A new kernel-based approach for linear system identification

Gianluigi Pillonetto [a], Giuseppe De Nicolao [b]

[a] *Department of Information Engineering*
University of Padova, Padova (Italy)

[b] *Dipartimento di Informatica e Sistemistica*
University of Pavia, Pavia (Italy)

Abstract

This paper describes a new kernel-based approach for linear system identification of stable systems. We model the impulse response as the realization of a Gaussian process whose statistics, differently from previously adopted priors, include information not only on smoothness but also on BIBO-stability. The associated autocovariance defines what we call a stable spline kernel. The corresponding minimum-variance estimate belongs to a reproducing kernel Hilbert space which is spectrally characterized. Compared to parametric identification techniques, the impulse response of the system is searched for within an infinite-dimensional space, dense in the space of continuous functions. Overparametrization is avoided by tuning few hyperparameters via marginal likelihood maximization. The proposed approach may prove particularly useful in the context of robust identification in order to obtain reduced order models by exploiting a two-step procedure that projects the nonparametric estimate onto the space of nominal models. The continuous time derivation immediately extends to the discrete time case. On several continuous and discrete time benchmarks taken from the literature the proposed approach compares very favorably with existing parametric and nonparametric techniques.

Key words: linear system identification; kernel-based methods; Bayesian estimation; regularization; Gaussian processes; robust identification; stochastic embedding

1 Introduction

We consider the problem of identifying the impulse response of a BIBO stable linear and time-invariant system, fed with a known input, from noisy and discrete output measurements. The usual identification approaches use finite-dimensional parametric models, whose order has to be relatively low when identification is motivated by robust-control design purposes [23,36,40]. The standard methods to select the "best" model order rely on complexity criteria such as Akaike (AIC), Generalized Cross Validation (GCV) or Minimum Description Length (MDL). In general, not only the obtained nominal model will be uncertain due to the variance of its estimated parameters but it will also be biased due to undermodeling. Robust identification has to do with the joint assessment of variance and bias affecting the estimated nominal model. It is well known that, in presence of undermodeling, the form of the input signal may significantly affect the estimate and the reliability of the model in frequencies relevant to the intended use. Prefiltering of input and output data is often adopted as a remedy even if the choice of the operating frequency range may be nontrivial [50].
In order to characterize the variance and bias error, robust identification has been developed along three main directions. Two of them, namely stochastic embedding [14,15] and model-error modeling [22,41], share a probabilistic background, see also [16] for another statistical approach. The third approach, namely set-membership identification [12,13,28,29], relies on a deterministic worst-case paradigm [27]. The starting point of all these methods is the identification of a low-order nominal model by standard techniques such as maximum likelihood or prediction error methods. The subsequent step is the assessment of bias and variance for the nominal model. The stochastic embedding approach models the bias error as the realization of a stochastic process, e.g. white noise with decreasing variance over the time domain [15] or a random walk over the frequency domain [14]. The model-error modeling approach exploits residual analysis in order to characterize undermodel-

[1] This paper was not presented at any IFAC meeting. Corresponding author Gianluigi Pillonetto Ph. +390498277607

Email addresses: giapi@dei.unipd.it (Gianluigi Pillonetto), giuseppe.denicolao@unipv.it (Giuseppe De Nicolao).

ing, whereas set-membership identification determines the worst-case error associated with the nominal model [36]. A schematic representation of the identification scheme common to the three approaches is reported in Fig. 1a. After a possible prefiltering phase, the data are passed to a parameter estimation module which yields the nominal model. The nominal model and the data are then processed by a model-error estimation module in order to quantify the bias and variance error.

Even if stochastic embedding has some connection with Bayesian estimation, only few contributions are available so far, see e.g. [18,19]. Differently from previous contributions, in this paper the probabilistic prior is formulated directly on the unknown impulse response, rather than on the bias error, and the impulse response is assumed to be the realization of a Gaussian process. As such, it belongs to an infinite-dimensional function space [4,35,39,51]. The prior is defined as an integrated Wiener process over a suitable transformation of the time-axis. Such prior prevents overfitting and accounts for continuity and nonstationarity of the impulse response. Moreover, information on BIBO-stability is incorporated within the prior, whose realizations are proven to be almost surely BIBO-stable. Connections with Tikhonov-type regularization [6,34,45,46] and Reproducing Kernel Hilbert Spaces (RKHS) [3,9,49] are extensively discussed. Among other things, it is shown that the estimate belongs to a space which is dense in the space of continuous functions.

If a low-order model is desired for some specific use, a two-step procedure can be adopted. First, a low-bias nonparametric estimate of the impulse response is computed by the proposed method. Then, the desired parametric model is obtained by projecting the nonparametric estimate onto a suitable low-order finite-dimensional space. A formal proof of optimality of this two-step procedure is also given (Proposition 4). It is worth noting that such result translates to the Bayesian context Hyalmarsson's advice "always first model as well as possible" based on the invariance/separation principle, see Section 4.2 in [17]. The new robust identification scheme is schematically illustrated in Fig. 1b, where the output of the nonparametric estimator is fed into a projection module yielding the nominal model and its uncertainty. Compared to Fig. 1a, note that prefiltering is no more needed. Even if schemes (a) and (b) in Fig. 1 share the same common objective of finding a low-order model suitable for robust control, there is a substantial difference between them. In fact, the former yields a low order model whose amount and type of bias depend on the experimental design, e.g. choice of the input. For instance, if the system is excited by a low frequency input, bias will be concentrated at high frequencies. The second procedure, conversely, first uses all the available information, i.e. data and prior knowledge on impulse response, to obtain the best possible estimate. Then, the subsequent projection step is not directly affected by experimental design conditions.

The paper is organized as follows. In Section 2, the identification problem is formulated and regression via Gaussian processes in RKHS is concisely overviewed. In Section 3, it is shown how to obtain a nominal model by projecting the Bayes estimate onto a finite-dimensional space. In Section 4, we propose a new Gaussian prior for system identification by defining a suitable Mercer kernel K. In Section 5, a spectral characterization of K is provided. It is also shown that realizations from the new prior are almost surely associated with BIBO-stable systems and that the RKHS defined by K is dense in the space of continuous functions. In Section 6, we use simulated benchmarks taken from the literature to demonstrate the effectiveness of the proposed approach. Conclusions end the paper. Proofs are gathered in the appendix.

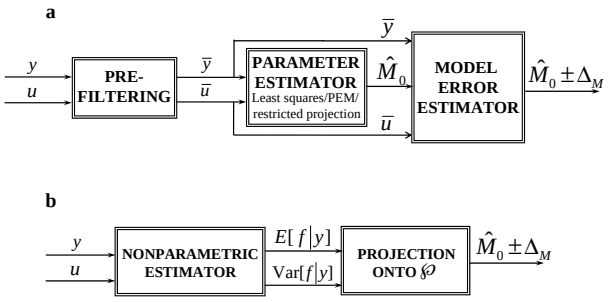

Fig. 1. *(a)* Identification scheme common to stochastic embedding, model-error modeling and set membership approaches. Notation u indicates the system input while output measurements are y, with their filtered versions denoted by $\bar{u}$ and $\bar{y}$, respectively. $\hat{M}_0$ denotes the estimate of the nominal model while Δ_M is uncertainty associated with bias and variance error *(b)* New identification scheme proposed in this paper. $E[f|y]$ and $Var[f|y]$ denote the posterior mean and autocovariance of the impulse response, respectively, while $\mathcal{P}$ is the space of nominal models.

2 Preliminaries

We are given a finite set of noisy data sampled from the output of a continuous-time linear dynamic system fed with a known input $u(t)$. We will mainly refer to such continuous-time setting even if the proposed approach can deal with discrete-time problems just by replacing integral operators with suitable discrete convolutions. In the sequel, f represents the unknown impulse response and $\mathbf{N}(\mu, \Sigma)$ denotes a Gaussian density of mean μ and covariance matrix Σ. Let $q(t)$ denote the noiseless output

defined as follows

$$q(t) \doteq L_t^u[f] = \int_0^t f(t-\tau)u(\tau)d\tau, \quad t \in D \qquad (1)$$

where D is an interval on the real line. The associated measurement model is

$$y_i = L_{t_i}^u[f] + v_i, \quad i = 1, \cdots, n \qquad (2)$$

where t_i, $i = 1, \ldots, n$ are the sampling instants while the errors v_i, $i = 1, \ldots, n$, are independently distributed with $v_i \sim \mathbf{N}(0, \sigma^2)$ $\forall i$. In the sequel, $y \doteq [y_1 \quad y_2 \ldots y_n]^T$, $v \doteq [v_1 \quad v_2 \ldots v_n]^T$ and the shorthand notation L_i^u is used in place of $L_{t_i}^u$.
Adopting a Bayesian viewpoint, it is assumed that there exists a prior for f which consists of a Gaussian measure in an infinite-dimensional function space. In particular, $\tilde{f}$ denotes a zero-mean Gaussian process with autocovariance $cov(f(t_i), f(t_j)) = \lambda^2 K(t_i, t_j)$. Here, λ^2 is a possibly unknown scale factor while K represents a Mercer kernel, i.e. a mapping $K : D \times D \mapsto \mathbf{R}$ which is continuous, symmetric and positive definite.Let I_n be the $n \times n$ identity matrix. The statistical model for f reads as follows

$$\begin{aligned} f(t) &= \sum_{i=1}^{d} \theta_i \psi_i(t) + \tilde{f}(t), \qquad t \in D \\ \theta &\sim \mathbf{N}(0, \rho I_d), \qquad \rho \to +\infty \end{aligned} \qquad (3)$$

where θ is independent of $\tilde{f}$ and v, while $\{\psi_i\}_{i=1}^d$ are assigned functions that account for components on whose amplitude no prior knowledge is assumed. In the sequel, $\mathcal{B}$ will denote the subspace spanned by $\{\psi_i\}$. A careful choice of $\mathcal{B}$ helps minimizing the bias when estimating certain classes of impulse responses, e.g. those with a dominant pole. For what concerns linear system identification, this will be extensively discussed in section 4.
Since f and v are assumed jointly normal, the posterior of f given y is Gaussian as well. Our target estimate is the minimum variance estimate of f, i.e. the posterior mean $E[f|y]$. To define such estimate in rigorous mathematical terms, it is useful to recall that a Mercer Kernel K can be associated with a unique RKHS $\mathcal{H}$, with norm $\|\cdot\|_{\mathcal{H}}$, containing scalar continuous functions on D, see [3,49] for details. If the dimension of $\mathcal{H}$ is infinite, it can be proven that realizations from $\tilde{f}$ do not fall in $\mathcal{H}$ with probability one [26,48]. Nevertheless, the following result points out that, for known y, the minimum variance estimate of f belongs to the direct sum of $\mathcal{H}$ and $\mathcal{B}$ (denoted as $\mathcal{H} \oplus \mathcal{B}$) and can be obtained as the solution of a Tikhonov-type variational problem. Below, and in the sequel, it is assumed that $L_i^u : \mathcal{H} \mapsto \mathbf{R}$ is continuous $\forall i$.

Proposition 1 *[48] Assume that f is independent of v. Let $\boldsymbol{P}[g]$ denote the orthogonal projection of g onto $\mathcal{H}$, in $\mathcal{H} \oplus \mathcal{B}$ and let also $\gamma = \sigma^2/\lambda^2$. For known y and γ, the minimum variance estimate of f is given by*

$$\hat{f} = \arg \min_{g \in \mathcal{H} \oplus \mathcal{B}} \quad \sum_{i=1}^{n} \left(y_i - L_i^u[g]\right)^2 + \gamma \|\boldsymbol{P}[g]\|_{\mathcal{H}}^2 \qquad (4)$$

Remark 2 *The above proposition states the duality between Gaussian processes and RKHS [48,26], which will be further exploited in the sequel. In particular, we will use f to denote Gaussian processes, with the abbreviated notation f_t often used in place of $f(t)$, while g will indicate deterministic functions and $\dot{g}$ the corresponding first derivative.*

In (4), besides the choice of $\mathcal{B}$, also K and γ will greatly influence the quality of the estimate. The former reflects our prior knowledge about f and will determine fundamental properties of $\mathcal{H}$ such as its capability of approximating a wide class of functions. The latter is the so-called regularization parameter that controls the balance between expected regularity of the solution and adherence to experimental data (the so called bias/variance trade off). The main contribution of the present paper is the suggestion of a specific choice of $\mathcal{B}$ and K for linear system identification such that, by a proper tuning of λ^2 and σ^2 (and hence γ), the solution of (4) has favorable bias and variance properties.
As far as K is concerned, typical choices are Gaussian or polynomial kernels [39]. In particular, when the signal is just known to be smooth, the most popular approach is to model f as an integrated Wiener process with completely unknown initial conditions. Under these statistical assumptions, one has that [30]

$$W(s,\tau) \doteq Cov(\tilde{f}(s), \tilde{f}(\tau)) = \begin{cases} \frac{s^2}{2}\left(\tau - \frac{s}{3}\right) & s \leq \tau \\ \frac{\tau^2}{2}\left(s - \frac{\tau}{3}\right) & s > \tau \end{cases} \qquad (5)$$

This kernel underlies also the Bayesian interpretation of cubic smoothing splines [48]. For the subsequent derivation, it is useful to focus on the cubic spline kernel $W(s,\tau)$ defined over the domain $S \times S$ where $S = [0, 1]$. Since the RKHS $\mathcal{H}_W$ associated with the kernel W is a Sobolev space of functions g with $g(0) = \dot{g}(0) = 0$ [1,7], it is convenient to select ψ_1 and ψ_2 as a constant and a linear function, respectively. In this way, $\theta \in \mathbf{R}^2$ and

$$\mathcal{B}_W = \text{span}\{1, t\} \qquad t \in S \qquad (6)$$

In the practical application of Gaussian regression, a hierarchical approach is adopted. Few high level parameters (called *hyperparameters*), e.g. λ^2 and σ^2, are regarded as fixed and deterministic in order to obtain closed form formulas for the estimate $\hat{f}$. According to the so-called Empirical Bayes method, the tuning of the hyperparameters grounds on statistical criteria based on

the stochastic interpretation underlying Problem (4), as described in section 4.

3 Mean-square optimal finite-dimensional approximation

In this section $\mathcal{L}$ indicates the set of all functions [1] mapping D into the real line, with generic element denoted by g. In our context $\mathcal{L}$ will represent the set of all possible models while $\mathcal{P} \subset \mathcal{L}$ will be used to represent the set of nominal models. For example, $\mathcal{P}$ may contain all the first-order approximations of our dynamic system, i.e.

$$\mathcal{P} = \left\{ g : g(t) = Ae^{-at}, A \in \mathbf{R}, a \in \mathbf{R}^+, t \in \mathbf{R}^+ \right\}$$

Let Γ be an operator mapping the observation vector y into functions g, i.e. $\Gamma : \mathbf{R}^n \mapsto \mathcal{L}$. Furthermore, we use $\Gamma_t : \mathbf{R}^n \mapsto \mathbf{R}$ to represent $\Gamma(y)$ evaluated at t, i.e. if $\Gamma : y \mapsto g$ then $\Gamma_t : y \mapsto g(t)$, $t \in D$. Finally, $\mathbf{w}(t)$, $t \in D$, is a strictly positive weighting function.
The next two results do not require Gaussianity of f.

Proposition 3 *Let $\hat{\Gamma}^B$ satisfy*

$$\hat{\Gamma}^B \doteq \arg\min_{\Gamma} \int_D \boldsymbol{E}[(f_t - \Gamma_t(y))^2 | y] \boldsymbol{w}(t) dt \qquad \forall y$$

Then,

$$\hat{\Gamma}^B_t(y) = \boldsymbol{E}[f_t | y] = \int_{\mathbf{R}} f_t \boldsymbol{p}_t(f_t | y) df_t \qquad \forall y$$

where $\boldsymbol{p}_t(f_t, y)$ is the joint density of f_t and y.

It is worth remarking that the above result shows that when there is no restriction on the range of Γ, the optimal estimate does not depend on the weighting function $\mathbf{w}(t)$. Let instead $\Gamma^{\mathcal{P}}$ be an operator that maps vectors y into functions $g \in \mathcal{P}$, i.e. $\Gamma^{\mathcal{P}} : \mathbf{R}^n \mapsto \mathcal{P}$. The next result shows that the optimal estimate of f_t restricted to $\mathcal{P}$ is given by a projection, weighted by $\mathbf{w}$, of the Bayes estimate onto the set $\mathcal{P}$ of nominal models. The result is an extension of that obtained in [52] where f is restricted to be a Gaussian process.

Proposition 4 *Let*

$$\hat{\Gamma}^{\mathcal{P}} \doteq \arg\min_{\Gamma^{\mathcal{P}}} \int_D \boldsymbol{E}[(f_t - \Gamma^{\mathcal{P}}_t(y))^2 | y] \boldsymbol{w}(t) dt \qquad \forall y$$

Then,

$$\hat{\Gamma}^{\mathcal{P}}(y) = \arg\min_{g \in \mathcal{P}} \int_D \left(\hat{\Gamma}^B_t(y) - g(t) \right)^2 \boldsymbol{w}(t) dt \qquad \forall y \tag{7}$$

[1] Here, and in the sequel, Lebesgue measurability is implicitly assumed.

The above proposition provides a simple way to approximate the impulse response within a desired finite-dimensional space. It states that the mean squared error is minimized by looking for the approximating function that best fits the Bayes estimate $\hat{\Gamma}^B$, thus suggesting a two-stage procedure, i.e. regularized Bayesian estimation followed by projection onto the finite-dimensional space. It is worth noting that the projection step is just a continuous least squares problem. The weighting function $\mathbf{w}(t)$ can be used to specify where a more accurate approximation is needed. The use of frequency weighting, e.g. to obtain a low frequency approximation, is also easily implementable.

4 System identification using a new Gaussian prior

4.1 Modeling the unknown impulse response

Regularization methods which rely upon the kernel W defined in (5) are widely employed in nonparametric function estimation, see e.g. the extensive literature on cubic spline regression [48,47,44,43]. However, this kernel is not suitable to reconstruct the impulse response of a stable dynamic system because of the following limitations:

- The Tikhonov estimator (4), with $\mathcal{H}_W \oplus \mathcal{B}_W$ defining our hypothesis space, coincides with the cubic smoothing spline estimator [48]. As such, it is is able to fit straight lines without bias. However, in system identification one would better obtain unbiased estimates of exponentials on the noncompact domain $X = [0, +\infty)$.
- The variance of the process associated with kernel W increases over time. But, for stable systems, *a priori* impulse response uncertainty is likely to decrease with time. In particular, a prior is needed on X which incorporates the BIBO-stability constraint.

The following definition will prove useful in the derivation of a new prior specifically suited for system identification.

Definition 5 *A prior on f preserves a family of functions $\mathcal{F}$ if there exists $\overline{n}$ such that, for any distinct times $t_1, \ldots, t_n$, $n \geq \overline{n}$, it holds that*

$$\boldsymbol{E}[f | f(t_1) = g(t_1), \ldots, f(t_n) = g(t_n)] = g, \quad \forall g \in \mathcal{F}$$

For instance, if a prior preserves lines, this means that, given sampled observations of the unknown function, the Bayes estimator projects lines onto themselves. In other words, the estimate draws all information on the linear trend from the data without biasing the estimate towards prior knowledge. It is well known that the Wiener

prior associated with linear splines preserves constant functions whereas the integrated Wiener prior associated with cubic splines preserves lines. However, when estimating impulse responses, it is convenient to adopt a prior that preserves exponentials. Below, we will introduce a mapping which converts X into the unit interval $S = [0, 1]$ such that the prior which preserves exponentials in the old coordinates preserves straight lines in the new ones. In other words, the time-transformation maps an exponential, with rate constant β, into a straight line. It will be also shown that impulse response stability is guaranteed by imposing that in the new coordinates the function value at zero is null (Proposition 10).
A prior on S enjoying all the desired features is the integrated Wiener process with zero initial value and arbitrary first-order derivative at zero. Summarizing, the desired time-transformation is

$$\tau = e^{-\beta t} \qquad t \in X$$

In the original coordinates, the prior for the unknown impulse response is thus defined as follows

$$f(t) = \begin{cases} 0 & \text{if } t < 0 \\ \theta e^{-\beta t} + \tilde{f}(t) & \text{if } t \in X \end{cases} \tag{8}$$

where $\theta \sim \mathbf{N}(0, \infty)$ and $\tilde{f}(t)$, independent of θ, is a zero-mean Gaussian process with autocovariance

$$Cov(\tilde{f}(s), \tilde{f}(t)) \doteq \lambda^2 K(s, t; \beta) \qquad (s, t) \in X \times X \tag{9}$$

where

$$K(s, t; \beta) \doteq W(e^{-\beta s}, e^{-\beta t}) \qquad (s, t) \in X \times X \tag{10}$$

Finally,

$$\mathcal{B}_K = \text{span}\{e^{-\beta t}\} \qquad t \in X \tag{11}$$

The kernel K will be hereafter named "stable spline kernel", given its connection with the cubic spline kernel and its intrinsic ability, when coupled with the bias space $\mathcal{B}_K$, to preserve a family of stable exponential functions.

Remark 6 *For the sake of simplicity, we will restrict our attention to a bias space which is the span of a single exponential. However, in principle, $\mathcal{B}_K$ could be easily extended to include the span of two or more exponential functions, although as demonstrated in the example section, even the simple model (11) performs very satisfactorily in a variety of situations.*

Finally, when dealing with discrete-time systems, the model for f becomes the sampled version of (8), i.e. for $k \in Z$ we have

$$f(k) = \begin{cases} 0 & \text{if } k < 0 \\ \theta e^{-\beta k} + \tilde{f}(k) & \text{for } k = 0, 1, 2, 3, ... \end{cases} \tag{12}$$

4.2 Estimating hyper-parameters and impulse response

The impulse response estimate is provided by the Tikhonov estimator (4) with hypothesis space $\mathcal{H} \oplus \mathcal{B}$ replaced by $\mathcal{H}_K \oplus \mathcal{B}_K$. However, such estimator requires the knowledge of the hyperparameter vector $\xi = [\lambda, \beta, \sigma]$.
According to the empirical Bayes approach, ξ is obtained by maximizing the marginal likelihood, i.e. the probability of y obtained by integrating out f from the joint probability of y and f. In the following, we give formulas for the computation of the log marginal likelihood. For this purpose, define

$$C(\xi) \doteq (L_1^u[h] \dots L_n^u[h])^T, \quad h = e^{-\beta s}$$
$$M(\xi) \doteq Var[y|\theta, \xi]$$

Note that the (i, j)-entry of M is

$$M(\xi)|_{i,j} = \lambda^2 L_i^u L_j^u [K(\cdot, \cdot; \xi)] + \sigma^2 \delta_{ij} \tag{13}$$

with δ_{ij} the Kronecker delta. In (13), $L_i^u L_j^u[K]$ means that L_j^u is first applied to $K(\cdot, \cdot)$ as a function of one of its arguments. This leads to a well defined function in $\mathcal{H}_K$ to which the second functional is applied. Ambiguity is avoided by the symmetry of the kernel.
In the following, dependence of C and M on ξ is sometimes omitted to simplify the notation. If $\theta \sim \mathbf{N}(0, \rho)$, using Lemma 19 in [5] we have

$$\begin{aligned} \det(Var[y|\xi]) &= \det(M + \rho C C^T) \\ &= \rho \det(M)(\rho^{-1} + C^T M^{-1} C) \end{aligned} \tag{14}$$

When $\theta \sim \mathbf{N}(0, \infty)$, one has

$$\begin{aligned} b(\xi) &\doteq \lim_{\rho \to \infty} \ln\left(\det(Var[y|\xi])\right) - \ln(\rho) \\ &= \ln\left(\det(M)\right) + \ln(C^T M^{-1} C) \end{aligned} \tag{15}$$

In addition, using eq. (1.5.12) in [48] we also have

$$\begin{aligned} A(\xi) &\doteq \lim_{\rho \to \infty} (Var[y|\xi])^{-1} \\ &= M^{-1}(I_n - C(C^T M^{-1} C)^{-1} C^T M^{-1}) \end{aligned} \tag{16}$$

Using (15,16), we obtain the following optimization problem

$$\hat{\xi} = \arg\min_{\xi} J(y; \xi) \tag{17}$$

where the cost function

$$J(y; \xi) = \frac{1}{2} b(\xi) + \frac{1}{2} y^T A(\xi) y \tag{18}$$

is equal to the opposite of the asymptotic log-marginal likelihood apart from terms independent of ξ. According to the empirical Bayes approach the estimate of f

is obtained through the Tikhonov estimator (4) with hyperparameters ξ replaced by their maximum likelihood estimate $\hat{\xi}$. Explicit formulas for the solution of (4) with the stable spline kernel are reported below, in eqs. (20,21,22). They are obtained using the so called representer theorem, see e.g. section 1.1.2 in [49] or also the proof of Theorem 1.5.3 in [48]. The new identification algorithm reads as follows.

Nonparametric system identification algorithm
The input to this algorithm includes the input and output sequences $\{u_k\}$ and $\{y_k\}$. The output of this algorithm is the estimate $\hat{f}$ of the impulse response of the system.

- Determine the estimate of the hyperparameter vector ξ and θ as follows

$$\hat{\xi} = \arg\min_{\xi} J(y;\xi) \tag{19}$$

$$\hat{\theta} = \frac{C(\hat{\xi})^T M(\hat{\xi})^{-1} y}{C(\hat{\xi})^T M(\hat{\xi})^{-1} C(\hat{\xi})} \tag{20}$$

where

$$\begin{aligned} J(y;\xi) &= \frac{1}{2} b(\xi) + \frac{1}{2} y^T A(\xi) y \\ b(\xi) &= \ln\left(\det(M)\right) + \ln(C^T M^{-1} C) \\ A(\xi) &= M^{-1}(I_n - C(C^T M^{-1} C)^{-1} C^T M^{-1}) \\ C(\xi) &= (L_1^u[h] \dots L_n^u[h])^T, \quad h = e^{-\beta s} \\ M(\xi)|_{i,j} &= \lambda^2 L_i^u L_j^u \left[K(\cdot,\cdot;\xi)\right] + \sigma^2 \delta_{ij} \end{aligned}$$

- Calculate the estimate of the system impulse response according to the formula

$$\hat{f}(t) = \hat{\theta} e^{-\hat{\beta} t} + \hat{\lambda}^2 \sum_{i=1}^{n} c_i L_i^u \left[K(\cdot, t; \hat{\beta})\right] \tag{21}$$

where $\{c_i\}$ are the elements of vector $c \in \mathbf{R}^n$ given by

$$c = (M(\hat{\xi}))^{-1}(y - C(\hat{\xi})\hat{\theta}) \tag{22}$$

Needless to say, in a discrete-time context the same approach can be followed provided that integral operators are replaced by their discrete counterparts.

4.3 Computing confidence intervals

Assume that hyper-parameters λ, β and σ are known or set to their maximum likelihood estimates. Our first aim is to compute the autocovariance of the noiseless output q, conditional on y, sampled on an arbitrarily dense grid $\Omega = \{s_i\}_{i=1}^N$ which contains the sampling grid $\{t_i\}_{i=1}^n$.

By omitting the dependence on ξ, to simplify notation, the noiseless output $q(t)$ can be written as

$$q(t) = \theta a(t) + b(t)$$

where

$$\begin{aligned} a(t) &:= L_t^u[e^{-\beta(\cdot)}], \qquad b(t) := L_t^u[\tilde{f}(\cdot)] \\ Cov[b(s), b(t)] &= \lambda^2 L_s^u L_t^u \left[K(\cdot,\cdot)\right] \end{aligned}$$

Given a function $g(t)$, its sampled version on Ω is

$$g_\Omega = [g(s_1) \quad g(s_2) \dots g(s_N)]^T$$

This allows us to write

$$q_\Omega = [a_\Omega \quad I_N] \begin{bmatrix} \theta \\ b_\Omega \end{bmatrix} \doteq Q \begin{bmatrix} \theta \\ b_\Omega \end{bmatrix}$$

$$y = P \begin{bmatrix} \theta \\ b_\Omega \end{bmatrix} + v$$

where $P \in \mathbf{R}^{n \times (N+1)}$ is obtained from Q by keeping only rows associated with actual output observations in y. Using standard properties of Gaussian random variables, see e.g. [2], we finally obtain

$$Var\left[q_\Omega | y\right] = Q\left(\sigma^{-2} P^T P + V\right)^{-1} Q^T$$

where, since the prior variance of θ is infinite,

$$V \doteq \left(Var \begin{bmatrix} \theta \\ b_\Omega \end{bmatrix}\right)^{-1} = \begin{pmatrix} 0 & 0_{1\times N} \\ 0_{N\times 1} & \left(Var\left[b_\Omega\right]\right)^{-1} \end{pmatrix}$$

with $0_{1\times N}$ the $1 \times N$ matrix with zero entries. [2]
Once the posterior autocovariance of q has been computed, confidence intervals for linear transformations of q can be easily obtained. For instance, when confidence intervals over the frequency domain are needed, let $F(j\omega)$ denote the Fourier transform of f with $Re[F(j\omega)]$ and $Im[F(j\omega)]$ indicating its real and imaginary part, respectively. The problem amounts now to computing $Var\left[(Re[F(j\omega)] \quad Im[F(j\omega)])^T | y\right]$ for any given ω. Letting

$$\mathcal{F}_R^\omega : q \mapsto Re[F(j\omega)], \qquad \mathcal{F}_I^\omega : q \mapsto Im[F(j\omega)]$$

[2] Existence of the inverse of matrix $\sigma^{-2} P^T P + V$ can be established by the same arguments as in the proof of Proposition 4 in [30]

(corresponding to composition of the Fourier transform with the inverse of L_t^u) one has

$$Var\begin{bmatrix} Re[F(j\omega)] \\ Im[F(j\omega)] \end{bmatrix} |y\Bigg] = \begin{pmatrix} F_R^\omega F_R^\omega \left[K_q^y(\cdot,\cdot)\right] & F_R^\omega F_I^\omega \left[K_q^y(\cdot,\cdot)\right] \\ F_R^\omega F_I^\omega \left[K_q^y(\cdot,\cdot)\right] & F_I^\omega F_I^\omega \left[K_q^y(\cdot,\cdot)\right] \end{pmatrix}$$

where

$$K_q^y(s,t) \doteq Cov\left[q(s), q(t)|y\right]$$

5 Stable spline kernel: spectral analysis

In this section, we report a complete spectral analysis of the stable spline kernel K defined by (10). The scope of the section is twofold. First, it is shown that realizations drawn from the new prior are almost surely the impulse response of a BIBO-stable system (Proposition 10). Second, a spectral characterization of the RKHS associated with the stable spline kernel K is derived (Proposition 11). We start with some definitions and a proposition which can be derived from results contained in [9,11,31].

Definition 7 *Define the sequence $\{\lambda_i\}$, with $\lambda_{i+1} \leq \lambda_i$, as*

$$\lambda_i = (1/\alpha_i)^4 \quad i = 1, 2, ... \tag{23}$$

where α_i denotes the solution of

$$1/\cosh(\alpha) + \cos(\alpha) = 0 \tag{24}$$

which is closest to $(i - 1/2)\pi$.
In addition, define functions $\{\phi_i\}$ and $\{\rho_i\}$ as follows

$$\begin{aligned} \phi_i(t;\alpha_i) &= C_1(\alpha_i)\cos(\alpha_i t) + C_2(\alpha_i)\sin(\alpha_i t) \\ &\quad + C_3(\alpha_i)e^{-\alpha_i(1-t)} + C_4(\alpha_i)e^{-\alpha_i t} \qquad t \in S \end{aligned} \tag{25}$$

$$\rho_i(\tau;\alpha_i) = \phi_i(e^{-\beta\tau};\alpha_i) \quad \tau \in X \tag{26}$$

where $\{C_k\}$ are scalars satisfying

$$\begin{aligned} C_4(\alpha) &= \Big(\int_S [C_1(1)\cos(\alpha t) + C_2(1)\sin(\alpha t) \\ &\quad + C_3(1)e^{-\alpha(1-t)} + e^{-\alpha t}]^2 dt\Big)^{-1/2} \\ C_3(C_4) &= C_4(\alpha)\left[\frac{2}{1+e^{-2\alpha}} - 1\right]/\sin(\alpha) \\ C_2(C_4) &= C_4(\alpha) - C_3(C_4)e^{-\alpha} \\ C_1(C_4) &= -C_4(\alpha) - C_3(C_4)e^{-\alpha} \end{aligned}$$

Below, $\mathbf{L}^2(S)$ denotes the classical Lebesgue space on S equipped with the inner product $< \cdot, \cdot >_2$.

Proposition 8 *[9] Let W be defined by (5). Then, it holds that*

$$\begin{aligned} < \phi_j, \phi_k >_2 &= \begin{cases} 1 \ \textit{if } j = k \\ 0 \ \textit{otherwise} \end{cases} \\ \lambda_j \phi_j(s) &= \int_S W(s,t)\phi_j(t)dt \\ W(s,t) &= \sum_{j=1}^{\infty} \lambda_j \phi_j(s)\phi_j(t) \end{aligned}$$

where the above sum converges uniformly with respect to $(s,t) \in S \times S$. In addition, letting $\mathcal{H}_W$ be the RKHS associated with the cubic spline kernel W

$$\mathcal{H}_W = \left\{ g \in \mathbf{L}^2(S) \quad | \quad g = \sum_{j=1}^{\infty} a_j \phi_j, \ \sum_{j=1}^{\infty} \frac{a_j^2}{\lambda_j} < \infty \right\} \tag{27}$$

Hereafter, $\mathbf{L}_\nu^2(X)$ is used to indicate the space of square integrable functions on X where the (probability) measure ν admits the density $\beta e^{-\beta t}$ ($\beta > 0$ and $t \geq 0$) with respect to Lebesgue measure. The inner product on $\mathbf{L}_\nu^2(X)$ is denoted as $< ., . >_\nu$.

Proposition 9 *The integral operator on $\mathbf{L}_\nu^2(X)$ associated with the kernel K in (10) and defined by*

$$\int_X K(x,\tau)f(\tau)d\nu(\tau) \qquad x \in X$$

is a bounded, compact and positive trace-class (nuclear) integral operator mapping $\mathbf{L}_\nu^2(X)$ into $C(X)$. We also have

$$< \rho_j, \rho_k >_\nu = \begin{cases} 1 \ \textit{if } j = k \\ 0 \ \textit{otherwise} \end{cases} \tag{28}$$

$$\lambda_j \rho_j(s) = \int_X K(s,t)\rho_j(t)d\nu(t) \tag{29}$$

$$K(s,t) = \sum_{j=1}^{\infty} \lambda_j \rho_j(s)\rho_j(t) \tag{30}$$

where $\{\rho_j\}$ are defined by (26) and the sum above converges uniformly with respect to $(s,t) \in X_1 \times X_2$, X_1 and X_2 being any compact subset of X.

The next result highlights the nature of the proposed prior on the impulse response of the system.

Proposition 10 *Let $\mathbf{L}^p(X)$ denote the classical Lebesgue space of p-power integrable functions on X. Let $\tilde{f}(t)$, with $t \in X$, be a zero-mean Gaussian process with stable spline autocovariance K. Then, realizations*

from $\tilde{f}(t)$ belong to $\mathbf{L}^p(X)$, with $p \geq 1$, almost surely, i.e. realizations from $\tilde{f}(t)$ are almost surely the impulse response of a BIBO linear system.

Recalling (8), it is immediate to see that stability with probability one of realizations of $\tilde{f}(t)$ implies that of realizations of $f(t)$.
As already mentioned, the optimal estimate given the data belongs to $\mathcal{H}_K \oplus \mathcal{B}_K$. The next proposition characterizes such hypothesis space showing that within the RKHS $\mathcal{H}_K$ any continuous-time impulse response can be approximated arbitrarily well in the uniform topology.

Proposition 11 *It holds that*

$$\mathcal{H}_K = \left\{ g \in \mathbf{L}^2_\nu(X) \quad | \quad g = \sum_{j=1}^{\infty} a_j \rho_j, \sum_{j=1}^{\infty} \frac{a_j^2}{\lambda_j} < \infty \right\} \tag{31}$$

Further, $\mathcal{H}_K$ is dense in the space of continuous functions defined on any compact subset of X, i.e. given any continuous function g on the compact $X_1 \subset X$ and any scalar $\epsilon > 0$, there exists $g_\epsilon \in \mathcal{H}_K$ such that

$$\sup_{\tau \in X_1} |g(\tau) - g_\epsilon(\tau)| < \epsilon$$

The eigenfunctions associated with some of the largest eigenvalues of $\mathcal{H}_W$ and $\mathcal{H}_K$ (with β set to 1) are displayed in Fig. 2. They give an interesting insight into the nature of the hypothesis space chosen for system identification. In fact the unknown impulse response is seen as the linear combination of eigenfunctions through independent weights with decreasing variance.

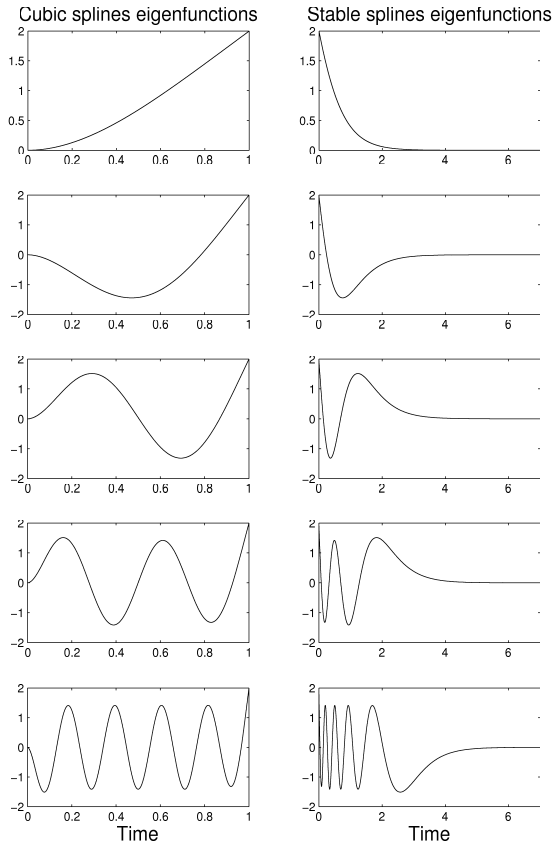

Fig. 2. Eigenfunctions $\{\phi_j\}$ of the standard cubic spline kernel W (left) and eigenfunctions $\{\rho_j\}$ of the novel stable spline kernel K (right) for j=1,2,3,5,10

6 Examples

6.1 Discrete-time test functions

The proposed nonparametric identification scheme is first applied to the identification of discrete-time dynamic systems from noisy output data. In particular, as a benchmark we consider 5 simulated impulse responses displayed in the left (and right) panels of Fig. 3 (solid line). They are listed below, where all, but the third one, are given in the z-transform domain

$$F_1(z) = \frac{0.0355z^2 + 0.02465z}{z^3 - 1.273z + 0.333}$$
$$F_2(z) = \frac{0.36z}{5(z^2 + 0.24 + 0.36)}$$
$$f_3(k) = e^{-\frac{k^2}{100}}/\sqrt{2\pi}, \quad k = 1, 2, \ldots$$
$$F_4(z) = \frac{0.01z^4 + 0.0074z^3 + 0.000924z^2 - 0.000017642z}{z^5 - 2.14z^4 + 1.5549z^3 - 0.4387z^2 + 0.042025z}$$
$$F_5(z) = \frac{z^3 + 0.5z^2}{z^4 - 2.2z^3 + 2.42z^2 - 1.87z + 0.7225}$$

The first two represent second-order systems taken from [15], while the third one is proportional to a normal density with support only on the positive axis. The last two impulse responses are a fifth- and a fourth-order model, taken from Example 5.1 in [50] and Section 8.6

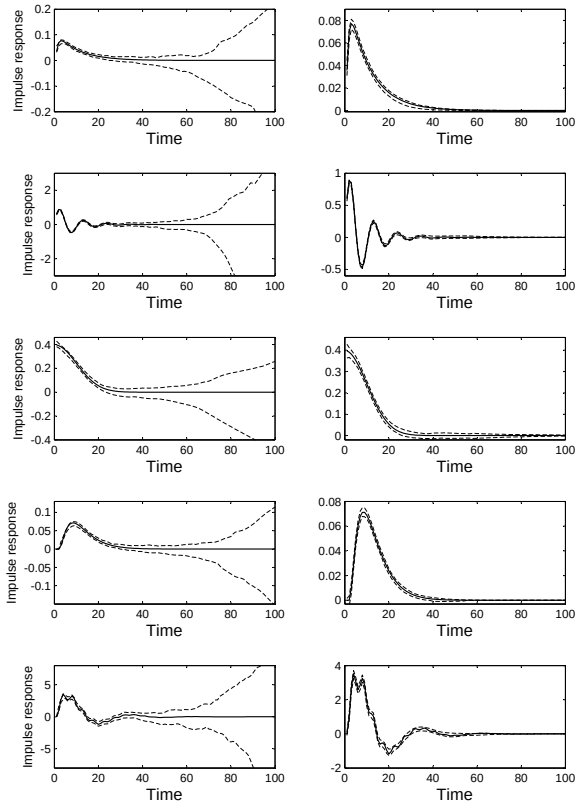

Fig. 3. Discrete-time test functions (Section 6.1). Results from Monte Carlo simulation using white noise as system input: true impulse response (solid line) and 99% variability bands of the estimates (dashed lines) obtained modeling the unknown function using the classical cubic spline kernel W (left) and the stable spline kernel K (right)

of [22], respectively. The system input is white noise of unit intensity. System identification has to be performed starting from 100 output noisy samples. In particular, system initial conditions are null at instant 0 and the forcing input is applied starting from instant 1. Measurement noise is white and normal with standard deviation set to 5% of the maximum absolute value of the generated noiseless output samples. Measurements are collected at instants $k = 1, 2, \ldots, 100$.

We consider 5 Monte Carlo (MC) studies, one for any test function, consisting of 300 runs with independent noise realizations. The prior model of the system impulse response is the sampled version of either the cubic spline prior with unknown initial conditions or the new stable spline prior. The number of reconstructed impulse response coefficients is equal to 100. Noise standard deviation σ and parameters λ, β are unknown and estimated from data.

In the left panels of Fig. 3, results obtained by using the cubic spline kernel W are depicted (left column). The true function (solid line) and the 99% variability bands (dashed lines) of the 300 estimates are visible. It is apparent that variability bands are rather wide. Reconstructed curves suffer from oscillations in the final part of the experiment because the prior model does not include asymptotic information on system stability. In the right panels of Fig. 3 we display results obtained by exploiting the stable spline kernel K (right column). In addition to the improved quality of the estimates, variability bands are much narrower and always close to the true function. Comparing these results with those reported in Section 7 of [15], one can notice that the second-order impulse responses are much better estimated. Furthermore, we have used far fewer output measurements (100 in place of 1000). In particular, the proposed regularization method removes the oscillations, due to ill-conditioning, which instead affect the estimates reported in Fig. 5, 6 and 7 of [15].

Given an estimate of f obtained at the j-th run, namely $\hat{f}_j$, the reconstruction error and the average reconstruction error are denoted by err_j and Err, respectively, and defined by

$$err_j = \sum_{k=1}^{\infty} \sqrt{(f_j(k) - \hat{f}_j(k))^2}, \quad Err = \frac{\sum_{j=1}^{300} err_j}{300} \tag{32}$$

In addition to the cubic and stable spline methods, Table 1 compares Err values obtained in the same MC studies described above by two other nonparametric approaches and two parametric ones. In particular, employed estimators are

(1) *empirical transfer function estimation* (ETFE) as implemented in the *etfe.m* function of the MATLAB System Identification Toolbox [24]. The smoothing parameter is chosen by an "oracle", i.e. setting the value to that minimizing Err at any monte Carlo study (this is an ideal tuning yielding a lower bound on the realistically achievable performance)

(2) regularized impulse response estimation using the standard cubic spline kernel W with hyperparameters tuned via maximum likelihood (already discussed)

(3) the same with a Gaussian kernel G defined by (see e.g. [35])

$$G(j,k) = \lambda^2 e^{\frac{(j-k)^2}{\varpi^2}}, \quad j,k = 1,2,\ldots$$

with λ and the kernel width ϖ estimated via maximum likelihood

MC study	$ETFE + oracle$	W	G	K	$PEM + AIC$	$PEM + oracle$
#1	3.4e-2	17e-2	4e-2	0.82e-2	1.9e-2	0.47e-2
#2	2.7e-1	24e-1	4.3e-1	1.3e-1	2.3e-1	0.38e-1
#3	17e-2	31e-2	24e-2	4e-2	15e-2	3.5e-2
#4	2e-2	12.1e-2	6.1e-2	0.67e-2	2.5e-2	0.53e-2
#5	1.6	9.1	1.9	0.73	1.1	0.3

Table 1
Discrete-time test functions (Section 6.1). Results from Monte Carlo simulation using white noise as system input: Err using ETFE with oracle (first column), classical cubic spline kernel W (second column), the Gaussian kernel (third column), the new stable spline kernel K (fourth column), PEM with Akaike (fifth column) and with oracle (sixth column).

MC study	$ETFE + oracle$	W	G	K	$PEM + AIC$	$PEM + oracle$
#1	11e-2	5.4e-2	5.5e-2	2.4e-2	24e-2	0.86e-2
#2	6.2e-1	5.5e-1	6.1e-1	3.6e-1	21e-1	0.8e-1
#3	53e-2	8.4e-2	8.7e-2	4.5e-2	25.5e-2	6e-2
#4	1.8e-1	1.3e-1	0.46e-1	0.12e-1	9.7e-1	0.1e-1
#5	4.4	2.1	2.1	1.5	11.4	1.3

Table 2
Discrete-time test functions (Section 6.1). Results from Monte Carlo simulation using square wave as system input: Err using ETFE with oracle (first column), classical cubic spline kernel W (second column), the Gaussian kernel (third column), the new stable spline kernel K (fourth column), PEM with Akaike (fifth column) and with oracle (sixth column).

(4) the same with the new stable spline kernel K (already discussed)
(5) the classical *prediction error method* (PEM) as implemented in the oe.m function of the MATLAB System Identification Toolbox [24]. Model orders $\hat{m}_1$ and $\hat{m}_2$ of the two polynomials defining the output error structure are chosen by the Akaike criterion (AIC), i.e.

$$(\hat{m}_1, \hat{m}_2) = \arg \min_{m_1 \in M, m_2 \in M} 2(m_1 + m_2) + n \ln[RSS(m_1, m_2)] \quad (33)$$

where $n = 100$, $M = \{1, 2, \ldots, 15\}$ and RSS is the residual sum of squares. The latter is computed using the predicted output of the estimated model obtained by the *predict.m* MATLAB function.
(6) the same with model order chosen by the oracle which minimizes Err obtainable by PEM

It is seen that the stable spline kernel outperforms all approaches but PEM+oracle with respect to which it performs almost as well. Table 2 is similar to Table 1 except that system input for identification is a square wave which alternates between levels 1 and 0, with period 10. It is apparent that the new nonparametric approach still outperforms the other nonparametric approaches and PEM+AIC while is only marginally worse than PEM+oracle.
In Table 3, we give the root mean square error obtained by applying the stable spline kernel K on reduced sampling grids (20, 40, 60 and 80 samples randomly chosen from the original 100 ones) with data generated using white noise as system input. Again, 300 MC runs for any subsampled schedule were performed. It is apparent that, even under these reduced sampling schedules, the impulse responses are accurately reconstructed. It is worth remarking that standard nonparametric spectral approaches like ETFE cannot handle nonuniform sampling schedules, which are routinely adopted in some fields, e.g. biomedical modeling.

MC study	K (20)	K (40)	K (60)	K (80)	K (100)
#1	2.8e-2	1.3e-2	1.1e-2	9.5e-3	8.2e-3
#2	3.8e-1	2.2e-1	1.7e-1	1.5e-1	1.3e-1
#3	1.07e-2	6.7e-2	5.6e-2	5e-2	4e-2
#4	1.5e-2	1e-2	8.3e-3	7.3e-3	6.7e-3
#5	1.9	1.08	8.3e-1	7.5e-1	7.3e-1

Table 3
Dicrete-time test functions (Section 6.1). Results from Monte Carlo simulation using white noise as system input: Err using the new stable spline kernel K with reduced and full sampling grids (number of samples are within brackets).

6.2 Randomly generated discrete-time test functions

In this subsection, we consider a more probing simulated study where, at any of the 300 runs, a discrete-time system of order 30 is randomly generated. In particular, the coefficients of the numerator of the transfer function are realizations of white noise with variance 4. The denominator is instead generated by using the MATLAB function *drmodel.m* with system poles constrained to lie inside the circle of radius 0.9.
System is at rest at instant 0 and the forcing input is white noise of unit variance. Measurement noise is white and Gaussian with standard deviation set to 10% of the maximum absolute value of the generated noiseless output samples. The identification data set consists of 150 output measurements collected at instants $1, 2, \ldots, 150$. In this case, given an estimate $\hat{f}_j$ obtained at the j-th run, it is useful to define the relative error

$$err_j = \sqrt{\frac{\sum_{k=1}^{\infty}(f_j(k) - \hat{f}_j(k))^2}{\sum_{k=1}^{\infty} f_j^2(k)}} \tag{34}$$

and Err as in (32). Employed estimators are

(1) regularized estimation of the first 100 impulse response coefficients using the the new stable spline kernel K. Hyperparameters are tuned via maximum likelihood.
(2) PEM with model order of the two polynomials defining the output error structure chosen by AIC with $M = 1, 2, \ldots, 35$ and $m_1 = m_2$ in (33)
(3) the same with model order chosen by BIC
(4) the same with model order chosen by the oracle which minimizes Err obtainable by PEM

At any Monte Carlo run j, we also computed the 95% confidence interval around the nonparametric estimate (see subsection (4.3)) and let χ_j indicate the fraction of samples of $\{f(k)\}_{k=1}^{100}$ that belong to such interval.

Table 4 displays Err values. Remarkably, the proposed nonparametric estimator outperforms PEM equipped with AIC and BIC. Moreover, its performance is very close to that of PEM equipped with the oracle. In addition, the average value for χ_j is 0.937, indicating that confidence intervals obtained from the nonparametric estimator are highly informative under these experimental conditions.

6.3 Other model selection examples

We now consider a discrete time second-order system with frequency response $F(z)$ given by

$$F(z) = \frac{2(z-0.3)^2}{5(z^2 - 1.4z + 0.65)} \tag{35}$$

K	$PEM + AIC$	$PEM + BIC$	$PEM + oracle$
0.23	0.35	0.32	0.21

Table 4
Randomly generated discrete-time test functions (Section 6.2). Results from Monte Carlo simulation: Err using the new stable spline kernel K (first column), PEM with AIC (third column), PEM with BIC (third column) and with oracle (fourth column).

As in [21], the real part of the two complex poles of the system is 0.7 and the problem consists of reconstructing f using a step function as input applied to the system at rest. In particular, estimation has to be performed from 40 noisy measurements corrupted by a noise with standard deviation $\sigma = 0.04$ which is assumed unknown. For the sake of comparison, we will also consider identification of f by means of finite Laguerre expansions, i.e.

$$F(z, \eta) = \sum_{k=1}^{m} \eta_k L_k(z), \quad L_k(z) = \frac{\sqrt{1-p^2}}{z-p}\left(\frac{1-pz}{z-p}\right)^{k-1}$$

where value for p is either 0 (corresponding to FIR models) or is optimally chosen and set to 0.7.
We perform 10 MC studies, each consisting of 300 runs with independent realizations of the noise. The studies use

(1) least-squares estimation of the Laguerre coefficients with $p = 0$ and model order m chosen by AIC with maximum allowed value equal to 15
(2) the same with model order chosen by BIC
(3) the same with model order chosen by an oracle which minimizes the reconstruction error Err defined in (32)
(4) the same except that $p = 0.7$ and model order is chosen by AIC
(5) the same except that $p = 0.7$ and model order is chosen by BIC
(6) the same except that $p = 0.7$ and model order is chosen by an oracle
(7) regularized impulse response estimation using the stable spline kernel K
(8) PEM with model order chosen by AIC, as described in (33) but with $M = \{1, \ldots, 6\}$
(9) the same with model order chosen by BIC
(10) the same with model order chosen by an oracle

In Fig. 4, box-plots of the errors achieved by the 10 estimators are shown. Remarkably, the proposed nonparametric approach outperforms AIC- and BIC-based estimators also when basis functions include knowledge on pole position and when PEM is used. Furthermore, results are better than those obtained by combining an oracle and FIR models and are close to those achieved by PEM+oracle and by combining the oracle with setting p to the optimal value 0.7.
Estimation of Laguerre coefficients by least-squares

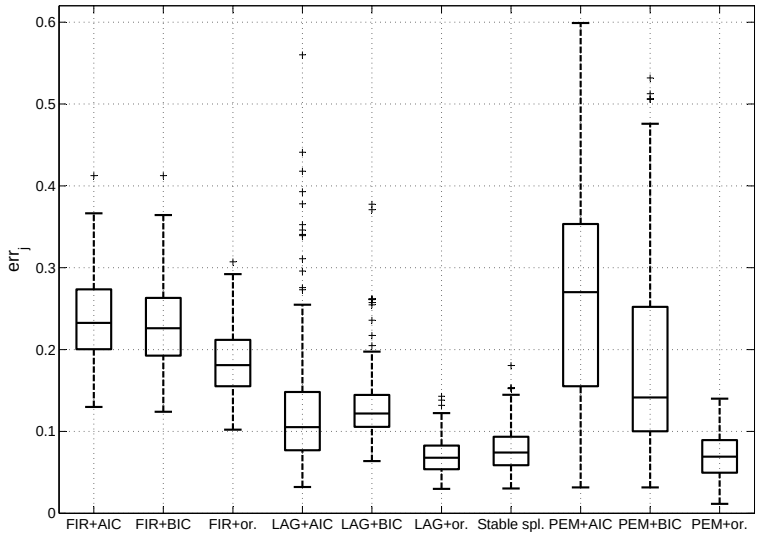

Fig. 4. Model selection example (Section 6.3). Boxplots of errors err_j (see eq. (32)) relative to the 10 estimators used to reconstruct the impulse response of eq. (35)

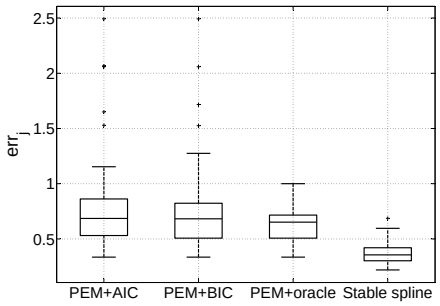

Fig. 5. Model selection example (Section 6.3). Boxplots of errors err_j (see eq. (32)) relative to the 4 estimators used to reconstruct the Runge function reported in eq. (36)

is not robust in this case because is subject to ill-conditioning. This problem is exacerbated when FIR models are used since they do not include any information regarding regularity of the impulse response. Conversely, when Laguerre polynomials are optimally chosen, smoothness information on f is included in the model. However, AIC- and BIC-based model selection is not satisfactory. At first sight, it may seem that the new stable spline estimator differs from parametric ones only in the choice of the basis functions (see Fig. 2). As a matter of fact, the difference is more substantial. In fact, the basis functions are not fixed but are adapted to the specific data set through the tuning of the hyperparameters. Moreover, the coefficients of the basis functions are not found by regression but rather through regularization which dampens high frequency basis functions. Seen in another way, model complexity is controlled by the regularization parameter γ while β encodes information on stability.

To further illustrate flexibility of stable spline basis functions, let us consider the reconstruction of an infinite-dimensional system whose impulse response is a translated and scaled version of the well known Runge function [37]

$$f(k) = \left(1 + 25\left(\frac{k-20}{20}\right)^2\right)^{-1}, \quad k = 1, 2, \dots \quad (36)$$

The system has to be reconstructed from 100 noisy measurements using a step as input to the system which is initially at rest. Noise standard deviation is 2% of the maximum absolute value of the generated noiseless output samples. We perform 4 MC studies, each consisting of 300 runs, where the following estimators are used

(1) PEM with model order chosen by AIC, as described in (33) with $M = \{1, \dots, 15\}$
(2) the same with model order chosen by BIC
(3) the same with model order chosen by an oracle
(4) regularized impulse response estimation using the stable spline kernel K

Fig. 5 displays boxplots of errors err_j as defined in (32). In this case, the oracle performs worse than the nonparametric estimator. As a matter of fact, Err values are 0.63 and 0.37 using the oracle and the stable spline kernel, re-

spectively, while those obtainable using PEM+AIC and PEM+BIC are similar and around 0.8.

6.4 First-order low-frequency approximation of a continuous-time second order system

Consider a continuous-time second-order system whose frequency response $F(s)$ is given by

$$F(s) = \frac{5s + 15}{s^2 + 21s + 20}$$

The impulse response is displayed in the top (and bottom) left panels of Fig. 6 (thick line) while the Bode plot of the magnitude is displayed in the top (and bottom) panel of Fig. 7 (thick line). In Fig. 6, we plot 200 noisy output samples generated by using as input either a comb function with noise standard deviation equal to 0.08 (top right panel) or a step function with $\sigma = 0.02$ (bottom right panel). Suppose now that for control purposes it is desirable to achieve a first-order approximation of the system for use at low frequencies. In the left panels of Fig. 6 we plot the estimates of the impulse response obtained by fitting a first-order model to data via least squares (dashed lines) while the corresponding Bode plots are visible in Fig. 7 (dashed lines). One can see that the result obtained by using the comb function is very inaccurate at low frequencies. This result could be improved by resorting to pre-filtering methods but this would require a careful choice of the bandwidth. In the left panels of Fig. 6 and in Fig. 7 the estimates obtained by the new nonparametric approach proposed in this paper are shown (thin lines). One can notice that the estimate is less sensitive to the type of system input due to the infinite-dimensional nature of the stable spline hypothesis space. In particular, it closely approximates the true magnitude plot over a wide frequency range. The desired lower order model can be derived from the regularized estimate via Proposition 4. For instance, Fig. 7 plots the magnitude plot of a first-order model obtained by projecting the nonparametric estimate onto a first-order model using a weighting function which, over the frequency domain, is constant on $[0, 1]$ rad/sec and 0 elsewhere (dash-dot line). Finally, in Fig. 8, we display the true magnitude and phase and the nonparametric estimates together with 99% confidence intervals (dashed lines). Given this information, it is possible to obtain the robustness margin Δ_M to be used for control design, see e.g. [15]. In the simplest case, if a symmetric deterministic error bound around the nominal model is desired, it suffices to take the smallest Δ_M such that both the lower and upper confidence limits of the nonparametric estimate (with prescribed confidence level, e.g. 1%) are encompassed.

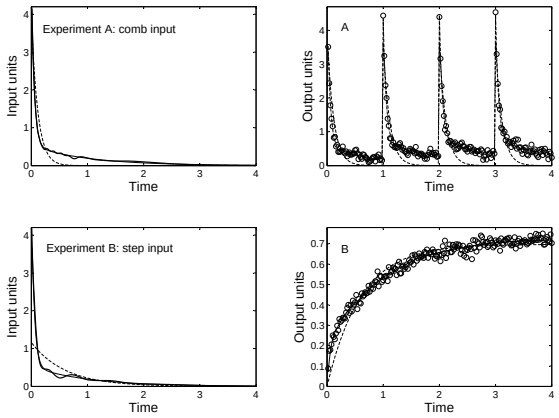

Fig. 6. Continuous-time second-order system (Section 6.4). Left: true impulse response (thick line), estimated impulse response obtained by fitting a first-order model to data (dashed lines) and by the new nonparametric approach (continuous line). Right: noisy output samples and reconstructed output. System input is a comb (top panels) or a step function (bottom panels).

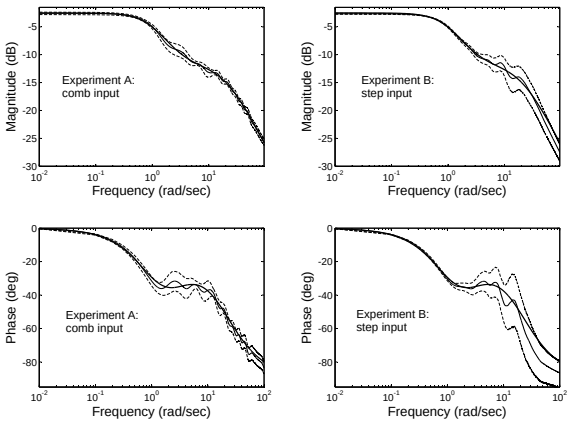

Fig. 8. Continuous-time second-order system (Section 6.4). *Top:* True magnitude Bode plot (thick line), nonparametric estimate (continuous line) and 99% confidence intervals (dashed lines). *Bottom:* True phase Bode plot (thick line), nonparametric estimate (continuous line) and 99% confidence intervals (dashed lines). System input is a comb (left panels) or a step (right panels).

7 Conclusions

Methods which are currently used for robust identification start with a low-order nominal model identified by standard techniques such as least-squares and prediction error methods. Then, on the basis of the nominal model, bias and variance errors are quantified. In this paper, we have embedded this problem in a fully

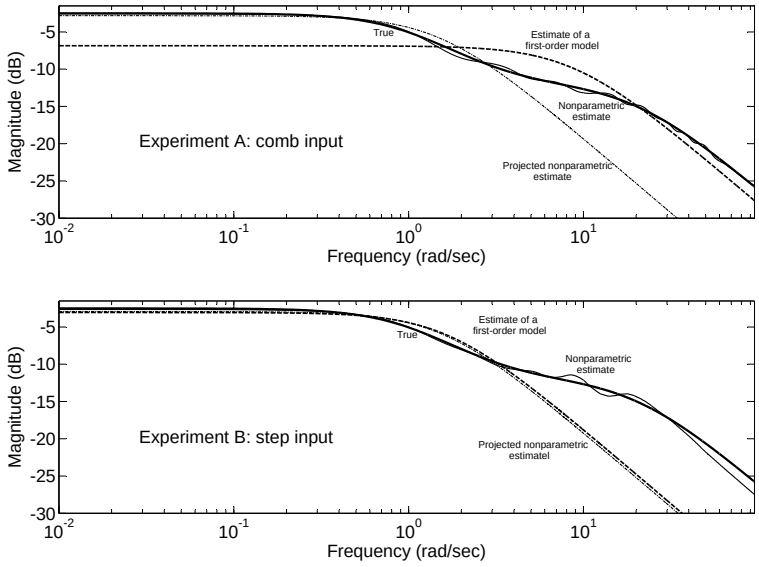

Fig. 7. Continuous-time second-order system (Section 6.4). True magnitude (thick line), estimated magnitude obtained by fitting a first-order model to data (dashed lines), using the new nonparametric approach (continuous line), and projecting the regularized estimate onto a first-order model (dash-dot lines). System input is a comb (top panel) or a step (bottom panel).

Bayesian framework. In particular, a new probabilistic prior has been formulated directly on the unknown impulse response f, rather than on the bias error. This prior, in some sense, is the least committing one that incorporates information on both continuity of f and system BIBO-stability. The actual degree of continuity, as measured by the norm of the intensity of the white noise entering the prior, is regulated by a hyperparameter which is tuned from the data. The rate of asymptotic exponential decay is also estimated from the data. The mean square estimate is the solution of a Tikhonov regularization problem formulated on a suitable RKHS which has been fully characterized and shown to be dense in the space of continuous functions. If a nominal low-order model is needed, first, a virtually unbiased estimate of f is computed in such an RKHS and then the desired nominal model is obtained by projecting the regularized estimate onto a finite-dimensional space. Simulated benchmarks taken from the literature demonstrate the effectiveness of the proposed approach.

The results obtained in this paper, in particular those reported in Tables 1-4, could appear surprising. In fact, even when PEM+AIC is applied to candidate models which contain the true one, searching the estimate within the stable spline infinite-dimensional space leads to much better results. The reasons of the superiority of the proposed nonparametric approach are threefold. First of all, Akaike-like criteria rely on approximations of the likelihood that are only asymptotically exact. On the contrary, in our approach, the likelihood of the hyperparameters is exact, irrespective of the sample size. Second, it is well known that a drawback of Akaike-like criteria is that they neglect uncertainty of the estimated parameters [20]. Instead, the approach of this paper fully accounts for impulse response uncertainty because the hyperparameter likelihood is obtained after marginalizing with respect to the random impulse response. Finally, the issue of local maxima of the likelihood is far less critical in our nonparametric setting. In fact, the presence of only 3 unknown variables in (18) makes it possible even to use grid methods for hyper-parameter tuning [25]. Conversely, Akaike-like methods are faced with optimization in larger dimensional spaces (the joint likelihood is a function of all model parameters) and therefore more exposed to local maxima. For what concerns the computational complexity of the new method, it depends on the cost of evaluating the marginal log likelihood (18), which in general is an $O(n^3)$ problem. When dealing with large data sets, a simple yet effective strategy to reduce computational complexity is to determine the hyper-parameters using only a subset of the measurements, subsequently exploiting the entire data set to achieve $\hat{f}$ in (21), see e.g. [35]. A more sophisticated option is to combine the spectral analysis in Section (5) and the efficient computational schemes developed in [5,31]. These results exploit the fact that accurate approximations of regularized estimates in RKHS typically belong to subspaces spanned by few kernel eigenfuctions, i.e. with dimension $\breve{n}$ much smaller than n, see also [53,10]. In practice, this permits both computation of $\hat{f}$ and evaluation of the objective (18) with only $O(\breve{n}^3)$ operations, see [5,31] for details.

As for the asymptotic properties of the stable spline estimator, it can be shown that for n tending to infinity and under suitable technical conditions, a consistency property holds for a wide class of impulse responses, dense in the space of continuous functions. This result can be derived by extending the error analysis reported in [38]. A detailed derivation will be the object of future work.
Finally, it is worth stressing that the proposed method can be used also for identification of MIMO systems. In particular, this can be obtained by replacing the projection module depicted in Fig. 1 with a subspace algorithm fed with a stable spline estimator of the one-step-ahead predictor. Preliminary results on this can be found in [32,8].

Acknowledgments

This research has been partially supported by FIRB Project "Learning theory and application" and by the PRIN Projects "Artificial pancreas: physiological models, control algorithms and clinical test" and "Metodi e algoritmi innovativi per la stima Bayesiana e l'identificazione e il controllo adattativo e distribuito".

Appendix

Proof of Proposition 3

The problem can be written as finding

$$\arg\min_{\Gamma} \int_D \left(\int_{\mathbf{R}} \left(f_t - \Gamma_t(y)\right)^2 \mathbf{p}_t(f_t|y) df_t \right) \mathbf{w}(t) dt$$

which is equivalent to solving for any value of t

$$\arg\min_{\Gamma_t} \int_{\mathbf{R}} \left(f_t - \Gamma_t(y)\right)^2 \mathbf{p}_t(f_t|y) df_t$$

i.e., minimization is independent of $\mathbf{w}$ since the objective can be optimized pointwise. In particular, for any t the solution is the conditional expectation and this completes the proof.

Proof of Proposition 4

We have

$$\begin{aligned}
&\mathbf{E}[(f_t - \Gamma_t^{\mathcal{P}}(y))^2|y] \\
&= \mathbf{E}[(f_t - \Gamma_t^{\mathcal{P}}(y) + \hat{\Gamma}_t^B(y) - \hat{\Gamma}_t^B(y))^2|y] \\
&= \mathbf{E}[(f_t - \hat{\Gamma}_t^B(y))^2|y] + \mathbf{E}[(\hat{\Gamma}_t^B(y) - \Gamma_t^{\mathcal{P}}(y))^2|y] \\
&+ 2\mathbf{E}[(f_t - \hat{\Gamma}_t^B(y))(\hat{\Gamma}_t^B(y) - \Gamma_t^{\mathcal{P}}(y))|y]
\end{aligned}$$

The first term in the RHS does not depend on $\Gamma_t^{\mathcal{P}}$. As for the second term,

$$\begin{aligned}
&\mathbf{E}[(f_t - \hat{\Gamma}_t^B(y))(\hat{\Gamma}_t^B(y) - \Gamma_t^{\mathcal{P}}(y))|y] = \\
&= \int_{\mathbf{R}} (f_t - \hat{\Gamma}_t^B(y))(\hat{\Gamma}_t^B(y) - \Gamma_t^{\mathcal{P}}(y)) \mathbf{p}_t(f_t|y) df_t \\
&= (\hat{\Gamma}_t^B(y) - \Gamma_t^{\mathcal{P}}(y)) \int_{\mathbf{R}} (f_t - \hat{\Gamma}_t^B(y)) \mathbf{p}_t(f_t|y) df_t \\
&= (\hat{\Gamma}_t^B(y) - \Gamma_t^{\mathcal{P}}(y))(\int_{\mathbf{R}} f_t \mathbf{p}_t(f_t|y) df_t - \mathbf{E}[f_t|y]) = 0
\end{aligned}$$

Hence, one is reduced to solve

$$\begin{aligned}
&\arg\min_{\Gamma^{\mathcal{P}}} \int_D \mathbf{E}[(\hat{\Gamma}_t^B(y) - \Gamma_t^{\mathcal{P}}(y))^2|y] \mathbf{w}(t) dt \\
&\quad = \arg\min_{\Gamma^{\mathcal{P}}} \int_D (\hat{\Gamma}_t^B(y) - \Gamma_t^{\mathcal{P}}(y))^2 \mathbf{w}(t) dt
\end{aligned}$$

Proof of Proposition 9

By definition, $K(s,t) = W(e^{-\beta s}, e^{-\beta t})$. Hence, K is a positive definite kernel. Since W is continuous on the compact domain $S \times S$, there exists a scalar M such that

$$\sup_{(s,t) \in S \times S} W(s,t) < M < +\infty \tag{37}$$

and thus we have

$$\begin{aligned}
&\int_X \int_X |K(s,t)|^2 d\nu(s) d\nu(t) \\
&= \int_X \int_X |W(e^{-\beta s}, e^{-\beta t})|^2 \beta^2 e^{-\beta s} e^{-\beta t} ds dt \\
&= \int_S \int_S |W(s,t)|^2 ds dt \leq M^2 < +\infty
\end{aligned}$$

Furthermore, for any $x \in X$

$$\begin{aligned}
\int_X |K(x,\tau)|^2 d\nu(\tau) &= \int_X |W(e^{-\beta x}, e^{-\beta \tau})|^2 \beta e^{-\beta \tau} d\tau \\
&= \int_S |W(e^{-\beta x}, t)|^2 dt \leq M^2 < +\infty
\end{aligned} \tag{38}$$

which shows that for any $x \in X$, $K(x,.) \in \mathbf{L}_\nu^2(X)$. Further, by defining $k(x) = \int_X |K(x,\tau)|^2 d\nu(\tau)$, from (38) one also obtains that $k(x)$ is bounded on any $X_i \subset X$. The first part of the thesis now follows by exploiting Propositions 1,2 and 3 in [42].
As for (28,29), they can be easily obtained using the fact that integration on X involving kernel K may be converted into integration on S involving kernel W and exploiting (26) and Proposition 8. Finally, (30) derives from Mercer theorem on noncompact domains, see Theorem 2 in [42].

Proof of Proposition 10

We must show that

$$\int_X |\tilde{f}(t)|^p dt < +\infty \qquad a.s. \tag{39}$$

Since $\tilde{f}(t) = f_W(e^{-\beta t})$, where f_W is integrated Wiener process, it holds that

$$\int_X |\tilde{f}(t)|^p dt = \int_X |f_W(e^{-\beta t})|^p dt = \frac{1}{\beta}\int_S \frac{|f_W(\tau)|^p}{\tau} d\tau$$

Since $f_W(\tau)$ is almost surely continuous, it suffices to study how $|f_W(\tau)|^p/\tau$ behaves near zero to assess if (39) holds. In view of Proposition 8, we obtain the following Karhunen-Loeve expansion of f_W on S

$$f_W(t) = \sum_{i=1}^{+\infty} \frac{z_i}{\alpha_i^2}\phi_i(t)$$

where $\{z_i\}$ are zero-mean and independent Gaussian variables of unit variance. Now, define for $t \in S$

$$h_1(t;\alpha_i) = \left[\frac{\cos(\alpha_i t) - 1}{\alpha_i t}\right], \qquad h_2(t;\alpha_i) = \left[\frac{\sin(\alpha_i t)}{\alpha_i t}\right]$$
$$h_3(t;\alpha_i) = e^{-\alpha_i}\left[\frac{e^{\alpha_i t} - 1}{\alpha_i t}\right], \qquad h_4(t;\alpha_i) = \left[\frac{e^{-\alpha_i t} - 1}{\alpha_i t}\right]$$

By exploiting (25) and the fact that $\phi_i(0) = 0$, $\forall i$, it holds that

$$\begin{aligned} 0 &= C_1(\alpha_i) + e^{-\alpha_i} C_3(\alpha_i) + C_4(\alpha_i) \qquad i = 1, 2, ... \\ \frac{\phi_i(t)}{\alpha_i t} &= C_1(\alpha_i)h_1(t;\alpha_i) + C_2(\alpha_i)h_2(t;\alpha_i) \\ &\quad + C_3(\alpha_i)h_3(t;\alpha_i) + C_4(\alpha_i)h_4(t;\alpha_i) \qquad t \in S \end{aligned}$$

Recalling also that $|\sin(\alpha_i)| > \sqrt{1 - 4e^{-\pi}}$, $\forall i$, see [31], one easily obtains that there exists $M < +\infty$ independent of indices i, k and of $t \in S$ such that

$$\begin{aligned} &|C_k(\alpha_i)| < M \qquad k = 1, 2, 3, 4, \quad i = 1, 2, ... \\ &|h_k(t, \alpha_i)| < M \qquad t \in S, \quad k = 1, 2, 3, 4, \quad i = 1, 2, ... \end{aligned}$$

Thus, we obtain

$$\frac{f_W(t)}{t} = \sum_{i=1}^{+\infty} \frac{\phi_i(t)}{t}\frac{z_i}{\alpha_i^2} \leq 4M^2\vartheta, \qquad \vartheta = \sum_{i=1}^{+\infty} \frac{z_i}{\alpha_i}$$

In view of the definition of α_i given in (23,24), for i tending to $+\infty$, α_i tends to $+\infty$ not slower than i. Thus, ϑ is a zero-mean Gaussian with a finite variance. It emerges that realizations from $|f_W(\tau)|/\tau$ are almost surely continuous on S, and hence also those drawn from $|f_W(\tau)|^p/\tau$.

Proof of Proposition 11

As far as (31) is concerned, it can be immediately obtained by exploiting Proposition 9 and results on separability of RKHSs defined on noncompact sets, see e.g. Corollary 1 in [42]. As for density of $\mathcal{H}_K$ in the space of continuous functions, we start noticing that $\mathcal{H}_W$ is associated with the Green's function of a self-adjoint differential operator. Hence, functions in $\mathcal{H}_W$ (plus a term able to accommodate a failure of the boundary condition at zero) can approximate arbitrarily well any continuous function on a compact $S_1 \subset S$ in the sup-norm topology, see [1] and also Proposition C.1 in [33]. The result is then obtained by noticing from (27) and (31) that $\mathcal{H}_W$ and $\mathcal{H}_K$ are isometrically isomorphic, the isometry being established by a transformation $\Psi : H_W \mapsto H_K$ which maps $h(t)$, $t \in S$ into $g(\tau) = h(e^{-\beta\tau})$, $\tau \in X$.

References

[1] R.A. Adams and J. Fournier. *Sobolev Spaces*. Academic Press, 2003.

[2] B. D. O. Anderson and J. B. Moore. *Optimal Filtering*. Prentice-Hall, Englewood Cliffs, N.J., USA, 1979.

[3] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404, 1950.

[4] D. Barry. Nonparametric Bayesian regression. *The Annals of Statistics*, 14:934–953, 1986.

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

[6] M. Bertero. Linear inverse and ill-posed problems. *Advances in Electronics and Electron Physics*, 75:1–120, 1989.

[7] V.I. Burenkov. *Sobolev Spaces on Domains*. Teubner-Texte zur Mathematik, 1998.

[8] A. Chiuso, G. Pillonetto, and G. De Nicolao. Subspace identification using predictor estimation via Gaussian regression. In *Proceedings of the IEEE Conf. on Dec. and Control, Cancun, Mexico*, 2008.

[9] F. Cucker and S. Smale. On the mathematical foundations of learning. *Bulletin of the American mathematical society*, 39:1–49, 2001.

[10] G. Ferrari-Trecate, C. K. I. Williams, and M. Opper. Finite-dimensional approximation of Gaussian processes. In M. Kearns, S. Solla, and D. Cohn, editors, *Advances in Neural Information Processing Systems*. MIT Press, 1999.

[11] D. Freedman. On the Bernstein-Von Mises theorem with infinite-dimensional parameters. *The Annals of Statistics*, 27:1119–1140, 1999.

[12] A. Garulli, A. Vicino, and G. Zappa. Conditional central algorithms for worst-case set-membership identification and filtering. *IEEE Trans. on Automatic control*, 45(1):14–23, 2000.

[13] L. Giarre', M. Milanese, and M. Taragna. H_∞ identification and model quality evaluation. *IEEE Trans. on Automatic Control*, 42:188–199, 1997.

[14] G.C. Goodwin, J.H. Braslavsky, and M.M. Seron. Non-stationary stochastic embedding for transfer function estimation. *Automatica*, 38:47–62, 2002.

[15] G.C. Goodwin, M. Gevers, and B. Ninness. Quantifying the error in estimated transfer functions with application to model order selection. *IEEE Trans. on Automatic control*, 37(7):913–928, 1992.

[16] R. G. Hakvoort and P. M. J. Van den Hof. Identification of probabilistic system uncertainty regions by explicit evaluation of bias and variance errors. *IEEE Trans. on Automatic Control*, 42:1516–1528, 1997.

[17] H. Hjalmarsson. From experiment design to closed-loop control. *Automatica*, 41:393–438, 2005.

[18] H. Hjalmarsson and F. Gustafsson. Composite modeling of transfer functions. *IEEE Trans. on automatic Control*, 40:820–832, 1995.

[19] T.A. Johansen. On tikhonov regularization, bias and variance in nonlinear system identification. *Automatica*, 33:441–446, 1997.

[20] R.E. Kass and A.E. Raftery. Bayes factors. *J. Amer. Statist. Assoc.*, 90:773–795, 1995.

[21] A. Lecchini and M. Gevers. Explicit expression of the parameter bias in identification of Laguerre models from step responses. *Systems and Control Letters*, 52:149–165, 2004.

[22] L. Ljung. Model validation and model error modeling. In *Proceedings of the Astrom symposium on control, Studentliteratur, Lund, Sweden*, pages 15–42, 1999.

[23] L. Ljung. *System Identification - Theory For the User*. Prentice Hall, 1999.

[24] L. Ljung. *System Identification Toolbox V7.1 for Matlab*. Natick, MA: The MathWorks, Inc., 2007.

[25] D.G. Luenberger. *Linear and nonlinear programming*. Addison Wesley, 1989.

[26] M.N. Lukic and J.H. Beder. Stochastic processes with sample paths in reproducing kernel Hilbert spaces. *Trans. Amer. Math. Soc.*, 353:3945–3969, 2001.

[27] P.M. Makila, J.R. Partington, and T. Gustafsson. Worst-case control-relevant identification. *Automatica*, 31:1799–1820, 1995.

[28] M. Milanese, J. P. Norton, H. Piet-Lahanier, and E. Walter. *Bounding approaches to system identification*. Plenum Press, New York, NY, USA, 1996.

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

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

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

[32] G. Pillonetto, A. Chiuso, and G. De Nicolao. Predictor estimation via Gaussian regression. In *Proceedings of the IEEE Conf. on Dec. and Control, Cancun, Mexico*, 2008.

[33] T. Poggio and F. Girosi. Networks and the best approximation property. *Biological Cybernetics*, 63:169–176, 1990.

[34] T. Poggio and F. Girosi. Networks for approximation and learning. *Proceedings of the IEEE*, 78:1481–1497, 1990.

[35] C.E. Rasmussen and C.K.I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

[36] W. Reinelt, A. Garulli, and L. Ljung. Comparing different approaches to model error modeling in robust identification. *Automatica*, 38(5):787–803, 2002.

[37] C. Runge. Uber empirische funktionen und die interpolation zwischen aquidistanten ordinaten. *Zeitschrift fr Mathematik und Physik*, 46:224–243, 1901.

[38] S. Smale and D.X. Zhou. Learning theory estimates via integral operators and their approximations. *Constructive approximation*, 26:153–172, 2007.

[39] A. J. Smola and B. Schölkopf. Bayesian kernel methods. In S. Mendelson and A. J. Smola, editors, *Machine Learning, Proceedings of the Summer School, Australian National University*, pages 65–117, Berlin, Germany, 2003. Springer-Verlag.

[40] T. Soderstrom and P. Stoica. *System Identification*. Prentice Hall, 1989.

[41] A. Stenman and F. Tjarnstrom. A nonparametric approach to model error modeling. In *Proceedings of the 12th IFAC Symposium on System Identification, Santa Barbara, USA*, pages 157–162, 2000.

[42] H. Sun. Mercer theorem for RKHS on noncompact sets. *Journal of Complexity*, 21:337–349, 2005.

[43] J.A.K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle. *Least Squares Support Vector Machines*. World Scientific, Singapore, 2002.

[44] J.R. Thompson and R.A. Tapia. *Nonparametric function estimation, modelling and simulation*. Philadelphia, PA. SIAM, 1990.

[45] A.N. Tikhonov and V.Y. Arsenin. *Solutions of Ill-Posed Problems*. Washington, D.C.: Winston/Wiley, 1977.

[46] M. Vidyasagar. *A Theory of Learning and Generalization*. Springer, 1996.

[47] G. Wahba. Practical approximate solutions to linear operator equations when the data are noisy. *SIAM journal on numerical analysis*, 14:651–667, 1977.

[48] G. Wahba. *Spline models for observational data*. SIAM, Philadelphia, 1990.

[49] G. Wahba. Support vector machines, reproducing kernel Hilbert spaces and randomized GACV. Technical Report 984, Department of Statistics, University of Wisconsin, 1998.

[50] B. Wahlberg and L. Ljung. Design variables for bias distribution in transfer function estimation. *IEEE Trans. on Automatic control*, 31(2):134 – 144, 1986.

[51] C. K. I. Williams and C. E. Rasmussen. Gaussian processes for regression. In David S. Touretzky, Michael C. Mozer, and Michael E. Hasselmo, editors, *Proc. Conf. Advances in Neural Information Processing Systems, NIPS*, volume 8, pages 514–520. MIT Press, 1996.

[52] H. Zhu and R. Rohwer. Bayesian regression filters and the issue of priors. *Neural computing and applications*, 4:130–142, 1995.

[53] H. Zhu, C. K. I. Williams, R. Rohwer, and M. Morciniec. Gaussian regression and optimal finite dimensional linear models. In C.M. Bishop, editor, *Neural networks and machine learning*. Springer-Verlag, 1998.