfactory, and it suffers from overfitting. This kind of result is confirmed by a Monte Carlo study made of 300 runs (see Appendix A for details). In particular, the mean of the 300 values of the relative root mean square errors, denoted by $\overline{Err}$ and defined in (55), is reported in Table 1. In this table, in addition to the results of PEM complemented with AIC-based order selection, we also report results obtained by an *oracle* that, at any run, selects the ideal model order which minimizes the reconstruction error. Obviously, this provides an upper bound for the performance of the PEM. One can notice that the oracle yields $\overline{Err} = 0.031$ while the AIC gives $\overline{Err} = 0.098$.

These findings are not completely unexpected. In fact, the tendency of the AIC to overfit is well documented in the statistical literature (Hurvich & Tsai, 1989; Kass & Raftery, 1995). In this regard, it is useful to summarize the limitations of parametric approaches equipped with the AIC.

Table 1
Monte Carlo study (Section 2): Average over 300 runs of the relative root mean square error (see (54) in Appendix A) obtained by PEM + AIC, PEM + oracle, and by the nonparametric approaches relying on the stable spline and the cubic spline kernels.

Estimator	PEM + AIC	PEM + oracle	Stable spline	Cubic spline
$\overline{Err}$	0.098	0.031	0.033	0.064

- The AIC is based on an approximation of the likelihood that is only asymptotically exact. This undermines the applicability of the theory when the ratio n/d^M is not sufficiently large. Indeed, the possible negative bias in the estimate of the Kullback–Leibler divergence can lead to overfitting (Hurvich & Tsai, 1989).
- From (3) and (4), it is also evident that the AIC selects the optimal model without taking into account the uncertainty of the estimated parameters.
- The AIC evaluation calls for the solution of several nonlinear optimization problems, one for each model M in $\mathscr{M}$. Since these problems are defined in possibly large-dimensional spaces, computational complexity and local maxima can be an issue.

2.3. The nonparametric Gaussian regression approach to system identification

Under the framework of Gaussian regression (Rasmussen & Williams, 2006), instead of postulating a parametric structure for h, the impulse response is regarded as the realization of a stochastic process f (Pillonetto & De Nicolao, 2010). In particular, f is modeled as a continuous-time zero-mean Gaussian process with autocovariance $\lambda_f^2 \Sigma(t, \tau)$, where $\Sigma : \mathbb{R}^+ \times \mathbb{R}^+ \mapsto \mathbb{R}$ and $\lambda_f \in \mathbb{R}^+$ is an unknown hyperparameter.[2] As such, h is assumed to belong to an infinite-dimensional function space. The new paradigm is graphically depicted in Fig. 2 (middle). In comparison with the parametric scenario, see Fig. 2 (left), the first notable difference is that the vector θ^M is now replaced by the hyperparameter vector ζ_f, whose dimension is fixed,[3] which contains λ_f and other possible parameters characterizing the autocovariance of f. In addition, while θ^M fully defines h^M in deterministic terms, ζ_f defines only the statistics of f, namely its autocovariance. This explains why, in Fig. 2, ζ_f is connected with f by a stochastic relationship.

Two important quantities can be obtained in closed form starting from the nonparametric model depicted in Fig. 2. The first one is the minimum variance estimate of f conditional on y^+, ζ_f, and σ. In fact, exploiting well-known results on Gaussian processes (Anderson & Moore, 1979), one obtains

$$\mathbb{E}[f(t)|y^+, \zeta_f, \sigma] = \lambda_f^2 \sum_{i=1}^{n} c_i \Sigma(t, t_i), \tag{5}$$

where c_i is the i-th component of the vector

$$c = \Sigma_y^{-1} y^+ \tag{6}$$

and $\Sigma_y \in \mathbb{R}^{n \times n}$, with the (i, j)-th entry given by

$$[\Sigma_y]_{i,j} = \lambda_f^2 \Sigma(t_i, t_j) + \sigma^2 \delta_{ij}, \tag{7}$$

[2] According to the statistical literature, the term *hyperparameter* is here used to indicate a parameter of a prior distribution.

[3] A generalization of this model where such dimension is allowed to vary by introducing a very restricted number of competitive kernels for f will be discussed in Section 4. In this section, for the sake of simplicity, the model is not equipped with the so-called *bias space*, i.e. parametric structures able to enrich the nonparametric model; see also Section 4.1 in Pillonetto and De Nicolao (2010).

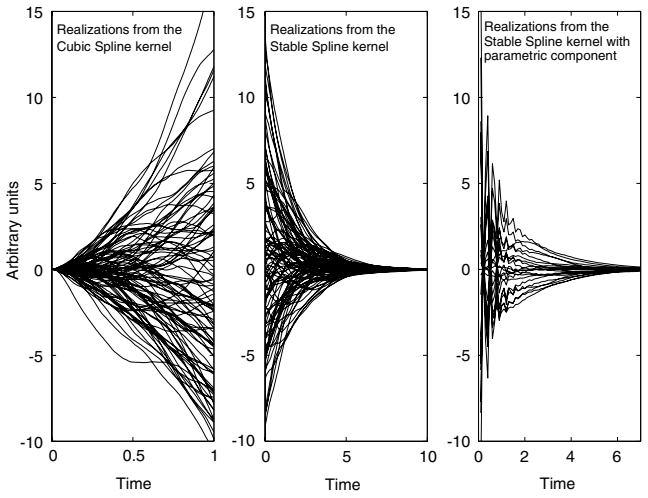

Fig. 3. Realizations of a stochastic process f with autocovariance proportional to the standard cubic spline kernel (left), the new stable spline kernel (middle), and the sampled version of the stable spline kernel enriched with a parametric component defined by the poles $-0.5 \pm 0.6\sqrt{-1}$ (right; see also Section 4.2).

with δ_{ij} the Kronecker delta. The conditional expectation (5) shows that the estimate exhibits the structure of a regularization network, i.e. it is the linear combination of n basis functions $\Sigma(t, t_i)$ that, in this case study, are the kernel sections at the sampling instants.

The second key quantity is the marginal likelihood of y^+, i.e. the marginalization with respect to f of the joint density of y^+ and f. After simple computations, one obtains

$$\mathbf{p}(y^+|\zeta_f, \sigma) = \frac{\exp\left(-\frac{1}{2}(y^+)^T \Sigma_y^{-1} y^+\right)}{\sqrt{\det(2\pi\,\Sigma_y)}}. \tag{8}$$

The model obtained after marginalization is graphically depicted in Fig. 2(right).

To estimate the unknown impulse response, the so-called empirical Bayes approach can be used (Maritz & Lwin, 1989; Pillonetto & De Nicolao, 2010). The first step refers to the network on the right side of Fig. 2 and obtains the hyperparameters by maximizing the marginal likelihood:

$$(\hat{\zeta}_f, \hat{\sigma}) = \arg \max_{\zeta_f,\ \sigma \in \Omega} \mathbf{p}(y^+|\zeta_f, \sigma). \tag{9}$$

where Ω is a suitable parameter set. The second step considers the network in the middle of Fig. 2, conditional on y^+, $\hat{\zeta}_f$, and $\hat{\sigma}$. Thus, the estimate of the impulse response is $\mathbb{E}[f(t)|y^+, \hat{\zeta}_f, \hat{\sigma}]$, given by (5).

The quality of the estimate coming from the nonparametric scheme will crucially depend on the kernel chosen to model the autocovariance of f. In the literature on Gaussian regression, the prior distribution usually reflects the knowledge that the unknown function, and possibly some of its derivatives, are continuous with bounded energy. In this case, the most widely used approach models f as the m-fold integral of white Gaussian noise. Thus, the autocovariance of f is assumed to be proportional to

$$W_m(s, t) = \int_0^1 G_m(s, u) G_m(t, u) \mathrm{d}u, \tag{10}$$

where

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

This is the autocovariance associated with the Bayesian interpretation of m-th-order smoothing splines (Wahba, 1990). In particular, when $m = 2$, one obtains the cubic spline kernel (see e.g. DeNicolao, and Marchesi (2007)):

$$W_2(s, t) = \frac{st \min\{s, t\}}{2} - \frac{(\min\{s, t\})^3}{6}. \tag{11}$$

When this kernel is used, from (5) we see that the estimate of the function is given by a linear combination of n basis functions of the type $W_2(t_k, \cdot)$ which, in view of (11), are just cubic splines, i.e. polynomials up to third-order.

In the system identification scenario, the main drawback of the kernel (10) is that it does not account for impulse response stability. In fact, the variance of f increases over time. This can be easily appreciated by looking at Fig. 3(left), which displays 100 realizations drawn from a zero-mean Gaussian process with autocovariance proportional to $\Sigma = W_2$. If we consider impulse response identification, the consequences of this drawback are illustrated in the middle panel of Fig. 1, where the solid line indicates the estimate of the impulse response obtained by using the cubic spline kernel ($\Sigma = W_2$ in (5)–(7)), with hyperparameters λ_f and σ determined via marginal likelihood optimization; see (9). In this case, two additional parameters are added to the model to account for the initial value $h(0)$ and the initial derivative $\dot{h}(0)$; see subsections 1.3–1.5 in Wahba (1990) for all the details. The obtained estimate suffers from oscillations in the final part of the experiment due to the fact that the model does not include the stability constraint. The result of the Monte Carlo study of 300 runs is $\overline{Err} = 0.064$; see Table 1.

2.4. Stable spline kernels

One of the key contributions of Pillonetto and De Nicolao (2010) is the definition of a kernel specifically suited to linear system identification which leads to an estimator with favorable bias and variance properties. In particular, it is immediate to see that, if the autocovariance of f is proportional to (11), the variance of $f(t)$ is zero at $t = 0$ and tends to ∞ as t increases. However, if f represents a stable impulse response, we would rather let it have a finite variance at $t = 0$ which goes exponentially to zero as t tends to ∞. This property can be ensured by considering autocovariances that are proportional to the class of kernels given by

$$K_m(s, t) = W_m(\mathrm{e}^{-\beta s}, \mathrm{e}^{-\beta t}), \quad s, t \in \mathbb{R}^+, \tag{12}$$

where β is a positive scalar governing the decay rate of the variance (Pillonetto & De Nicolao, 2010). In practice, β will be unknown, so it is convenient to treat it as a further hyperparameter to be included in the vector ζ_f which is estimated via the likelihood maximization (9). It can be easily seen that

$$K_m(s, t) = \int_0^\infty H_m(s, u) H_m(t, u) \beta \mathrm{e}^{-\beta u} \mathrm{d}u \tag{13}$$

$$H_m(r, u) = G_m(\mathrm{e}^{-\beta r}, \mathrm{e}^{-\beta u}), \tag{14}$$

which shows that the new kernel K_m models $f(t)$ as multiple integration from t to $+\infty$ of a white noise whose variance decays exponentially to zero.

In view of (11), (12), if $m = 2$, the autocovariance becomes the stable spline kernel originally introduced in Pillonetto and De Nicolao (2010):

$$K_2(t, \tau) = \frac{\mathrm{e}^{-\beta(t+\tau)} \mathrm{e}^{-\beta \max(t,\tau)}}{2} - \frac{\mathrm{e}^{-3\beta \max(t,\tau)}}{6}. \tag{15}$$

The following proposition, taken from Pillonetto and De Nicolao (2010), shows that assuming an autocovariance proportional to K_2 constrains all the realizations to be asymptotically stable.

Proposition 1. *Let f be zero-mean Gaussian with autocovariance given by the stable spline kernel. Then, with probability one, the realizations of f are continuous impulse responses of BIBO stable dynamic systems.*

Notice that the above result is different from that reported in McVinish et al. (2006), where one can find a prior that guarantees the stability of impulse responses only on average, with no smoothness enforced.

The effect of the stability constraint can be appreciated by looking at Fig. 3(middle), which displays 100 realizations drawn from a zero-mean Gaussian process with autocovariance proportional to K_2 with $\beta = 0.4$.

In the right panel of Fig. 1, the solid line indicates the estimate of the impulse response obtained by using the empirical Bayes approach described in the previous subsection with f modeled using the stable spline kernel. Notice that the estimate is very close to the true profile. Table 1 reveals that this approach outperforms all other methods: $\overline{Err}$ from the Monte Carlo study is 0.033, very close to the performance of the PEM equipped with the oracle ($\overline{Err} = 0.031$). This result is also notable, as the stable spline estimator does better than the parametric approach performing model selection using the AIC among a set of competitive model structures including also the true model that generated the data. The reason is that the stable spline estimator overcomes the drawbacks of the parametric approaches listed at the end of Section 2.2. To be more precise:

- in the nonparametric approach, the marginal likelihood, which is a function of the hyperparameters λ_f, β, and σ, is exact, irrespective of the sample size;
- the stable spline estimator accounts for impulse response uncertainty because the hyperparameter likelihood is obtained after marginalizing with respect to the stochastic impulse response;
- due to marginalization, the domain of the marginal likelihood is a three-dimensional space. Thus, the issue of local maxima is far less critical. In fact, instead of solving several nonlinear optimization problems, one is faced with only one optimization problem in a very low-dimensional domain.

Remark 2. The focus of the present paper is on the comparison of parametric prediction error methods against nonparametric identification techniques. Although other approaches like the instrumental variable (IV) one are beyond our horizon, we applied IV identification to the Monte Carlo study described in Section 2.2, performing model selection with either the BIC or the YIC (Young, 2000); see Appendix A for implementation details. Compared to the average estimation error reported in Table 1, the two IV methods yielded 0.077 and 0.079, respectively. As suggested by an anonymous reviewer, we have also employed a continuous-time implementation where the unknown impulse response is modeled as the convolution of a rectangular signal with support on [0, 0.01] and a rational transfer function whose order is estimated by the BIC. In this case, the error decreases to 0.041. These results are better than those with PEM + AIC, but still confirm the potential advantage ensuing from nonparametric identification techniques.

3. Statement of the predictor identification problem

Now, the predictor identification problem is considered. In what follows, ℓ_1 denotes the space of impulse responses $\{f_k\}_{k=0}^{+\infty}$ of BIBO stable discrete-time causal SISO systems. In addition, I_n will indicate the $n \times n$ identity matrix.

All the subsequent derivation is carried out in discrete time. Nevertheless, the continuous-time results of the previous section can be immediately applied by regarding discrete-time stochastic processes as the sampled version of continuous-time ones.

We are given a finite set of input–output data $\{u_k\}$, $\{y_k\}$ that, with some abuse of notation, will both denote jointly stationary stochastic processes with zero mean and finite variance and their sample values. It is a standard fact that the second-order statistics of $\{u_k\}$, $\{y_k\}$ are compatible with a linear dynamical model of the form

$$y_t = \sum_{k=1}^{\infty} q_k u_{t-k} + \sum_{k=0}^{\infty} w_k e_{t-k}, \tag{16}$$

where $\{e_k\}$ is the one-step-ahead (linear) prediction error of y_t given the joint past of u and y. Our problem is to estimate the one-step-ahead predictor of y_t starting from inputs $\{u_k\}$ and outputs $\{y_k\}$. As is well known, provided that the joint spectrum of y and u is bounded away from zero, this predictor is (BIBO) stable.

For the sake of simplicity, in what follows we will assume that u is a stochastic process independent of $\{e_k\}$. However, it is worth stressing that all the outcomes obtained in what follows would still hold if u were a deterministic signal or if output feedback were present in the system, apart from minor modifications in the proofs. We also stress that all the probability densities reported in what follows are conditional on the system input, but we omit this dependence to simplify the notation.

4. Prior for predictor coefficients

Let $\hat{y}(t)$ denote the one-step-ahead prediction of y_t. The typical approach to estimate predictor coefficients relies on a finite-dimensional parameterization of $\hat{y}(t)$:

$$\hat{y}(t;\theta) = \sum_{k=1}^{\infty} a_k(\theta) y_{t-k} + \sum_{k=1}^{\infty} b_k(\theta) u_{t-k}, \tag{17}$$

where $\theta \in \mathbb{R}^p$, $a : \mathbb{R}^p \mapsto \ell_1$, and $b : \mathbb{R}^p \mapsto \ell_1$, with $a_k(\cdot)$ and $b_k(\cdot)$ denoting the predictor impulse responses evaluated at instant k. In contrast with (17), according to the nonparametric scenario described in Section 2, we let the predictor impulse responses belong to infinite-dimensional function spaces. To be more specific, the predictor now takes on the form

$$\hat{y}(t;\zeta) = \sum_{k=1}^{\infty} f_k(\zeta) y_{t-k} + \sum_{k=1}^{\infty} g_k(\zeta) u_{t-k} \tag{18}$$

$$f_t(\zeta) = \sum_{k=1}^{t} a_k(\zeta) \bar{f}_{t-k} \qquad g_t(\zeta) = \sum_{k=1}^{t} b_k(\zeta) \bar{g}_{t-k}. \tag{19}$$

The relevant variables in (18) and (19), which will be fully specified in the remaining part of the section, are

- $\bar{f} = \{\bar{f}_k\}$ and $\bar{g} = \{\bar{g}_k\}$, which indicate zero-mean Gaussian processes, mutually independent and independent of $\{e_k\}$. Their auto-covariances (kernels) are

$$cov(\bar{f}_i, \bar{f}_j) = \lambda_f^2 K_{\bar{f}}(i,j) \tag{20}$$

$$cov(\bar{g}_i, \bar{g}_j) = \lambda_g^2 K_{\bar{g}}(i,j), \tag{21}$$

  where $K_{\bar{f}}, K_{\bar{g}} : \mathbb{N} \times \mathbb{N} \mapsto \mathbb{R}$, while λ_f and λ_g are unknown hyperparameters contained in ζ;
- the sequences a and b, which belong to ℓ_1 and represent finite-dimensional components of the model, parameterized by ζ. Their knowledge, together with that of $\lambda_f^2 K_{\bar{f}}$ and $\lambda_g^2 K_{\bar{g}}$, fully defines the autocovariances of f and g;
- the vector ζ containing unknown hyperparameters.

4.1. Choice of the kernels associated with $\bar{f}$ and $\bar{g}$

Hereafter, the notation K indicates the stable spline kernel obtained setting $m = 2$, i.e. $K = K_2$. As far as the choice of the autocovariances of $\bar{f}$ and $\bar{g}$ is concerned, $K_{\bar{f}}$ and $K_{\bar{g}}$ are the sampled versions of K, i.e.

$$K_{\bar{f}}(k,j) = K(k,j;\beta_f) \tag{22}$$

$$K_{\bar{g}}(k,j) = K(k,j;\beta_g), \quad \forall k,j \in \mathbb{N}. \tag{23}$$

The hyperparameters β_f and β_g thus represent the asymptotic exponential decay rates of the variance of $\bar{f}$ and $\bar{g}$ which will be tuned from data together with the scale factors λ_f^2 and λ_g^2.

4.2. Choice of the finite-dimensional components

The sequences a and b in (19), i.e. the finite-dimensional components of f and g, are used to enhance the flexibility of the predictor. In fact, their role is to capture dynamics, such as high-frequency poles, which are hardly represented by the smooth processes $\bar{f}$ and $\bar{g}$. In particular, in the scenario of ARMAX model identification, it is convenient to set $\beta \doteq \beta_f = \beta_g$ in (22) and (23). Then, we let $a_k(\theta) = b_k(\theta), \forall \zeta, k$, where the subvector $\theta \in \mathbb{R}^l$ of ζ specifies poles with negative real part. To be more specific, θ enters the zeta-transforms $F_a(z)$ and $F_b(z)$ of a and b as follows:

$$F_a(z) = F_b(z) = \frac{z^l}{P_\theta(z)}, \quad P_\theta(z) = z^l + \sum_{j=1}^{l} \theta_j z^{l-j}, \tag{24}$$

and is such that all the roots of $P_\theta(z)$ belong to the open left unit semicircle in the complex plane.

4.3. The prior model for f and g

With a slight abuse of notation, we think of the autocovariances $K_{\bar{f}}, K_{\bar{g}}$ of $\bar{f}, \bar{g}$ as elements of $\mathbb{R}^{\infty\times\infty}$, where the i-th columns of $K_{\bar{f}}$ and $K_{\bar{g}}$ are the sequences $K(\cdot, i; \beta_f)$ and $K(\cdot, i; \beta_g)$, $i \in \mathbb{N}$, respectively. In addition, notation of ordinary finite-dimensional algebra is used to handle infinite-dimensional objects where convergence will be guaranteed by the exponential decay of the elements of the stable spline kernel. Thus, for instance, if $w \in \mathbb{R}^\infty$, the j-th element of $K_{\bar{f}} w$ is $\sum_{i=1}^{\infty} [K_{\bar{f}}]_{ji} w_i$.

In view of (19), the kernels of f and g are $\lambda_f^2 K_f$ and $\lambda_g^2 K_g$, respectively, where

$$K_f = F_a K_{\bar{f}} F_a^T \tag{25}$$

$$K_g = F_b K_{\bar{g}} F_b^T, \tag{26}$$

with F_a and F_b lower-triangular Toeplitz infinite-dimensional matrices with first column given by a and b, respectively.

To better appreciate the role of the finite-dimensional component of the model, Fig. 3(right panel) shows some realizations (with samples linearly interpolated) drawn from a discrete-time zero-mean normal process having autocovariance proportional to K_f with $\beta = 0.4$ and F_a defined by $\theta = [1 \quad 0.61]$ in (24). In this way, an oscillatory behavior is introduced in the realizations by enriching the stable spline kernel $K_{\bar{f}}$ with the poles $-0.5 \pm 0.6\sqrt{-1}$.

5. Estimation of hyperparameters, predictor coefficients and system impulse responses

In practical applications, the hyperparameters β_f and β_g entering $K_{\bar{f}}$ and $K_{\bar{g}}$, the scale factors λ_f and λ_g, the vector θ entering F_a and F_b, and the innovation variance σ^2 have to be estimated from data together with the predictor coefficients. In addition, the complexity of F_a and F_b, e.g. the number of high-frequency poles to be introduced in the prior, may not be known in advance. For these reasons, it is useful to introduce the vector ζ^M, which gathers all the unknown parameters of the nonparametric model once a certain structure M for a and b is postulated.

5.1. Handling initial condition effects

We start by considering a situation in which ζ^M is perfectly known. To simplify the notation, the dependence on such a vector is often omitted. Now, let $A \in \mathbb{R}^{n\times\infty}$ and $B \in \mathbb{R}^{n\times\infty}$, where

$$[A]_{ji} = y_{j-i}, \tag{27}$$

$$[B]_{ji} = u_{j-i}, \quad j = 1,\ldots,n, i \in \mathbb{N}. \tag{28}$$

In view of (18), it holds that

$$y^+ = A(y^+, y^-)f + B(u)g + e, \tag{29}$$

where u is the input sequence, while

$$y^+ = [y_1 \quad y_2 \quad \ldots \quad y_n]^T \tag{30}$$

$$y^- = [y_0 \quad y_{-1} \quad y_{-2} \ \ldots]^T \tag{31}$$

$$e = [e_1 \quad e_2 \quad \ldots \quad e_n]^T, \tag{32}$$

where, as in Section 2, e contains the innovations. It is useful also to introduce the vector

$$y_a^- = [y_0 \quad y_{-1} \quad \ldots \quad y_{-r+1}]^T, \tag{33}$$

which contains only the r components of y^- which are measured.

Since y^- is never completely known, i.e. $r < \infty$ in (33), a solution to handle the initial conditions consists of setting its unknown components to zero, introducing an error which goes to zero as n increases; see e.g. Section 3.2 in Ljung (1999). Thus, in what follows we assume perfect knowledge of $A(y^+, y^-)$, i.e. $A(y^+, y^-) = A(y^+, y_a^-)$. Furthermore, we will exploit the following approximation for the joint density of y^+, f, g, and y^-:

$$\begin{aligned}\mathbf{p}(y^+, f, g, y^-|\zeta) &= \mathbf{p}(y^+|f, g, y^-, \zeta)\mathbf{p}(f, g|y^-, \zeta)\mathbf{p}(y^-|\zeta)\\ &\approx \mathbf{p}(y^+|f, g, y^-, \zeta)\mathbf{p}(f, g|\zeta)\mathbf{p}(y^-);\end{aligned} \tag{34}$$

i.e. the past y^- is assumed neither to affect the prior on f and g nor to carry information on the hyperparameters.

To simplify the notation, dependence on y^- is hereafter omitted as well as dependence of A and B on y^+ and u.

5.2. Estimation of the predictor coefficients for known ζ^M

In the following, $\mathscr{H}_f$ and $\mathscr{H}_g$ denote the reproducing kernel Hilbert spaces (Aronszajn, 1950) of deterministic functions on $\mathbb{N}$, associated with K_f and K_g, with norms denoted by $\|\cdot\|_{\mathscr{H}_f}$ and $\|\cdot\|_{\mathscr{H}_g}$, respectively.

For a given model structure M, we use f^{MV} to indicate the minimum variance estimator of f, i.e. $f^{MV} = \mathbb{E}[f|y^+, \zeta^M]$. The minimum variance estimator g^{MV} is defined in the same way. The following result shows that, under mild conditions, a.s. these estimates belong to the spaces $\mathscr{H}_f$ and $\mathscr{H}_g$, are solutions of a Tikhonov-type variational problem, and admit the structure of a regularization network (Poggio & Girosi, 1990).

Proposition 3. *Let the roots of P_θ in (24) be strictly inside the unit circle with the kernels K_f and K_g defined as in (25) and (26). Also, let $\{y_t\}$ and $\{u_t\}$ be zero-mean, finite variance stationary stochastic processes. Then, under the approximation (34), a.s., we have*

$$\begin{aligned}(f^{MV}, g^{MV}) = \arg \min_{h_f\in\mathscr{H}_f,\ h_g\in\mathscr{H}_g} &\|y^+ - Ah_f - Bh_g\|^2\\ &+ \gamma_f\|h_f\|_{\mathscr{H}_f}^2 + \gamma_g\|h_g\|_{\mathscr{H}_g}^2,\end{aligned} \tag{35}$$

where $\|\cdot\|$ *is the Euclidean norm,* $\gamma_f = \sigma^2/\lambda_f^2$, *and* $\gamma_g = \sigma^2/\lambda_g^2$. *Moreover, a.s., we also have*

$$f^{MV} = \lambda_f^2 K_f A^T c \qquad g^{MV} = \lambda_g^2 K_g B^T c, \tag{36}$$

where

$$c = (\lambda_f^2 A K_f A^T + \lambda_g^2 B K_g B^T + \sigma^2 I_n)^{-1} y^+. \quad \square \tag{37}$$

5.3. Estimation of the hyperparameters and predictor structure

Given a predictor structure M, the hyperparameter vector ζ^M can be determined by maximizing the marginal likelihood of y^+, which is now obtained by integrating out f and g from the joint density of y^+, f, and g. This is described in the next proposition.

Proposition 4. *Under the same assumptions as in Proposition 3, the maximum marginal likelihood estimate of* ζ^M *is a.s. well defined and given by*[4]

$$\hat{\zeta}^M = \arg\min_{\zeta^M} J(y^+; \zeta^M), \tag{38}$$

where J, the opposite of the log-marginal likelihood of y^+, *is*

$$J(y^+; \zeta^M) = \frac{1}{2}\ln(\det[2\pi V[y^+]]) + \frac{1}{2}(y^+)^T (V[y^+])^{-1} y^+, \tag{39}$$

with

$$V[y^+] = \lambda_f^2 A K_f A^T + \lambda_g^2 B K_g B^T + \sigma^2 I_n. \quad \square \tag{40}$$

Among the possible nonparametric estimators identified by the choice of M, model selection is performed according to the Akaike criterion AIC, that is, by minimizing

$$\mathrm{AIC}(M) = 2J(y^+; \hat{\zeta}^M) + 2d^M, \quad M \in \mathscr{M}, \tag{41}$$

where d^M is the dimension of ζ^M.

5.4. Numerical algorithm for predictor estimation

As for the practical implementation of the numerical algorithm for predictor estimation, for computational reasons f^{MV} and g^{MV} are truncated to having only q non-zero elements. It is important to stress that the predictor length q is not critical, as it does not involve any kind of trade-off between bias and variance. It is just a value that is large enough to capture the dynamics of the predictor. We are now in a position to summarize the entire numerical procedure for predictor estimation.

Algorithm 1. The input to this algorithm includes the input and output sequences, u, y^+, y_a^-, and the predictor length q, together with a set $\mathscr{M}$ containing competitive structures which define a and b in (19). The outputs of this algorithm are the estimates of the predictor coefficients in (19). The steps are as follows.

(i) Choose the model $\hat{M}$ in $\mathscr{M}$ which minimizes (41).
(ii) Determine $\hat{\zeta}^{\hat{M}}$ using (38) conditional on $\hat{M}$.
(iii) According to the empirical Bayes approach, determine the predictor estimates f^{MV} and g^{MV} via the regularization network (36), with hyperparameters set to $\hat{\zeta}^{\hat{M}}$.

[4] In (38), it is implicit that optimization is restricted to the region where the variances are positive and θ is such that the all the roots of P_θ in (24) belong to the open left unit semicircle in the complex plane.

Remark 5. Note that, if the relationship between predictor impulse responses and system output admitted a linear state-space realization, marginal likelihood evaluation and computation of the solution in the items above could be efficiently computed using finite interval smoothing (Harvey, 1989; Weinert, 2001), as done for example in De Nicolao and Ferrari Trecate (2001), Pillonetto and Saccomani (2006), and Young and Pedregal (1999). Unfortunately this is not the case in general.

5.5. Model reduction

The estimated predictor impulse responses from Algorithm 1 belong, in principle, to an infinite-dimensional space and, as such, are not the impulse responses of a finite-dimensional linear system. In practice, as discussed in the previous section, f^{MV} and g^{MV} have only q non-zero elements, i.e. the predictor is a (long) autoregression of the form

$$y_t = \sum_{k=1}^{q} f_k^{MV} y_{t-k} + \sum_{k=1}^{q} g_k^{MV} u_{t-k} + e_t. \tag{42}$$

It is now convenient to rewrite Eq. (42) in the input–output form as in Eq. (16), i.e.

$$y_t = \frac{\sum_{k=1}^{q} g_k^{MV} z^{-k}}{1 - \sum_{k=1}^{q} f_k^{MV} z^{-k}} u_t + \frac{1}{1 - \sum_{k=1}^{q} f_k^{MV} z^{-k}} e_t,$$

where for convenience of notation the Z-transform formalism has been used. The transfer functions

$$\hat{Q}(z) = \frac{\sum_{k=1}^{q} g_k^{MV} z^{-k}}{1 - \sum_{k=1}^{q} f_k^{MV} z^{-k}}, \qquad \hat{W}(z) = \frac{1}{1 - \sum_{k=1}^{q} f_k^{MV} z^{-k}} \tag{43}$$

admit a minimal realization of dimension q, which may be large. Hence, if needed, one could also apply standard model reduction techniques to $\hat{Q}(z)$ and $\hat{W}(z)$ in order to obtain reduced-order approximations, e.g. deterministic balanced truncation (Pernebo & Silverman, 1982) for $\hat{Q}(z)$ and stochastic balanced truncation (Desai & Pal, 1984) for $\hat{W}(z)$.

6. Numerical experiments involving ARMAX models

The performance of the proposed approach was first evaluated by means of five Monte Carlo studies, each consisting of 500 runs. The aim is to estimate an ARMAX model from an identification (training) set and assess the performance of different estimators both in terms of predictive capability on new data (test set) and the quality of reconstruction of the system impulse responses.

6.1. Random generation of ARMAX models

At any run, an ARMAX model of the following form

$$y_t = \sum_{i=1}^{n_y} h_i^y y_{t-i} + \sum_{i=1}^{n_u} h_i^u u_{t-i} + e_t + \sum_{i=1}^{n_e} h_i^e e_{t-i} \tag{44}$$

is generated. To be specific, first, the value for n_y is randomly drawn from a discrete uniform distribution with support on $\{1, 2, \ldots, 20\}$. Then, the MATLAB function `drmodel.m` is used to generate a random stable discrete-time n_y-th-order model. The coefficients of the polynomial defining the denominator of the transfer function provide the coefficients $\{h_i^y\}$. The number and

Table 2
Features characterizing the five Monte Carlo studies.

Experiment	#1	#2	#3	#4	#5
Input for training set	WN	WN	SW	LP1	LP2
Noise for training set and test set	WN	WN	WN	WN	WN
Training set size	1000	200	500	500	200
Input for test set	WN	WN	WN	WN	WN
Test set size	1000	1000	1000	1000	1000

values of the non-zero elements of the numerator define n_u and $\{h_i^u\}$, respectively. In this way, h_1^u is always different from zero with input delay equal to one. Finally, we set $n_e = n_y$, and the function `drmodel.m` is used to obtain another Hurwitz polynomial whose coefficients define $\{h_i^e\}$. System and predictor poles are restricted to lie inside the circle of radius 0.95, while the output variance is bounded by 400 (`drmodel.m` is repeatedly called at any run until such requirements are fulfilled).

To obtain information regarding the signal to noise ratios in the identification data, at each run, we computed the quantity $\|q\|_2/\|w\|_2$ which, in view of (16), is the ratio of the 2-norms of the impulse responses governing the deterministic and stochastic parts of the system. The average value of such ratio was 2.1. In particular, the 5%, 25%, 50%, 75%, and 95% quantiles of the ratios were 0.08, 0.45, 1.2, 2.6, and 7.7, respectively.

6.2. Features of the five Monte Carlo experiments

At each run, the ARMAX model is used to generate a training set and a test set, consisting of data collected after getting rid of initial condition effects. Five different experimental conditions are considered. They are characterized in terms of the following aspects.

- *Input u generating the training set.* This is white noise of unit variance (WN) in the first two experiments, while in the third one it is a square wave of period 40, which alternates between levels 0 and 1 (SW). In the last two experiments, we use two different low-pass signals. In the fourth one, it is a random Gaussian signal (LP1) generated using the `idinput.m` MATLAB function with band [0, 0.8], where 0 and 0.8 are the lower and upper limits of the passband, expressed as fractions of the Nyquist frequency. In the last experiment, the input power is even more concentrated at low frequencies since it is the filtered Gaussian signal (LP2) with cut-off frequency equal to 10 kHz described in Schoukens, Suykens, and Ljung (2009).
- *Disturbance noise generating the training set and the test set.* In all cases this is WN with variance equal to one.
- *Size of training set.* In the five experiments this is 1000, 200, 500, 500, and 200.
- *Input u generating the test set.* This is always WN of unit variance.
- *Size of test set.* This is always equal to 1000.

All the relevant information concerning the 5 experiments is summarized in Table 2. Notice that the first and the last experiment exhibit, respectively, the most favorable condition (1000 output data generated using WN as input) and the most adverse one (training set size equal to 200 and strongly low-pass input).

6.3. Performance indices: quality of the estimated predictor and model

The first performance index regards predictive capability on new data. In particular, at each Monte Carlo run, the estimates $\{\hat{f}_k\}$ and $\{\hat{g}_k\}$ of the predictor coefficients are first obtained. Let $\{y_k^{new}\}_{k=1}^{1000}$ and $\{u_k^{new}\}_{k=1}^{1000}$ denote the test set consisting of new output and input data, respectively. Then, the one-step-ahead prediction $\hat{y}_t^{new}$ is computed by the estimated predictor. At the j-th Monte Carlo run, the root mean square one-step-ahead prediction error is

$$err_{j1} = \sqrt{\frac{\sum_{t=1}^{1000}(\hat{y}_t^{new} - y_t^{new})^2}{1000}}, \tag{45}$$

while the so-called coefficient of determination, which quantifies how much of the output variable variance is explained by the forecast, is

$$COD_j = 1 - \frac{\sum_{t=1}^{1000}(\hat{y}_t^{new} - y_t^{new})^2}{1000 \times V_j}, \tag{46}$$

where V_j denotes the sample variance of $\{y_t^{new}\}_{t=1}^{1000}$. We also define

$$\overline{Err}_1 = \frac{\sum_{j=1}^{500} err_{j1}}{500}, \qquad \overline{COD} = \frac{\sum_{j=1}^{500} COD_j}{500} \tag{47}$$

to quantify the average root mean square error in the one-step-ahead prediction on new output data and the average coefficient of determination. The indices $err_{j1}(k)$, $COD_j(k)$, $\overline{Err}_1(k)$ and $\overline{COD}(k)$ instead quantify the performance of the estimated k-step-ahead predictor. They are defined as in (45)–(47) except that the one-step-ahead prediction $\hat{y}_t^{new}$ is replaced by the k-step-ahead prediction, computed from the estimated model.

As regards the identification performance, at each Monte Carlo run, the estimates $\{\hat{q}_k\}$ and $\{\hat{w}_k\}$ of the system impulse responses are first obtained. Then, the quality of the estimated model is quantified by

$$err_{j2} = \frac{1}{2}\frac{\|\hat{q}_k - q_k\|_2}{\|q_k\|_2} + \frac{1}{2}\frac{\|\hat{w}_k - w_k\|_2}{\|w_k\|_2}. \tag{48}$$

Finally,

$$\overline{Err}_2 = \frac{\sum_{k=1}^{500} err_{k2}}{500} \tag{49}$$

measures the average 2-norm of the error relative to the reconstruction of the system impulse responses.

6.4. The evaluated estimators

The following estimators are compared via the Monte Carlo studies.

- *Stable spline.* This is the estimator based on the stable spline kernel. As for the prior on f and g, we set $\beta_1 = \beta_2 = \beta$. Then, hyperparameters $\{\lambda_f, \lambda_g, \beta, \sigma\}$, as well as the poles with negative real part enriching the kernel structure, are determined from data according to Algorithm 1. In particular, the number of poles which can be introduced ranges from 0 and 3, and is determined by the AIC; see (41). In experiments #2 and #5, where the data set size is 200, 50 predictor coefficients are estimated, i.e. we set the predictor length q to 50, while y_a^- and y^+ contain the first 50 and the last 150 available output samples, respectively, i.e. $r = 50$ in (33) and $n = 150$ in (30). In the other three experiments, we set $q = r = 60$. After obtaining the estimates of the predictor coefficients, the system impulse responses are obtained as described in Section 5.5 without adopting model order reduction. From them, for any k, it is straightforward to obtain the estimate of the k-step-ahead predictor (Ljung, 1999).

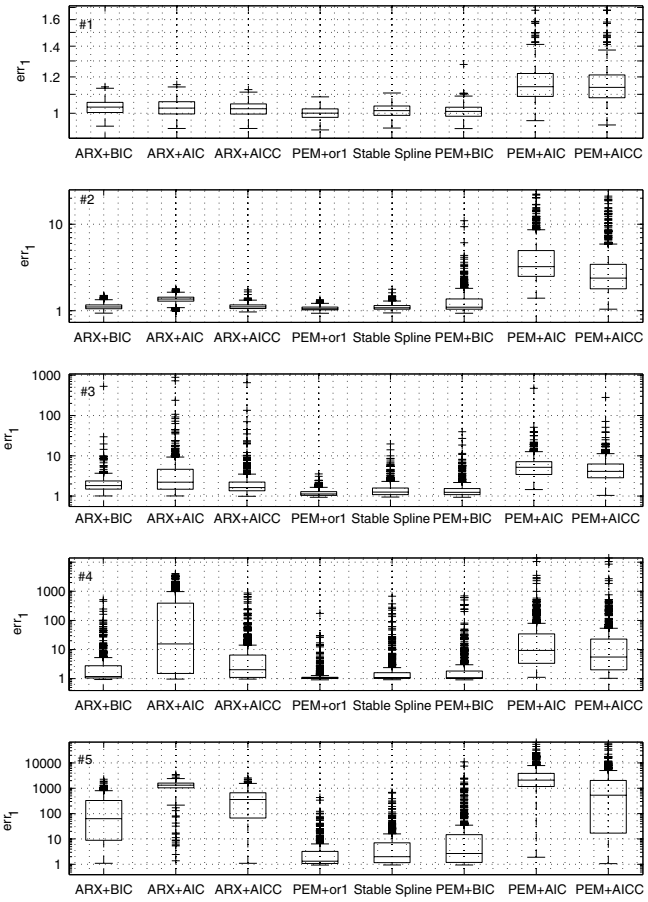

Fig. 4. Five Monte Carlo studies using ARMAX models: boxplot of prediction errors.

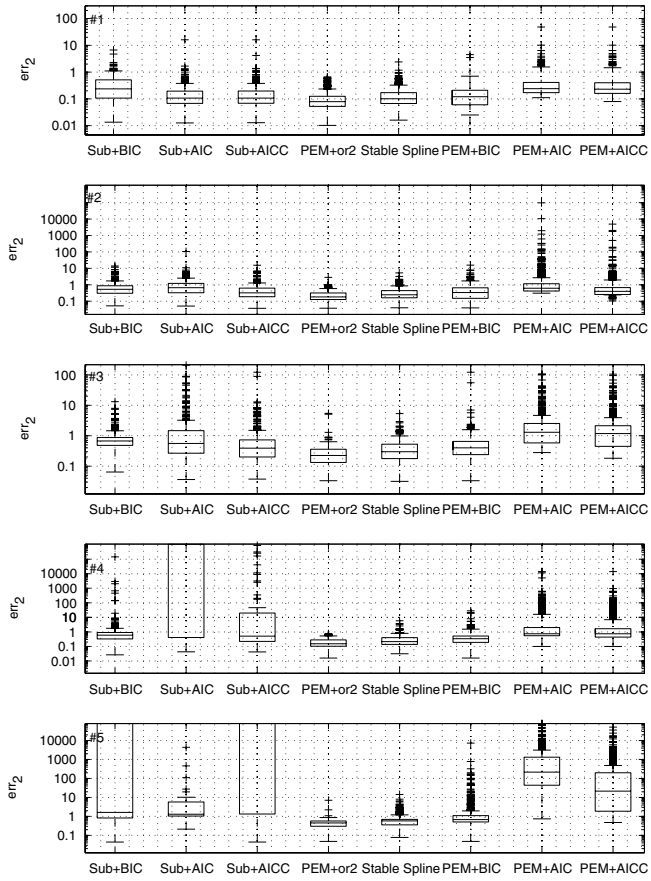

Fig. 5. Five Monte Carlo studies using ARMAX models: boxplot of errors in reconstruction of the system impulse responses.

- *PEM + or*1. The classical PEM approach, as implemented in the `pem.m` function of the MATLAB System Identification Toolbox (Ljung, 2007), equipped with an oracle for one-step-ahead prediction. To be specific, at any run j, this ideal tuning provides an upper bound for the prediction performance of the PEM by selecting those model orders n_y, n_u, and n_e, that minimize err_{j1} in (45).
- *PEM + or*2. The same as above, except that the oracle minimizes err_{j2} in (48). In this way, an upper bound for the performance obtainable by the PEM in system impulse response reconstruction is obtained.
- *PEM + AIC*. The same as above, but with the model orders n_y, n_u, and n_e chosen by the Akaike criterion AIC, i.e.

$$(\hat{n}_y, \hat{n}_u, \hat{n}_e) = \arg \min_{n_y, n_u, n_e \in I_{25}} 2n_p + n_t \ln[RSS(n_y, n_u, n_e)] \quad \text{s.t. } n_y = n_e,\ n_u \le n_y, \tag{50}$$

  where n_t is the size of the training set, $n_p = 1 + n_y + n_u + n_e$, $I_{25} = \{1, 2, \ldots, 25\}$, and *RSS* is the residual sum of squares. The latter is computed from the output predicted by the estimated model using the `predict.m` MATLAB function.
- *PEM + AICC*. The same as above, but now the model orders are chosen by the corrected version of the Akaike criterion AICC (Hurvich & Tsai, 1989), i.e.

$$(\hat{n}_y, \hat{n}_u, \hat{n}_e) = \arg \min_{n_y, n_u, n_e \in I_{25}} \frac{n_t(n_t + n_p)}{n_t - n_p - 2} + n_t \ln[RSS(n_y, n_u, n_e)] \quad \text{s.t. } n_y = n_e,\ n_u \le n_y.$$

- *PEM + BIC*. The same as above, but now the model orders are chosen by using the BIC, i.e.

$$(\hat{n}_y, \hat{n}_u, \hat{n}_e) = \arg \min_{n_y, n_u, n_e \in I_{25}} \log(n_t) n_p + n_t \ln[RSS(n_y, n_u, n_e)] \quad \text{s.t. } n_y = n_e,\ n_u \le n_y. \tag{51}$$

- *ARX + AIC*. This estimator determines the predictor coefficients via ARX modeling. The model orders of the two polynomials defining the predictor structure are equal and are chosen by the AIC. The maximum allowed lengths of the predictor impulse responses are 40 when the size of the data set is 200 and 60 in the other situations.
- *ARX + AICC*. The same, except that the model orders are chosen by the AICC.
- *ARX + BIC*. The same, except that the model orders are chosen by the BIC.
- *Sub + AIC*. This estimator determines the system impulse responses by using a subspace algorithm which relies upon the results achieved by ARX + AIC. The order of the state-space model is automatically determined using the singular value criterion; see also Chiuso, Pillonetto, and De Nicolao (2008), Chiuso (2007) and Bauer (2001).
- *Sub + AICC*. The same, except that the subspace algorithm relies on the predictor coefficients obtained by ARX + AICC.
- *Sub + BIC*. The same, except that the subspace algorithm relies on the predictor coefficients obtained by ARX + BIC.

6.5. Results

Figs. 4 and 5 illustrate the performance of the estimators. We start by commenting on the results regarding prediction performance. The panels of Fig. 4 report boxplots of $\{err_{j1}\}$ for the eight different estimators, while Table 3 reports the $\overline{Err}_1$

Table 3
Five Monte Carlo studies using ARMAX models. Average prediction error $\overline{Err}_1$ as a function of experiment number and employed estimator.

Experiment	ARX + BIC	ARX + AIC	ARX + AICC	PEM + or1	Stable spline	PEM + BIC	PEM + AIC	PEM + AICC
#1	1.03	1.03	1.02	1.00	1.01	1.01	1.17	1.16
#2	1.12	1.38	1.13	1.07	1.10	1.42	4.47	2.93
#3	3.34	8.58	6.52	1.22	1.61	1.74	7.21	6.12
#4	7.72	356	19.1	1.89	7.71	12.3	117	113
#5	208	130	437	6.77	11.1	103	3.93e3	1.93e3

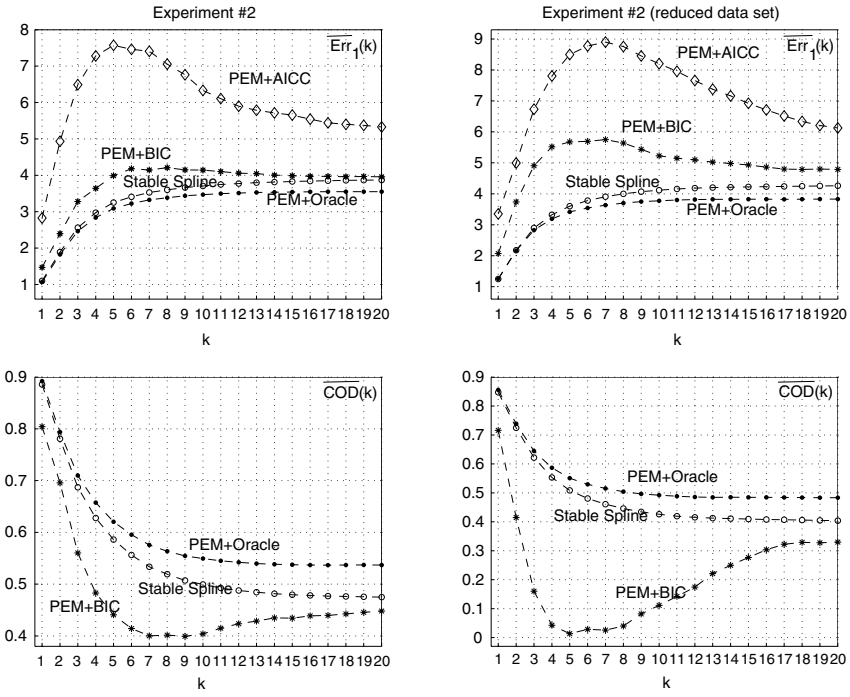

Fig. 6. Monte Carlo study #2 (left) and its variant with halved size of the data set (right): $\overline{Err}_1(k)$ (top), i.e. average error in k-step-ahead prediction, and $\overline{COD}(k)$ (bottom), i.e. average coefficient of determination relative to k-step-ahead prediction, using PEM + oracle (•), PEM + BIC (∗), PEM + AICC (◇) and Stable spline (○).

values obtained. The performance reference is represented by PEM + or1. The results obtained by the PEM suggest that, except in experiment #1, the AIC performance is really poor. This is in perfect agreement with what already discussed in Hurvich and Tsai (1989) and in Section 2 of the present paper: the negative bias affecting the estimate of the Kullback–Leibler divergence often leads to overfitting. When the AICC is used, the results improve, but only marginally. The performance of PEM + BIC is much better than that of the AICC-based predictor, even if in the last two experiments it performs much worse than the oracle. Similar considerations hold for ARX-based estimators with the AIC and AICC, whose performance is largely unsatisfactory when the training set is not generated using WN. The best performance is instead obtained by the Stable spline. In fact, the $\overline{Err}_1$ values achieved by using the approach proposed in this paper are always close, or at least comparable, to those of PEM + or1. Notice also that, in the last experiment, the Stable spline estimator's average performance is better than that of PEM + BIC by a tenfold factor.

To further illustrate the advantages of the stable spline estimator, we focus on Experiment #2, considering $\overline{Err}_1(k)$ relative to the k-step-ahead prediction. Fig. 6 (top left) displays $\overline{Err}_1(k)$ obtained by Stable spline, PEM + BIC, PEM + AICC, and PEM equipped with an oracle. Similarly to PEM + or1, the latter provides, for any value of k, the best k-step-ahead predictor obtainable using the PEM. It is apparent that Stable spline outperforms all the other implementable approaches, with its performance close to that of the oracle. Fig. 6 (bottom left) displays $\overline{COD}(k)$ obtained in the same experiment by Stable spline, PEM + BIC, and PEM equipped with an oracle. The stable spline performance is similar to that of the oracle. In addition, the forecast obtained by the stable spline estimator captures much more output variance than PEM + BIC. Consider now a variant of Experiment #2, where only half of the identification data are available to the estimators (100 output data in place of 200). In this case, Stable spline is implemented setting the predictor length q as well as r in (33) to 30. The results are reported in the right panels of Fig. 6 and confirm the robustness of the stable spline estimator.

Finally, we consider the results regarding performance in system impulse response reconstruction. The panels of Fig. 5 report boxplots of $\{err_{j2}\}$ obtained by the eight estimators, while Table 4 reports the corresponding $\overline{Err}_2$ values. The reference now is PEM + or2. Even in this scenario, Stable spline outperforms all the other implementable approaches, with its performance always close to that of the oracle. The performance of Sub is satisfactory only when the input u, used to generate the training set, is WN (experiment #1 and #2) and ARX + AICC is used. However, even in these cases, $\overline{Err}_2$ is almost twice as large as that achieved by Stable spline. In the other three experiments, Sub performs very poorly, irrespective of the model order selection criterion employed. The PEM performance is really unsatisfactory when the AIC or AICC is used, while PEM + BIC appears the best competitor of Stable spline. However, in experiments #2, #3, and #4, $\overline{Err}_2$ is 2 or even 3 times larger than that achieved by Stable spline, while it is 30 times larger in the last experiment. In particular, notice that even when the median of the distribution of the errors is comparable, PEM + BIC leads to longer tails of poor estimates.

Table 4
Five Monte Carlo studies using ARMAX models. Average system identification error $\overline{Err}_2$ as a function of experiment number and estimator employed.

Experiment	Sub + BIC	Sub + AIC	Sub + AICC	PEM + or2	Stable spline	PEM + BIC	PEM + AIC	PEM + AICC
#1	0.42	0.23	0.24	0.13	0.18	0.22	0.73	0.67
#2	0.85	1.24	0.57	0.24	0.32	0.62	9.06e5	21.2
#3	4.22e7	4.41e38	9.12e12	0.28	0.36	0.95	15.2	8.8
#4	3.51e54	3.42e56	3.48e56	0.21	0.31	0.72	78.2	42.6
#5	5.34e98	2.41e82	3.08e97	0.45	0.72	21.1	1.21e3	1.37e3

Table 5
Monte Carlo studies using the Runge function. $\overline{Err}_2$ (average system identification error) as a function of the input used to generate the data set for identification (WN or SW) and estimator employed.

Input	PEM + or2	Stable spline	PEM + BIC	PEM + AICC
WN	0.24	0.17	0.73	0.71
SW	0.35	0.23	1.3	1.9

Remark 6. We have also repeated the five numerical studies using Stable spline with the dimension of the parametric part of the model fixed to 2, i.e. $l = 2$ in (24). The results (not shown) are very close to those obtained when the number of poles with negative real part is chosen by the AIC.

7. Numerical experiments involving infinite-dimensional models

To further illustrate the flexibility of the stable spline estimator, we consider the identification of an output error model whose impulse response is infinite dimensional. In particular, the measurement model is

$$y_t = \sum_{i=1}^{\infty} h_i^u u_{t-i} + e_t, \tag{52}$$

where $\{h_i^u\}$ is a translated and scaled version of the well-known Runge function (Runge, 1901), i.e.

$$h_i^u = \left(1 + 25\left(\frac{i-\mu}{\mu}\right)^2\right)^{-1}, \quad i = 1, 2, \ldots, \tag{53}$$

with μ a positive scalar.

Two Monte Carlo studies of 500 runs are considered in which the system input is WN or SW, while the innovation variance is always 1. At each Monte Carlo run, μ in (53) is randomly drawn from a distribution which is uniform in [5, 25]. Then, the training set consists of 300 input–output pairs collected after getting rid of the initial condition effect.

Employed estimators are PEM + or2, PEM + BIC, PEM + AICC, and Stable spline, with $q = r = 60$, implemented as described in the previous section. In this way, information regarding the output error model structure is not provided, i.e. estimators search for a suitable model within the class of ARMAX models.

We just focus on performance regarding the quality of the impulse response reconstruction. The panels of Fig. 7 report boxplots of $\{err_{j2}\}$ for the four different estimators, while Table 5 reports the $\overline{Err}_2$ values. Remarkably, in both of the experiments, Stable spline performs better than the PEM equipped with the oracle. To understand this point, let us recall that there exists an extensive literature in the field of numerical analysis (see e.g. Berrut and Trefethen (2004); Wahba (1990)) showing that (smooth) functions can be efficiently and arbitrarily well approximated by combinations of splines (or of particular polynomials) centered in suitable nodes. Notice also that Proposition 3 shows that the estimate from the nonparametric estimator adopted in this paper is indeed the linear combination of (filtered versions of) stable splines, the number of nodes being equal to the number of measurements. Instead, insofar as parametric models are concerned, it is possible to assess that a good approximation of the Runge function, e.g. within a 5% relative error, requires a rational transfer function of order close to the value of μ. For instance, if $\mu = 20$, a model (searched inside the ARMAX class in this experiment) containing roughly 60 parameters is needed. On the downside, the estimated parameters of a high-dimensional model may have large variance, especially with small data sets. This not only exposes the PEM to the risk of local minima, but may bias the estimates of PEM + oracle. This explains why in this case PEM + oracle is also outperformed by the stable spline estimator.

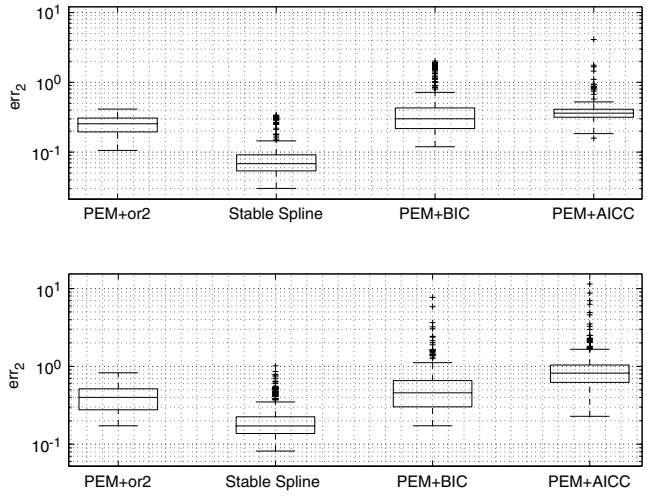

Fig. 7. Monte Carlo studies using the Runge function: boxplot of errors in the reconstruction of the system impulse responses using WN as input (top panel) and SW (bottom panel).

8. Conclusions

The approaches that are currently used for linear predictor design postulate finite-dimensional models which are identified by standard techniques such as least squares and the PEM. So far, nonparametric kernel-based approaches have been mainly used for nonlinear system identification; see e.g. Goethals, Pelckmans, Suykens, and De Moor (2005), Ha Quang, Pillonetto, and Chiuso (2009) and Young et al. (2001). In this paper, which is a follow up of Pillonetto and De Nicolao (2010), we have shown that linear system identification can also benefit from the flexibility of these methods. In particular, we have extended a recently proposed nonparametric paradigm to design the optimal predictor and to identify discrete-time models via prediction error minimization within infinite-dimensional spaces of candidate models. Predictor coefficients are modeled as the convolution between a Gaussian process, which incorporates information on BIBO stability, and a low-order finite-dimensional model which is used to capture high-frequency poles. The predictor structure, as well as the unknown

hyperparameters characterizing the prior, are estimated from data. Then, according to an empirical Bayes paradigm, estimates of the predictor and system impulse responses are obtained in closed form.

In the numerical experiments, the performance of the nonparametric approach is remarkable, also in view of the fact that in almost all the case studies the PEM approach considers a finite number of competitive models among which the true model is present. As discussed in Section 2 and also in Leeb and Pötscher (2005) and Pillonetto and De Nicolao (2010), the model selection required by the PEM may prove critical. In fact, a joint likelihood, associated with a much richer parametric structure, has to be handled. The nonparametric approach, instead, searches the estimate within a much larger and infinite-dimensional space. However, the choice of the regularization parameters, whose tuning plays a role similar to model selection, is performed by optimizing a marginal likelihood defined in a low-dimensional space. Furthermore, in the likelihood maximization, the uncertainty of the unknown impulse response is accounted for.

In future work, we will discuss the asymptotic performance of our nonparametric estimator by obtaining, in the spirit of Smale and Zhou (2007), explicit learning rates of important classes of impulse responses, such as sums of exponentials.

Appendix A

In Section 2, the unknown system impulse response displayed in Fig. 1 is the sum of ten exponentials, i.e.

$$h(t) = \sum_{i=1}^{10} A_i \exp(-\alpha_i t)$$

where

$$A = [19.6, -8.6, -3.3, -5.9, -3.1, 9.7, -19, 11, 7.15, 1.8]$$
$$\alpha = [9.9, 7.5, 5.1, 5.3, 5.7, 9.4, 7.3, 7.8, 7.4, 5.3].$$

In particular, $\{A_i\}$ and $\{\alpha_i\}$ are realizations from mutually independent random variables: each A_i is drawn from a zero-mean Gaussian with standard deviation (SD) equal to 10, while each α_i comes from a uniform distribution with support in [5, 10]. The number of available data is $n = 200$. The first 100 consist of pure measurement noise, and are used by the parametric estimators described below to handle initial condition effects. The last 100 are collected on an uniform grid on the unit interval after applying the Dirac delta at instant 0. The noise SD is 5% of the maximum absolute value of the noiseless output.

As for the Monte Carlo study, at each run a different impulse response h is randomly generated, and 200 output measurements are generated by the same stochastic machinery described above. As regards the parametric estimators, the estimate of h is obtained by PEM + AIC using the `oe.m` function of the System Identification Toolbox for MATLAB, by IV + YIC, and by IV + BIC using `rivid.m`[5] and `rivbjid.m`, respectively, as implemented in the Captain Toolbox for MATLAB (Pedregal, Taylor, Tych, & Young, 2009). The continuous-time implementation instead exploits the function `rivcbjid.m`, equipped with the BIC, still as implemented in the Captain Toolbox. The maximum allowed order of the two polynomials defining the output error model is 10, so 100 candidate models are considered. Regularized nonparametric estimation instead uses the cubic spline kernel (in this case just obtaining the first 100 impulse response coefficients and setting the other ones to zero) and the stable spline kernel. Let $\bar{h}$ and $\hat{h}_j$ be h sampled on 0.01, 0.02, ... and its estimate obtained at the j-th run, respectively. Then, the relative root mean square error at the j-th run is

$$err_j = \frac{\|\bar{h} - \hat{h}_j\|_2}{\|\bar{h}\|_2}, \tag{54}$$

where $\|\cdot\|_2$ is the ℓ_2-norm that is numerically approximated using the first 200 components of $\bar{h}$ and $\hat{h}_j$. Finally, the average error is defined by

$$\overline{Err} = \frac{1}{300}\sum_{j=1}^{300} err_j. \tag{55}$$

Appendix B. Preliminaries

First, some additional notation is introduced. Dependence on ζ^M, y_a^-, u, and sometimes also on y^+ is omitted in what follows to simplify the notation. Let us denote by $\breve{A} \in \mathbb{R}^{n\times q}$ and $\breve{B} \in \mathbb{R}^{n\times q}$ the matrices obtained by retaining only the first q columns of A and B, respectively. Similarly, $\breve{K}_f \in \mathbb{R}^{q\times q}$ and $\breve{K}_g \in \mathbb{R}^{q\times q}$ contain only the first q rows and columns of K_f and K_g, respectively. In addition, we use $\breve{f}$ and $\breve{g}$ to indicate q-dimensional random vectors in correspondence with f and g subject to the constraints $f_k = g_k = 0$ for $k > q$. If such constraints hold, one has

$$y^+ = \breve{A}(y^+)\breve{f} + \breve{B}\breve{g} + e. \tag{56}$$

Let us also recall the following "pseudo-autocovariance" of y^+, already introduced in (40):

$$V[y^+] = \lambda_f^2 A(y^+)K_f A(y^+)^T + \lambda_g^2 B K_g B^T + \sigma^2 I_n,$$

which, if (56) holds, is equivalent to

$$\breve{V}[y^+] = \lambda_f^2 \breve{A}(y^+)\breve{K}_f\breve{A}^T(y^+) + \lambda_g^2 \breve{B}\breve{K}_g\breve{B}^T + \sigma^2 I_n. \tag{57}$$

Proof of Proposition 3. We use A_i and B_i to denote the i-th row of the matrices A and B. The following preliminary lemma is needed.

Lemma 7. *Let the roots of P_θ in (24) be strictly inside the unit circle, with the kernels K_f and K_g defined as in (25) and (26). Then, provided that $\{y_t\}$ and $\{u_t\}$ are zero-mean, finite-variance stationary stochastic processes, the operators $\{A_i\}$ and $\{B_i\}$ are a.s. continuous in the topology of $\mathscr{H}_f$ and $\mathscr{H}_g$.*

Proof. Recall that a functional acting on a reproducing kernel Hilbert space (RKHS) is continuous if its application to the kernel yields a function belonging to the RKHS; see e.g. Aronszajn (1950). Then, it suffices to show that $h_k = A_1 K_f(\cdot, k) \in \mathscr{H}_f$ (a.s.). The same argument will hold for A_i and B_i.

Let us first consider

$$h_k^\ell = \sum_{j=1}^{\ell} A_{1j} K_f(j,k) = \sum_{j=1}^{\ell} y_{1-j} K_f(j,k).$$

This is a finite linear combination of shifted kernel functions and, using the fact that the y_k's are a.s. finite, it is (a.s.) in $\mathscr{H}_f$. In addition, in view of the so-called reproducing property (Aronszajn, 1950), $\|h_k^\ell\|^2_{\mathscr{H}_f} = \sum_{i=1}^{\ell}\sum_{j=1}^{\ell} y_{1-i}y_{1-j}K_f(i,j)$. A sufficient condition for almost sure convergence of a sequence of random variables x_n is that, $\forall \varepsilon > 0$,

$$\sum_{\ell=1}^{\infty} \sup_{m>\ell} \mathbb{P}[|x_m - x_\ell| > \varepsilon] < \infty \tag{58}$$

[5] Such a function implements three different versions of YIC. The one providing the best results has been selected.

(see e.g. Loève (1963) [Ex. 14(b), p. 174]). From (15) and (25), and the fact that P_θ has roots strictly inside the unit circle, we notice that the elements of K_f decay exponentially fast as a function of both row and column indices. Now, setting $x_\ell = h_k^\ell$ and $|\cdot| = \|\cdot\|_{\mathscr{H}_f}$ in (58), we have that

$$\begin{aligned}
&\sum_{\ell=1}^{\infty} \sup_{m\geq\ell} \mathbb{P}[\|h^m - h^\ell\|_{\mathscr{H}_f} > \varepsilon] \\
&= \sum_{\ell=1}^{\infty} \sup_{m\geq\ell} \mathbb{P}\left[\sum_{i=\ell}^{m}\sum_{j=\ell}^{m} y_{1-i}y_{1-j}K_f(i,j) > \varepsilon^2\right] \\
&\leq \sum_{\ell=1}^{\infty} \sup_{m\geq\ell} \frac{\sum_{i=\ell}^{m}\sum_{j=\ell}^{m}\mathbb{E}|y_{1-i}y_{1-j}||K_f(i,j)|}{\varepsilon^2} \\
&\leq \sum_{\ell=1}^{\infty} \frac{\rho^2}{\varepsilon^2}\sum_{i=\ell}^{\infty}\sum_{j=\ell}^{\infty}|K_f(i,j)| < \infty,
\end{aligned}$$

where the Markov inequality, a bound on the second moments ($\mathbb{E}|y_{1-i}y_{1-j}| \leq \rho^2$, where ρ^2 is the variance of y_i), and the exponential decay of $|K_f(i,j)|$ have been used. This shows that h_k^ℓ is (a.s.) a Cauchy sequence in the topology of $\mathscr{H}_f$, and thus has (a.s.) a limit in the space. □

Now, as regards the first part of the proof, we will derive the minimum variance estimates of predictor impulse responses under the truncated model given by (56). Then, we will obtain (36) and (37) by letting q tend to ∞ so that the finite-dimensional prior on $\breve{f}$ and $\breve{g}$ tends (a.s.) to the original one on f and g.

If (56) holds, using the factorization

$$\begin{aligned}
\mathbf{p}(y^+|\breve{f},\breve{g}) = \mathbf{p}(y_n|y_{n-1},\ldots,y_1,\breve{f},\breve{g})\ldots \\
\mathbf{p}(y_2|y_1,\breve{f},\breve{g})\mathbf{p}(y_1|\breve{f},\breve{g}),
\end{aligned}$$

one easily obtains

$$\begin{aligned}
L(y^+,\breve{f},\breve{g}) = -\log[\mathbf{p}(y^+,\breve{f},\breve{g})] &= \frac{\breve{f}^T\breve{K}_f^{-1}\breve{f}}{2\lambda_f^2} + \frac{\breve{g}^T\breve{K}_g^{-1}\breve{g}}{2\lambda_g^2} \\
&+ \frac{1}{2}\log\det(2\pi\lambda_f^2\breve{K}_1) + \frac{1}{2}\log\det(2\pi\lambda_g^2\breve{K}_2) \\
&+ \frac{1}{2}\log\det(2\pi\sigma^2 I_n) + \frac{\|y^+ - \breve{A}\breve{f} - \breve{B}\breve{g}\|^2}{2\sigma^2},
\end{aligned} \tag{59}$$

where the existence of $\breve{K}_f^{-1}$ and $\breve{K}_g^{-1}$ derives from the fact that the stable spline kernel is strictly positive; see Pillonetto and De Nicolao (2010) for details. Notice that, for known y^+, $\breve{f}$ and $\breve{g}$ are Gaussian, and that the Hessian of L with respect to $\breve{f}$ and $\breve{g}$ is

$$\partial^2_{\breve{f},\breve{g}} L(y^+,\cdot,\cdot) = \mathrm{diag}\{\lambda_1^{-2}\breve{K}_1^{-1}, \lambda_2^{-2}\breve{K}_2^{-1}\} + \sigma^{-2}[\breve{A}\breve{B}]^T[\breve{A}\breve{B}]. \tag{60}$$

Hence, the minimum variance estimates of $\breve{f}$ and $\breve{g}$ are

$$\begin{aligned}
\begin{pmatrix}\breve{f}^{MV}\\ \breve{g}^{MV}\end{pmatrix} &= \sigma^{-2}\left[\partial^2_{\breve{f},\breve{g}} L(y^+,.,.)\right]^{-1}[\breve{A}\breve{B}]^T y^+ \\
&= \begin{pmatrix}\lambda_f^2\breve{K}_f\breve{A}^T \\ \lambda_g^2\breve{K}_g\breve{B}^T\end{pmatrix}(\breve{V}[y^+])^{-1}y^+,
\end{aligned} \tag{61}$$

where the last equality exploits the matrix inversion lemma (Anderson & Moore, 1979). Finally, (36) and (37) are obtained by letting q tend to ∞ in (61) and noticing that in this way $\breve{V}[y^+]$, $\breve{K}_f\breve{A}^T$ and $\breve{K}_g\breve{B}^T$ tend to $V[y^+]$, K_fA^T and K_gB^T, respectively, and that all these limits are a.s. well defined under the same assumptions as in Lemma 7[6].

Let us now prove the correspondence with Tikhonov regularization in RKHS illustrated in (35). First, recall that an RKHS on X is the Hilbert space of functions which are the completion of the manifolds given by all the finite linear combinations $\sum_{i=1}^{l} m_iK(\cdot,t_i)$ for all choices of $l \in \mathbb{N}$, $m_i \in \mathbb{R}$, and $t_i \in X$, with inner product

$$< \sum_i m_iK(\cdot,t_i), \sum_j n_jK(\cdot,s_j) >_{\mathscr{H}} = \sum_{i,j} m_in_jK(t_i,s_j). \tag{62}$$

Now, we define an RKHS $\mathscr{H}$ of objects in $\mathbb{R}^{\infty\times 2}$ whose kernel $Q : (\mathbb{N}\times\{1,2\})\times(\mathbb{N}\times\{1,2\}) \mapsto \mathbb{R}$ is defined as follows:

$$[Q]_{i1,j1} = \lambda_f^2K_f(i,j), \tag{63}$$

$$[Q]_{i2,j2} = \lambda_g^2K_g(i,j), \tag{64}$$

$$[Q]_{i1,j2} = [Q]_{i2,j1} = 0, \quad i,j \in \mathbb{N}. \tag{65}$$

If $h = [h_1\ h_2]$, with $h_1, h_2 \in \mathbb{R}^\infty$, in view of (62)–(65), it holds that

$$\|h\|^2_{\mathscr{H}} = \frac{\|h_1\|^2_{\mathscr{H}_f}}{\lambda_f^2} + \frac{\|h_2\|^2_{\mathscr{H}_g}}{\lambda_g^2}.$$

Problem (35) can now be rewritten as follows:

$$h^{MV} = \arg\min_{h\in\mathscr{H}} \frac{\|y^+ - Ch\|^2}{\sigma^2} + \|h\|^2_{\mathscr{H}}, \tag{66}$$

where $C : \mathscr{H} \mapsto \mathbb{R}^n$, from Lemma 7, is a linear and (a.s.) continuous operator defined for any h by $Ah_1 + Bh_2$. Now, the same rationale as in the proof of Theorem 1.3.1 in Wahba (1990) can be followed to show that the solution is a linear combination of the representers of the n linear functionals, defined in our case by the rows of C. Recalling that the representer of a linear and continuous functional acting on an RKHS can be obtained by applying it to the reproducing kernel, we conclude that h^{MV} belongs to the subspace generated by the columns of QC^T. Replacing h and $\|h\|^2_{\mathscr{H}}$ in (66) with QC^Tc and c^TCQC^Tc, respectively, we obtain that c must satisfy

$$(CQC^T + \sigma^2 I_n)c = y^+.$$

Using (63)–(65), and the definition of C, (36) and (37) are obtained, and this completes the proof. □

Proof of Proposition 4. Given the function h on a finite-dimensional domain, whose Hessian matrix $\partial_x^2 h(x)$ is constant and positive definite, Laplace's method provides the following expression for calculating exponential integrals:

$$\int_{-\infty}^{+\infty} \mathrm{e}^{-h(x)}\mathrm{d}x = \det\left[\frac{\partial_x^2 h}{2\pi}\right]^{-1/2} \mathrm{e}^{-h(\hat{x})}, \tag{67}$$

where $\hat{x}$ minimizes $h(x)$ with respect to x; see e.g. De Bruijn (1961). Then, if (56) holds, using (67), we obtain

$$\begin{aligned}
-\log[\mathbf{p}(y^+)] = \frac{1}{2}\log\det\left(\frac{\partial^2_{\breve{f},\breve{g}} L(y^+,.,.)}{2\pi}\right) \\
+ L(y^+,\breve{f}^{MV},\breve{g}^{MV}),
\end{aligned} \tag{68}$$

[6] The proof uses the same arguments used in the proof of Lemma 7 and is therefore omitted.

with L given by (59), $\partial^2_{\check{f},\check{g}} L$ by (60), and $\check{f}^{MV}, \check{g}^{MV}$ by (61). Using Lemma 19 in Bell and Pillonetto (2004), one has

$$\begin{aligned}&\frac{1}{2}\log\det\left(\frac{\partial^2_{\check{f},\check{g}} L(y^+,.,.)}{2\pi}\right)+\frac{1}{2}\log\det(2\pi\lambda_f^2\check{K}_f)\\&\quad+\frac{1}{2}\log\det(2\pi\lambda_g^2\check{K}_g)+\frac{1}{2}\log\det(2\pi\sigma^2 I_n)\\&\quad=\frac{1}{2}\log\det(2\pi\check{V}[y^+]),\end{aligned}\tag{69}$$

while, after simple computations which exploit the equality

$$\begin{aligned}&I_n-\lambda_f^2\check{A}\check{K}_f\check{A}^T(\check{V}[y^+])^{-1}-\lambda_g^2\check{B}\check{K}_g\check{B}^T(\check{V}[y^+])^{-1}\\&\quad=\sigma^2(\check{V}[y^+])^{-1},\end{aligned}$$

we obtain

$$\begin{aligned}&\frac{(\check{f}^{MV})^T\check{K}_f^{-1}\check{f}^{MV}}{2\lambda_f^2}+\frac{(\check{g}^{MV})^T\check{K}_g^{-1}\check{g}^{MV}}{2\lambda_g^2}+\frac{\|y^+-\check{A}\check{f}-\check{B}\check{g}\|^2}{2\sigma^2}\\&\quad=\frac{1}{2}(y^+)^T(\check{V}[y^+])^{-1}y^+.\end{aligned}\tag{70}$$

This allows us to conclude that

$$-\log[\mathbf{p}(y^+)]=\frac{1}{2}\ln(\det[2\pi\check{V}[y^+]])+\frac{1}{2}(y^+)^T(\check{V}[y^+])^{-1}y^+.\tag{71}$$

By letting q tend to ∞, $\check{V}[y^+]$ tends to $V[y^+]$ in (71), where, as above, the limit is (a.s.) well defined under the same assumptions as in Lemma 7, and (39) is finally obtained. □

Remark 8. Compared with (39), (68) is especially efficient if $q \ll n$, since it allows one to compute the marginal likelihood by inverting matrices of size $q \times q$. On the other hand, using (39) may lead to a more numerically robust method, since it can be easily seen that the matrix (60) is generally more ill conditioned than (57).

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723.

Anderson, B. D. O., & Moore, J. B. (1979). *Optimal filtering*. Englewood Cliffs, NJ, USA: Prentice-Hall.

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society*, *68*, 337–404.

Barry, D. (1986). Nonparametric Bayesian regression. *The Annals of Statistics*, *14*, 934–953.

Bauer, D. (2001). Order estimation for subspace methods. *Automatica*, *37*, 1561–1573.

Bell, B. M., & Pillonetto, G. (2004). Estimating parameters and stochastic functions of one variable using nonlinear measurements models. *Inverse Problems*, *20*(3), 627–646.

Berrut, J. P., & Trefethen, L. N. (2004). Barycentric Lagrange interpolation. *SIAM Review*, *46*, 501–517.

Bertero, M. (1989). Linear inverse and ill-posed problems. *Advances in Electronics and Electron Physics*, *75*, 1–120.

Chiuso, A. (2007). The role of vector autoregressive modeling in predictor based subspace identification. *Automatica*, *43*(6), 1034–1048.

Chiuso, A., Pillonetto, G., & De Nicolao, G. (2008). Subspace identification using predictor estimation via Gaussian regression. In *Proceedings of 2008 conf. on decision and control*. Cancun, Mexico.

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

De Bruijn, N. G. (1961). *Asymptotic methods in analysis*. North-Holland.

De Nicolao, G., & Ferrari Trecate, G. (2001). Regularization networks: fast weight calculation via Kalman filtering. *IEEE Transactions on Neural Networks*, *12*, 228–235.

Desai, U. B., & Pal, D. (1984). A realization approach to stochastic model reduction. *IEEE Transactions on Automatic Control*, *29*, 1097–1100.

Doan, T., Litterman, R., & Sims, C. A. (1984). Forecasting and conditional projection using realistic prior distributions. *Econometric Reviews*, *3*, 1–100.

Goethals, I., Pelckmans, K., Suykens, J. A. K., & De Moor, B. (2005). Subspace identification of Hammerstein systems using least squares support vector machines. *IEEE Transactions on Automatic Control*, *50*, 1509–1519.

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

Hannan, E. J. (1980). The estimation of the order of an ARMA process. *The Annals of Statistics*, *8*, 1071–1081.

Ha Quang, M., Pillonetto, G., & Chiuso, A. (2009). Nonlinear system identification via Gaussian regression and mixtures of kernels. In *Proceedings of the 15th IFAC symposium on system identification, SYSID 2009*. Saint-Malo, France.

Harvey, A. C. (1989). *Forecasting structural time series models and the Kalman filter*. Cambridge University Press.

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

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.

Kitagawa, G., & Gersch, W. (1996). *Smoothness priors analysis of time series*. Springer-Verlag.

Kitagawa, G., & Gersh, H. (1984). A smoothness priors state space modeling of time series with trends and seasonalities. *Journal of the American Statistical Association*, *79*(386), 378–389.

Kitagawa, G., & Gersh, H. (1985). A smoothness priors long AR model method for spectral estimation. *IEEE Transactions on Automatic Control*, *30*(1), 57–65.

Leeb, H., & Pötscher, B. (2005). Model selection and inference: facts and fiction. *Econometric Theory*, *21*, 21–59.

Ljung, L. (1999). *System identification — theory for the user*. Prentice Hall.

Ljung, L. (2007). *System identification toolbox V7.1 for MATLAB*. Natick, MA: The MathWorks, Inc.

Loève, M. (1963). Probability theory. Van Nostrand Reinhold.

Maritz, J. S., & Lwin, T. (1989). *Empirical Bayes method*. Chapman and Hall.

McVinish, R.S., Braslavsky, J.H., & Mengersen, K.L. (2006). A Bayesian-decision theoretic approach to model error modeling. In *Proceedings of the 14th IFAC symposium on system identification, SYSID 2006*, (pp. 1015–1020) Newcastle, Australia.

Milanese, M., & Vicino, A. (1991). Optimal estimation theory for dynamic systems with set membership uncertainty: an overview. *Automatica*, *27*(6), 997–1009.

Neve, M., De Nicolao, G., & Marchesi, L. (2007). Nonparametric identification of population models via Gaussian processes. *Automatica*, *97*(7), 1134–1144.

Pedregal, D., Taylor, J., Tych, W., & Young, P. (2009). Captain toolbox for MATLAB. CRES, Lancaster University. http://www.es.lancs.ac.uk/cres/captain/.

Pernebo, L., & Silverman, L. M. (1982). Model reduction via balanced state space representations. *IEEE Transactions Automatic Control*, *27*, 382–387.

Pillonetto, G., & Bell, B. M. (2007). Bayes and empirical Bayes semi-blind deconvolution using eigenfunctions of a prior covariance. *Automatica*, *43*(10), 1698–1712.

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., Chierici, M., & Cobelli, C. (2009). Fast algorithms for nonparametric population modeling of large data sets. *Automatica*, *45*(1), 173–179.

Pillonetto, G., & Saccomani, M. P. (2006). Input estimation in nonlinear dynamic systems using differential algebra concepts. *Automatica*, *42*, 2117–2129.

Poggio, T., & Girosi, F. (1990). Networks for approximation and learning. *Proceedings of the IEEE*, *78*, 1481–1497.

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

Runge, C. (1901). Uber empirische funktionen und die interpolation zwischen aquidistanten ordinaten. *Zeitschrift fr Mathematik und Physik*, *46*, 224–243.

Schoukens, J., Suykens, J., & Ljung, L. (2009). Benchmark on nonlinear identification. http://wwwtw.vub.ac.be/elec/sysid09.htm.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461–464.

Smale, S., & Zhou, D. X. (2007). Learning theory estimates via integral operators and their approximations. *Constructive Approximation*, *26*, 153–172.

Smola, A. J., & Schölkopf, B. (2003). Bayesian kernel methods. In S. Mendelson, & A. J. Smola (Eds.), *Machine learning, proceedings of the summer school, Australian national university* (pp. 65–117). Berlin, Germany: Springer-Verlag.

Soderstrom, T., & Stoica, P. (1989). *System identification*. Prentice Hall.

Tikhonov, A. N., & Arsenin, V. Y. (1977). *Solutions of ill-posed problems*. Washington, DC: Winston, Wiley.

Wahba, G. (1990). *Spline models for observational data*. Philadelphia: SIAM.

Weinert, H. L. (2001). *Fixed-interval smoothing for state space models*. Kluwer.

Yang, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation. *Biometrika*, *92*(4), 937–950.

Young, P. C. (2000). Recursive estimation, forecasting and adaptive control. In *Control and Dynamic Systems* (pp. 119–166). San Diego: Academic Press.

Young, P. C., McKenna, P., & Bruun, J. (2001). Identification of nonlinear stochastic systems by state dependent parameter estimation. *International Journal of Control*, *74*, 1837–1857.

Young, P. C., & Pedregal, D. J. (1999). Recursive and en-bloc approaches to signal extraction. *Journal of Applied Statistics*, *26*, 103–128.

Gianluigi Pillonetto was born on 21 January 1975 in Montebelluna (TV), Italy. He received his Doctoral degree in Computer Science Engineering cum laude from the University of Padova in 1998 and his Ph.D. degree in Bioengineering from the Polytechnic of Milan in 2002. In 2000 and 2002 he was a visiting scholar and visiting scientist, respectively, at the Applied Physics Laboratory, University of Washington, Seattle. From 2002 to 2005 he was a 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, stochastic systems, deconvolution problems, nonparametric regularization techniques, learning theory and randomized algorithms.

Alessandro Chiuso is an Associate Professor in the Department of Management and Engineering, University of Padova, Italy.

He received his "Laurea" degree summa cum laude in Telecommunication Engineering from the University of Padova in July 1996 and his Ph.D. degree in System Engineering from the University of Bologna in 2000. He has held visiting positions with Washington University in St. Louis (USA), KTH (Sweden) and UCLA (USA).

Dr. Chiuso serves or has served as a member of several conference program committees and technical committees. He is an Associate Editor of IEEE Trans. on Automatic Control (from 2010), Automatica (from 2008), European Journal of Control (from 2011) and a member of the editorial board of IET Control Theory and Application (from 2007). He has also been an Associate Editor of the IEEE Conference Editorial Board (2004–2009).

His research interest are mainly in eEstimation, identification theory and applications. Further information can be found at his personal web page http://automatica.dei.unipd.it/people/chiuso.html.

Giuseppe De Nicolao was born in Padova, Italy, in 1962. In 1986, he received his Laurea (master degree) cum laude in Electronic Engineering 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", Milan. In 1988, he joined the Italian National Research Council (C.N.R.) as a researcher scientist at the Center of System Theory in Milan. From 1992 to 2000, he was an Associate Professor and, since 2000, he has been a 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 was a visiting fellow at the Department of Systems Engineering of the Australian National University, Canberra. He was a keynote speaker at the workshop on "Nonlinear model predictive control: Assessment and future directions for research", 1998, Ascona, Switzerland. He is a Senior Member of the IEEE. From 1999 to 2001, he was an Associate Editor of the IEEE Transactions on Automatic Control and, since 2007, he has been an 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. On these subjects, he has authored or coauthored more than 100 journal papers, and he is a coinventor of two patents.