2008 American Control Conference
Westin Seattle Hotel, Seattle, Washington, USA
June 11-13, 2008

A new kernel-based approach for system identification

Giuseppe De Nicolao and Gianluigi Pillonetto

Abstract— **We propose a new-kernel based approach for linear system identification. The impulse response is modeled as realization of a Gaussian process which includes information on smoothness and BIBO-stability. The corresponding minimum-variance estimate belongs to a Reproducing kernel Hilbert space which is given a spectral characterization and shown to be dense in the space of continuous functions. The approach may prove particularly useful in order to obtain reduced order models and assess the corresponding bias error in the context of robust identification. Several benchmarks taken from the literature demonstrate the effectiveness of the proposed approach.**

Index Terms— **linear system identification; kernel-based methods; Bayesian estimation; regularization; Gaussian processes; robust identification; stochastic embedding**

I. Introduction

We consider estimation of the impulse response of a BIBO-stable time-invariant linear system, fed with a known input, from a finite set of noisy output samples. The most used approach to solve such problem postulates a class of finite-dimensional models, possibly of low-order for control purposes [1], [2]. Then, criteria such as AIC or GCV are used to select the "best" model order. A crucial issue is quantification of both variance and bias error affecting the estimate of the nominal model. Furthermore, in presence of undermodeling, it is well known that the estimate may depend heavily on the form of the signal chosen as system input. Thus, prefiltering of output data is often advisable even if the choice of the operating frequency range may be nontrivial [3].

In the robust identification literature three main approaches have been proposed in order to characterize the variance and bias error. The first two approaches, namely stochastic embedding [4], [5] and model error modeling [6], rely on a probabilistic paradigm, while the third one, namely set-membership identification [7], [8], adopts a deterministic worst-case viewpoint. All the three methods start with the identification of a low-order nominal model by using standard techniques such as maximum likelihood or prediction error methods. Then, on the basis of the nominal model they proceed to quantify bias and variance errors. For example, in the stochastic embedding setting the bias error is described as the realization of a stochastic process, e.g. white noise with decreasing variance [4]. In the model error modeling approach, residual analysis is used to obtain information on undermodeling while set-membership identification relies on worst-case error associated with the nominal model [2].

Although the stochastic embedding has some connection with Bayesian estimation, a treatment of the robust identification problem in a fully Bayesian context is still lacking. In this paper, the main difference with respect to existing approaches is that the probabilistic prior is formulated directly on the unknown estimated impulse response rather than on the bias error. In particular, we assume that the impulse response is the realization from a Gaussian measure defined on an infinite-dimensional function space. Our new prior model avoids overfitting by including information on both continuity of the impulse response and BIBO-stability of the system. The minimum variance estimate is the solution of a Tikhonov-type regularization problem defined on a Reproducing Kernel Hilbert Space (RKHS) which is fully characterized and shown to be able to approximate a very wide class of functions. According to our strategy, first, a virtually unbiased estimate of the impulse response is obtained in such hypothesis space. Then, the desired low-order model, suitable for the intended use, may be derived from the regularized estimate. In this way, prefiltering of output data is completely avoided and replaced by projection of a regularized estimate onto a low-dimensional space.

The paper is organized as follows. In Section 2, the problem statement is given and regression via Gaussian processes [9] in RKHS [10] is concisely overviewed. In Section 3, we extend a result reported in [11] and show that the optimal estimate of a nominal model is obtained by projecting the Bayes estimate onto a finite-dimensional space. In Section 4, a new Gaussian prior for system identification is derived by defining a suitable Mercer kernel K. In Section 5, a spectral analysis of K is obtained. 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 simulated benchmarks taken from the literature are used to demonstrate the effectiveness of the proposed approach. Conclusions then end the paper.

II. Preliminaries

We are given a finite set of noisy output data from a continuous-time linear dynamic system fed with a known input $u(t)$. The measurements model is

$$y_i \doteq L_i^u[f;\Delta] + v_i = \int_0^{t_i} f(t_i - \tau - \Delta)u(\tau)d\tau + v_i \qquad (1)$$

where $\{t_i\}_{i=1}^n$ are the sampling instants, Δ may account for a possible time-delay in the system, $\{v_i\}$ is white Gaussian

G. De Nicolao is with the Dipartimento di Informatica e Sistemistica, Universita' di Pavia, Via Ferrata 1, 27100, Pavia, Italy giuseppe.denicolao@unipv.it

G. Pillonetto is with the Dipartimento di Ingegneria dell'Informazione, Universita' di Padova, Via Gradenigo 6, 35131 Padova, Italy giapi@dei.unipd.it

noise with variance σ^2. In addition, f represents the unknown system impulse response of the system which has to be estimated from the output data. In the sequel, we will mainly refer to such continuous-time setting even if the approach which will be developed can deal with discrete-time problems by just replacing integral operators with suitable discrete convolutions.

It is assumed that there exists a prior for f which consists of a Gaussian measure in an infinite-dimensional function space. To be specific, we use $\tilde{f}$ to indicate a zero-mean Gaussian process with auto-covariance $\lambda^2 K(\cdot,\cdot)$. Here, λ^2 is a possibly unknown scale factor while K represents a Mercer kernel, i.e. a mapping $K : D \times D \mapsto \Re$, where $D \subseteq \Re$, which is continuous, symmetric and positive definite[1]. Let $\mathbf{N}(\mu,\Sigma)$ denote a Gaussian density of mean μ and covariance Σ and let I_d be the $d \times d$ identity matrix. Then, the statistical model for f is specified as follows

$$\begin{aligned} f(t) &= \sum_{i=1}^{d} \theta_i \psi_i(t) + \tilde{f}(t) \qquad t \in D \\ \theta &\sim \mathbf{N}(0, \rho I_d) \qquad \rho \to +\infty \end{aligned} \tag{2}$$

where θ is independent of f and of $\{v_i\}$ while $\{\psi_i\}_{i=1}^d$ are assigned functions. In the sequel, we also use $\mathscr{B}$ to indicate the subspace spanned by $\{\psi_i\}$.

Since f and $\{v_i\}$ are assumed jointly normal, the posterior of f given $\{y_i\}$ is Gaussian as well. Our target estimate is the posterior mean. To define such estimate in rigorous mathematical terms, it is useful to recall that a Mercer Kernel K can be associated with a unique Hilbert space $\mathscr{H}$ of real valued functions, contained in the space $C(D)$ of continuous functions on D. This RKHS satisfies the following two properties

- pointwise evaluation is a linear bounded functional
- the inner product, denoted by $< \cdot,\cdot >_{\mathscr{H}}$, satisfies the reproducing property

$$< f(\cdot), K(\cdot,s) >_{\mathscr{H}} = f(s)$$

In view of the last property, K is also called *reproducing kernel* of $\mathscr{H}$ [10], [12]. In the sequel, we also use $\|\cdot\|_{\mathscr{H}}$ to denote the norm associated with $\mathscr{H}$.

Remarkably, if the dimension of $\mathscr{H}$ is infinite, it can be proved that realizations from $\tilde{f}$ do not fall in $\mathscr{H}$ with probability one [13], [14]. The following result points out that, for known $\{y_i\}$, the minimum variance estimate of f belongs to the direct sum of $\mathscr{H}$ and $\mathscr{B}$ (denoted as $\mathscr{H} \oplus \mathscr{B}$) and it can be obtained as the solution of a Tikhonov-type variational problem. Below, and in the sequel, it is assumed that $L_i^u : \mathscr{H} \mapsto \Re$ is continuous $\forall i$.

Proposition 1: Assume that f is given by (2) and is independent of $\{v_i\}$. Let $\mathbf{P}$ denote the orthogonal projection of h onto $\mathscr{H}$, in $\mathscr{H} \oplus \mathscr{B}$ and let also $\gamma = \sigma^2/\lambda^2$. For known $\{y_i\}$ and γ, the minimum variance estimate of f is given by

$$\hat{f} = \arg \min_{h \in \mathscr{H} \oplus \mathscr{B}} \quad \sum_{i=1}^{n} (y_i - L_i^u[h;\Delta])^2 + \gamma \|\mathbf{P}[h]\|_{\mathscr{H}}^2 \tag{3}$$

■

In (3), the choice of K and γ will have a major influence on the quality of the estimate. The former reflects our prior knowledge about f and will determine important properties of $\mathscr{H}$ such as its capability of approximating a wide class of functions. The latter is the so-called regularization parameter which has to correctly balance expected regularity of the solution and adherence to experimental data.

As far as K is concerned, typical choices are Gaussian or polynomial kernels. In particular, when the signal is just known to be regular, the most popular approach is to model f as an integrated Wiener process with completely unknown initial conditions. Under such statistical assumptions, denoting the corresponding kernel as W, we have [13], [15]

$$W(s,\tau) = cov(\tilde{f}(s), \tilde{f}(\tau)) = \begin{cases} \frac{s^2}{2}\left(\tau - \frac{s}{3}\right) & s \le \tau \\ \frac{\tau^2}{2}\left(s - \frac{\tau}{3}\right) & s > \tau \end{cases} \tag{4}$$

This type of kernel underlies the Bayesian interpretation of cubic smoothing splines [13]. In the sequel, let $S = [0,1]$. Then, the associated RKHS of functions on S, denoted as $\mathscr{H}_W$, is a Sobolev space whose boundary conditions are the values of the unknown function and its first-order derivative at zero. Thus, ψ_1 and ψ_2 are a constant and a linear function, respectively, so that $\theta \in \Re^2$ and

$$\mathscr{B}_W = \text{span}\{1, t\} \qquad t \in S \tag{5}$$

An explicit solution of the problem in (3) with hypothesis space $\mathscr{H}_W \oplus \mathscr{B}_W$ can be found in Chapter 1 of [12].

Once K is given, the Bayesian interpretation underlying Problem (3) can be exploited to determine γ. In particular, an effective approach is the so-called Empirical Bayes method: the unknown hyper-parameters, e.g. σ^2 and λ^2, are first estimated via a maximum likelihood approach. Then, they are set to their point estimates and $\hat{f}$ is computed from (3) as if they were perfectly known.

Finally, confidence intervals can be obtained by computing the posterior autocovariance $Var[f|y]$, see e.g. Section 4 in [16] for computational details.

III. Mean-square optimal finite-dimensional approximation

Let $y \in \Re^n$ denote a random vector. We also use $\mathscr{L}$ to indicate the space of functions mapping an interval D into the real line. A generic element of $\mathscr{L}$ is denoted by h, while $\mathscr{F} \subset \mathscr{L}$ represents the space of nominal models. As an example, $\mathscr{F}$ could contain all the first-order approximations of a stable time-invariant dynamic system, i.e.

$$\mathscr{F} = \left\{ h : h(t) = Ae^{-at}, A \in \Re, a \in \Re^+, t \in \Re^+ \right\}$$

Let Γ map vectors y into functions h, i.e. $\Gamma : \Re^n \mapsto \mathscr{L}$. Furthermore, $\Gamma_t : \Re^n \mapsto \Re$ is defined by $\Gamma(y)$ evaluated at t,

[1] A kernel K is positive definite if for any finite set $\{s_1, s_2, ..., s_k\} \subset D$ and for any real $r_1, r_2, \ldots, r_k$ we have

$$\sum_{i=1}^{k} \sum_{j=1}^{k} r_i K(s_i, s_j) r_j \ge 0$$

i.e. if $\Gamma : y \mapsto h$ then $\Gamma_t : y \mapsto h(t)$, $t \in D$. Further, we use $f(t)$, or the abbreviated notation f_t, to denote a real-valued stochastic process with $t \in D$ while $\mathbf{p}_t(f_t, y)$ indicates the joint density of f_t and y. Finally, $\mathbf{w}(t)$, $t \in D$, represents a strictly positive weighting function.

Proposition 2: If $\Gamma : y \mapsto \mathscr{L}$, the solution of the problem

$$\arg\min_{\Gamma} \int_{\Re^{n+1}\times D} (f_t - \Gamma_t(y))^2 \mathbf{p}_t(f_t, y)\mathbf{w}(t) df_t dy dt$$

is denoted by $\hat{\Gamma}^B$ and given by

$$\hat{\Gamma}_t^B(y) = \mathbf{E}[f_t|y] = \int_{\Re} f_t \mathbf{p}_t(f_t|y) df_t$$

■

Now, consider the situation where the range of Γ is restricted to the function space $\mathscr{F}$. The next result shows how the optimal estimate of f_t within such space is given by a projection (weighted by $\mathbf{w}$) of the Bayes estimate onto the space of nominal models.

Proposition 3: Let the range of Γ be restricted to $\mathscr{F}$, i.e. $\Gamma : y \mapsto \mathscr{F}$. Then, for known y, the solution of the problem

$$\arg\min_{\Gamma(y)} \int_{\Re\times D} (f_t - \Gamma_t(y))^2 \mathbf{p}_t(f_t|y)\mathbf{w}(t) df_t dt$$

is given by

$$\hat{\Gamma}(y) = \arg\min_{h\in\mathscr{F}} \int_D \left(\hat{\Gamma}_t^B(y) - h(t)\right)^2 \mathbf{w}(t) dt \qquad (6)$$

IV. System identification using a new Gaussian prior

A. Modeling the unknown impulse response

Regularization methods which exploit the kernel W in (4) are widely employed in nonparametric function estimation. However, this approach is not suitable to reconstruct the impulse response of a physical system because of the following limitations:

- Tikhonov estimator (3), with $\mathscr{H}_W \oplus \mathscr{B}_W$ as hypothesis space, is able to fit straight lines on $S = [0,1]$ without bias. However, in system identification one would like to 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, from physical constraints, it is known that impulse response variability is larger in the first time instants and then decreases over time. In particular, a prior is needed on X which includes the BIBO-stability constraint.

In the sequel, a prior is said to preserve a family of functions if the posterior expectation, given direct (i.e. input u in (1) is a Dirac delta) and noiseless samples (1) of any function f belonging to the family, coincides with the function itself. The problem that we pose is to find a mapping which converts X into S such that the prior which preserves exponentials in the old coordinates preserves straight lines in the new ones. The time-transformation has thus to map an exponential, with rate constant β, into a straight line. Once the change of coordinates is performed, we will show that impulse response stability is guaranteed by imposing that the function value at zero is null. The prior on S which has such features is exactly the integrated Wiener process with zero initial value and arbitrary first-order derivative at zero. From this discussion it comes that the desired time-transformation is

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

and the resulting kernel is

$$K(s,t) = W(e^{-\beta s}, e^{-\beta t}) \qquad (s,t) \in X \times X \qquad (7)$$

Then, our stochastic model ("stable spline" model) for the unknown impulse response becomes

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

where $\theta \in \Re$ is an infinite variance Gaussian variable and $\tilde{g}(t)$ is now a zero-mean Gaussian process, independent of θ, with auto-covariance $\lambda^2 K$. The process g is assumed independent of the measurement noise. Further, in place of (5), we define

$$\mathscr{B}_K = \text{span}\{e^{-\beta t}\} \qquad t \in X \qquad (9)$$

Finally, when dealing with discrete-time systems, one has just to consider the sampled version of the model (8).

B. Estimating hyper-parameters and impulse response

Our estimate for the impulse response is thus given by the Tikhonov estimator (3) with hypothesis space $\mathscr{H} \oplus \mathscr{B}$ replaced by $\mathscr{H}_K \oplus \mathscr{B}_K$. However, such estimator requires the knowledge of the parameter vector $\xi = [\lambda, \beta, \sigma, \Delta, \theta]$. We treat the elements of ξ as possibly unknown hyper-parameters to be determined by optimizing the marginal likelihood of y, i.e. the total probability of y, ξ and g where g is integrated out.[2].

Let $\hat{\xi}$ denote the estimate of ξ. By the same arguments as in the proof of Theorem 1.5.3 in [13], we obtain

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

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

$$c = Var\left[y|\xi = \hat{\xi}\right]^{-1} \psi(\hat{\xi})$$

where $Var[y|\xi = \hat{\xi}]$ denotes the auto-covariance of y given ξ and $\psi(\xi) \in \Re^n$ is the vector whose i-th component is $y_i - L_i^u[\theta e^{-\beta t}; \Delta]$. Needless to say, in a discrete-time context the same approach can be followed provided that integral operators are replaced by their discrete counterparts.

[2]An alternative is to treat θ as a nuisance parameter and then integrate it out to obtain the likelihood of $\{y_i\}$, see e.g. Section 1 of [13]. Our simulations, however, suggest that including also θ in ξ improves numerical stability of the computational scheme.

V. System identification kernel: spectral analysis

In this Section, also following [17], we derive a complete spectral analysis of the kernels W and K defined by (4) and (7). Relying upon this analysis, first we will demonstrate that realizations drawn from the new prior are almost surely the impulse response of a BIBO-stable system. Then, a characterization of the RKHS associated with K will be provided. We start with some definitions and a proposition which will prove useful in the sequel.

Definition 4: We define the sequence $\{\lambda_i\}$, with $\lambda_{i+1} \leq \lambda_i$, as

$$\lambda_i = (1/\alpha_i)^4 \quad i = 1,2,... \tag{10}$$

where α_i denotes that solution of

$$1/\cosh(\alpha) + \cos(\alpha) = 0 \tag{11}$$

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

$$\begin{aligned} \phi_i(t;\alpha_i) &= C_1(\alpha_i)\cos(\alpha_i t) + C_2(\alpha_i)\sin(\alpha_i t) \qquad (12)\\ &+ C_3(\alpha_i)e^{-\alpha_i(1-t)} + C_4(\alpha_i)e^{-\alpha_i t} \qquad t \in S \\ \rho_i(\tau;\alpha_i) &= \phi_i(e^{-\beta\tau};\alpha_i) \quad \tau \in X \qquad (13) \end{aligned}$$

where $\{C_k\}$ are suitable scalars, see [17] for details.

■

Proposition 5: Let W be defined by (4). Then, it holds that

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

where the sum above converges uniformly with respect to $(s,t) \in S \times S$. In addition, for $s \in S$, $\mathscr{H}_W$ is defined by

$$\mathscr{H}_W = \left\{ g \in \mathbf{L}^2(S) \quad | \quad g = \sum_{j=1}^{\infty} a_j\phi_j \ , \sum_{j=1}^{\infty} \frac{a_j^2}{\lambda_j} < \infty \right\} \tag{14}$$

■

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

Proposition 6: Associated with K, consider the integral operator on $\mathbf{L}^2_\nu(X)$ defined by

$$\Upsilon_k[f](x) = \int_X K(x,\tau)f(\tau)d\nu(\tau) \qquad x \in X$$

Then Υ_k is a bounded, compact and positive operator. In addition, for every $g \in \mathbf{L}^2_\nu(X)$, $\Upsilon_k[g] \in C(X)$.

■

The next proposition (whose proof is again omitted) shows that Υ_k is a trace-class (nuclear) integral operator on $\mathbf{L}^2_\nu(X)$, i.e. K has a spectrum composed entirely of a countable number of eigenvalues with a finite sum. In addition an explicit characterization of such spectrum is provided.

Proposition 7: We have

$$\begin{aligned} <\rho_j,\rho_k>_{L^2_\nu} &= \begin{cases} 1 & \text{if } j=k \\ 0 & \text{otherwise} \end{cases} \qquad (15)\\ \lambda_j\rho_j(s) &= \int_X K(s,t)\rho_j(t)d\nu(t) \qquad (16)\\ K(s,t) &= \sum_{j=1}^{\infty} \lambda_j\rho_j(s)\rho_j(t) \qquad (17) \end{aligned}$$

where $\{\rho_j\}$ are defined by (13) and the sum above converges uniformly with respect to $(s,t) \in X_1 \times X_2$, with X_1 and X_2 being any compact subset of X.

■

The next result provides information regarding the nature of the prior describing the impulse response of the system.

Proposition 8: Let $\mathbf{L}^p(X)$ denote the classical Lebesque spaces of p-power integrable functions on X. Let $g(t)$, with $t \in X$, be a zero-mean Gaussian process with autocovariance K. Then, realizations from $g(t)$ belong to $\mathbf{L}^p(X)$, with $p \geq 1$, almost surely. Hence, realizations from $g(t)$ are almost surely the impulse response of a BIBO linear system.
□

Recall that the optimal estimate belongs to $\mathscr{H}_K \oplus \mathscr{B}_K$. Then, by exploiting Proposition 7 and results on separability of RKHSs defined on noncompact sets (see e.g. Corollary 1 in [18]), we have

$$\mathscr{H}_K = \left\{ g \in \mathbf{L}^2_\nu(X) \quad | \quad g(s) = \sum_{j=1}^{\infty} a_j\rho_j(s) \text{ with } \sum_{j=1}^{\infty} \frac{a_j^2}{\lambda_j} < \infty \right\} \tag{18}$$

Some eigenfunctions relative to $\mathscr{H}_W$, as well as to $\mathscr{H}_K$ (with β set to 1), are displayed in Fig. 1. They provide an interesting insight into the nature of the hypothesis space chosen for system identification.
From (14) and (18) it also comes that $\mathscr{H}_W$ and $\mathscr{H}_K$ are isometrically isomorphic, the isometry being established by a transformation $\Psi : H_W \mapsto H_K$ which maps $f(t)$, $t \in S$ into $g(\tau) = f(e^{-\beta\tau})$, $\tau \in X$. Now, if h is a continuous function defined on X, f defined by $\Gamma^{-1}[h]$ turns out to be a continuous function on S. In addition, since $\mathscr{H}_W$ is associated with the Green's function of a self-adjoint differential operator, functions in $\mathscr{H}_W$ (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 Proposition C.1 in [19]. Then, the following result holds.

Proposition 9: $\mathscr{H}_K$ is dense in the space of continuous functions defined on any compact subset of X, i.e. given any continuous function h on the compact $X_1 \subset X$ and any scalar $\varepsilon > 0$, there exists $g \in \mathscr{H}_K$ such that

$$\sup_{\tau \in X_1} |g(\tau) - h(\tau)| < \varepsilon$$

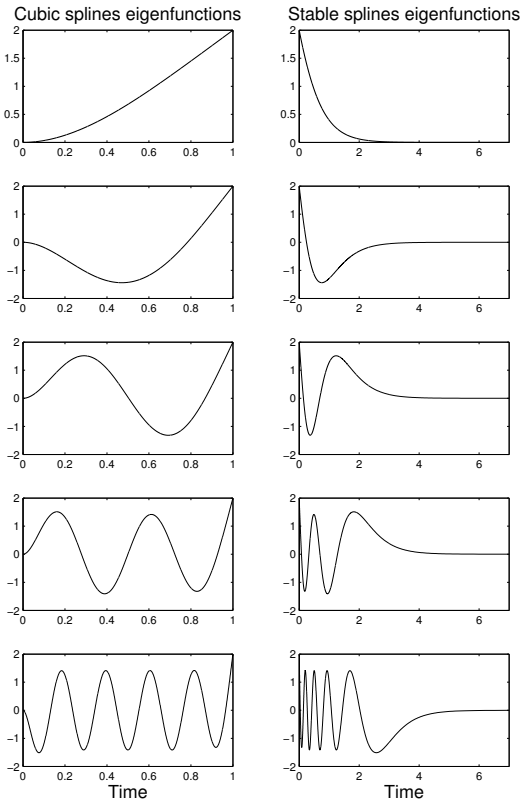

Fig. 1. $\{\phi_j\}$ (left) and $\{\rho_j\}$ (right) for j=1,2,3,5,10

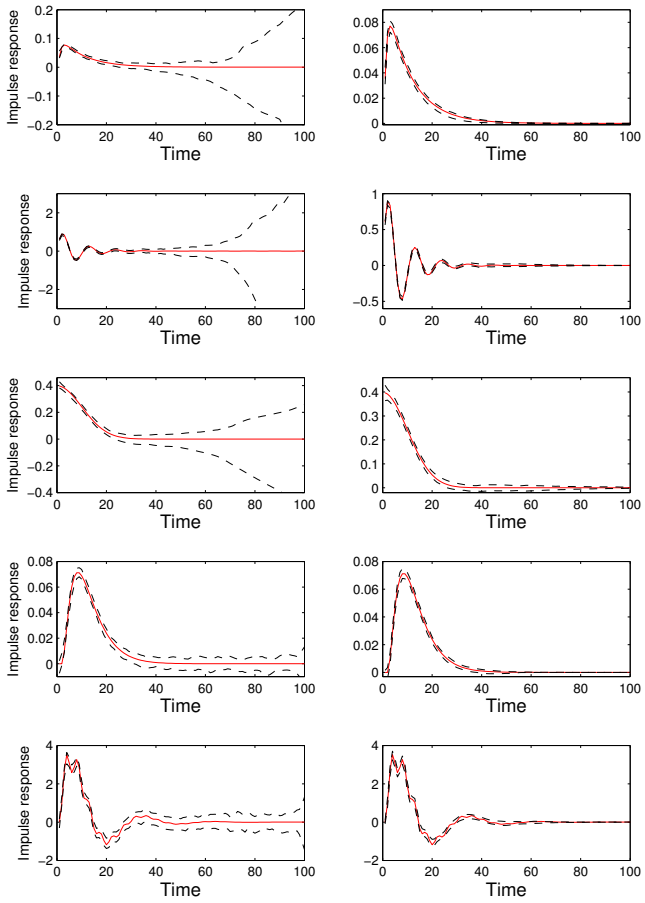

Fig. 2. Monte Carlo study (300 runs): true impulse response (solid line) and 99% variability bands of estimates (dashed lines) obtained modeling the unknown function using classical kernel W (left) and the new kernel K (right)

VI. Examples

A. Discrete-time test functions

The proposed nonparametric identification scheme is applied to identify discrete-time dynamic systems from noisy output data. In particular, we consider 5 classical simulated impulse responses which are visible in the left (and right) panels of Fig. 2 (solid line). The first two are second-order systems taken from [4], while the third one is proportional to a normal density with support only on the positive axis. The last two impulse responses are a third and a fourth order model, taken from Example 5.1 in [3] and Section 8.6 of [6], respectively. The input of the system is white noise of unit intensity in the first three cases, while in the last two cases it consists of a PRBS signal, with basic period equal to one sample. System identification has to be performed starting from 100 output noisy samples, in the time interval $[0,100]$, corrupted by a white noise. At each Monte Carlo run, SD of the noise is set to 5% of the maximum absolute value of the generated noiseless output samples. The SD of the measurement noise and the time-delay of the system are assumed to be unknown and have to be estimated from data together with λ and β. Estimation of hyper-parameters is performed by randomly choosing a starting point for the optimizer. This did not lead to any convergence problem.
For each test function, 300 Monte Carlo runs are considered. the prior model of the system impulse response is either the sampled version of the integrated Wiener process with unknown initial conditions or our new model. In the left panels of Fig. 2 we display results obtained by using kernel W. In particular, 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 information on system stability. In the right panels we display results obtained by exploiting the new kernel K. In addition to the improved quality of the estimates variability bands are much narrower and always close to the true function.

B. Continuous-time second-order system: estimate of a first-order nominal model suitable at low frequencies

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 visible in the top (and bottom) left panels of Fig. 3 (thick line) while the Bode plot of the magnitude is displayed in the top (and bottom) panel of Fig.

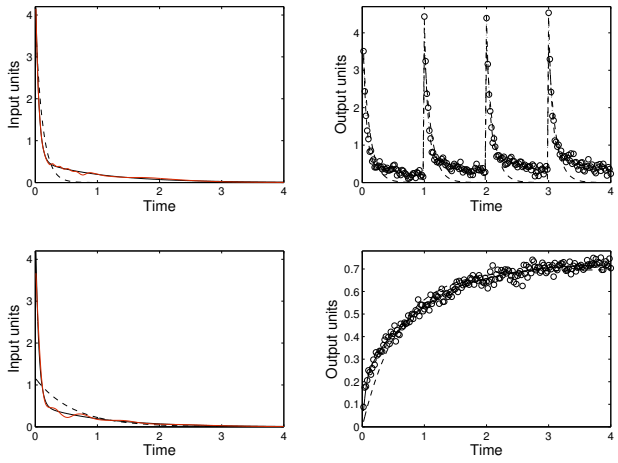

Fig. 3. Left: true impulse response (thick line), estimated impulse response obtained by fitting a first-order model to data (dashed lines) and using the new nonparametric approach (solid line). Right: noisy output samples and reconstructed output. System input is a comb (top) or a step function (bottom).

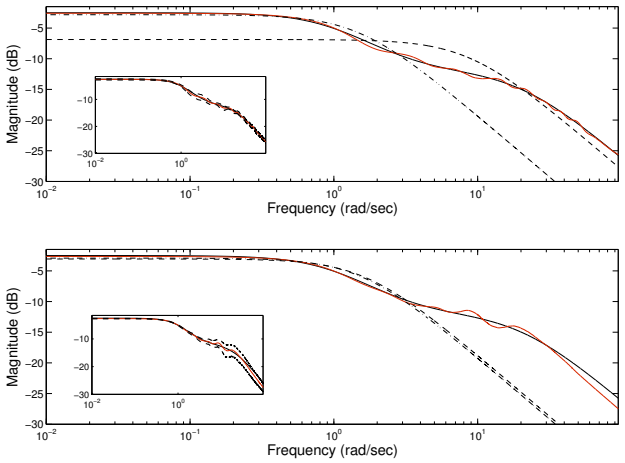

Fig. 4. True magnitude Bode plot (thick line), estimated magnitude obtained by fitting a first-order model to data (dashed lines), by using the new nonparametric approach (solid line) and by projecting the regularized estimate onto a first-order model (dash-dot lines). System input is a comb (top panel) or a step (bottom panel). Smaller plots show the true profile and the nonparametric estimate together with 99% confidence intervals (dashed lines).

4 (thick line). In Fig. 3, we plot 200 noisy output samples generated by using as input either a comb function with noise SD equal to 0.08 (top right panel) or a step function with SD = 0.02 (bottom right panel). Now, suppose 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. 3 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. 4 (dashed lines). It is apparent that when using the comb function the result 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 operating frequency range. In the left panels of Fig. 3 and in Fig. 4 we plot the estimates obtained by the new nonparametric approach proposed in this paper (solid line). One can notice that the estimate is not sensitive to the type of system input and closely approximates the magnitude plot over a wide frequency range. The desired finite order model can be derived from the regularized estimate in both situations. For instance, in Fig. 4 we display 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 the small plots of Fig. 4, we also display the true profile and the nonparametric estimate together with 99% confidence intervals (dashed lines).

C. Discrete-time second-order system and model selection issues

Consider now a discrete time second-order system taken from [20] with frequency response $F(z)$ given by

$$F(z) = \frac{z-0.6}{z^2-1.4z+0.65}$$

where poles have real part equal to 0.7. As in [20], the problem consists of reconstructing f using a step function as input. In particular, estimation has to be performed from 40 noisy measurements corrupted by a noise with a constant $SD = 0.04$ which is assumed unknown. For the sake of comparison, we consider also the identification of f by means of finite Laguerre expansions, i.e.

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

where value for p is either 0 (corresponding to FIR models) or is optimally chosen and set to 0.7 (see also Fig. 5).
We perform 5 Monte Carlo simulations consisting of 1000 runs where independent realizations of the noise are generated. In the first case study, at each Monte Carlo run f is estimated by the new nonparametric approach proposed in this paper. The other 4 studies, where least-squares estimation of the Laguerre coefficients is performed, differ from each other by the employed value for p (either FIR, that is $p=0$, or Laguerre, that is $p=0.7$) and the way model order m is selected (either AIC or "oracle"). For what concerns this latter point, letting $\hat{f}^m$ denote the estimate achieved in a certain run using m basis functions, model order is chosen either by Akaike's criterion AIC (with maximum allowed value for m equal to 20) or by using an "oracle" in which case m is given by

$$\arg\min_m \sum_{k=1}^{40} \left(f(k) - \hat{f}^m(k)\right)^2$$

In Fig. 6 we display box-plots of the root mean square errors (RMSE) achieved by the 5 estimators. Remarkably, the proposed nonparametric approach outperforms AIC-based estimators also when basis functions encode knowledge on

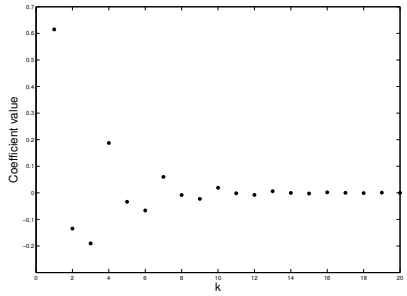

Fig. 5. Coefficients of the Laguerre expansion of f for $p = 0.7$

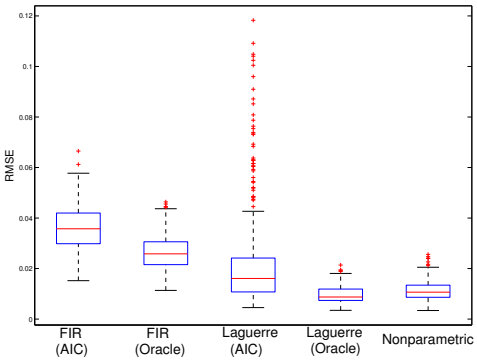

Fig. 6. Boxplots of RMSE relative to 5 estimators used to reconstruct f

pole position. Furthermore, results are better than those obtained by combining an oracle and FIR models and are close to those achieved by combining an oracle and setting p to 0.7. To understand these results, one has to consider that estimation of Laguerre coefficients by least-squares may be exposed to ill-conditioning. This problem is exacerbated when using FIR models, since they do not include any information about regularity of the impulse response. When Laguerre polynomials are optimally chosen, smoothness information on f is instead included in the model. However, AIC does not explicitly account for system stability when selecting number of basis functions to reconstruct f. In our nonparametric approach, model complexity is controlled by the regularization parameters γ and β and information on regularity and stability is incorporated in the prior for f. This explains why our identification procedure, which searches the estimate in an infinite-dimensional space, can prove more robust than finite-dimensional models.

VII. Conclusions

Current methods for robust identification start with a low-order nominal model identified by standard techniques such as least-squares. 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 Bayesian framework. In particular, a new probabilistic prior has been formulated directly on the unknown impulse response f, rather than on the bias error. The prior encodes information on both continuity of f and system BIBO-stability. The minimum variance estimate is given by a Tikhonov estimator defined on an RKHS which has been fully characterized and shown to be dense in the space of continuous functions. Following our strategy, first, a virtually unbiased estimate of f is obtained in such RKHS and then the desired nominal model is obtained by projecting the regularized estimate onto the desired finite dimensional space. Simulated benchmarks taken from the literature demonstrate the effectiveness of the proposed approach.

VIII. Acknowledgments

This research has been partially supported by FIRB Project "Learning theory and application" and by the PRIN Project "New Methods and Algorithms for Identification and Adaptive Control of Technological Systems".

References

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

[2] W. Reinelt, A. Garulli, and L. Ljung, "Comparing different approaches to model error modeling in robust identification," *Automatica*, vol. 38, no. 5, pp. 787–803, 2002.

[3] B. Wahlberg and L. Ljung, "Design variables for bias distribution in transfer function estimation," *IEEE Trans. on Automatic control*, vol. 31, no. 2, pp. 134 – 144, 1986.

[4] G. Goodwin, M. Gevers, and B. Ninness, "Quantifying the error in estimated transfer functions with application to model order selection," *IEEE Trans. on Automatic control*, vol. 37, no. 7, pp. 913–928, 1992.

[5] G. Goodwin, J. Braslavsky, and M. Seron, "Non-stationary stochastic embedding for transfer function estimation," *Automatica*, vol. 38, pp. 47–62, 2002.

[6] L. Ljung, "Model validation and model error modeling," in *Proceedings of the Astrom symposium on control, Studentliteratur, Lund, Sweden*, 1999, pp. 15–42.

[7] M. Milanese and A. Vicino, "Optimal estimation theory for dynamic systems with set membership uncertainty : An overview," *Automatica*, vol. 27, no. 6, pp. 997–1009, 1991.

[8] A. Garulli, A. Vicino, and G. Zappa, "Conditional central algorithms for worst-case set-membership identification and filtering," *IEEE Trans. on Automatic control*, vol. 45, no. 1, pp. 14–23, 2000.

[9] C. K. I. Williams and C. E. Rasmussen, "Gaussian processes for regression," in *Proc. Conf. Advances in Neural Information Processing Systems, NIPS*, D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, Eds., vol. 8. MIT Press, 1996, pp. 514–520.

[10] N. Aronszajn, "Theory of reproducing kernels," *Transactions of the American Mathematical Society*, vol. 68, pp. 337–404, 1950.

[11] H. Zhu and R. Rohwer, "Bayesian regression filters and the issue of priors," *Neural computing and applications*, vol. 4, p. 130142, 1995.

[12] G. Wahba, "Support vector machines, reproducing kernel hilbert spaces and randomized gacv," Department of Statistics, University of Wisconsin, Technical Report 984, 1998.

[13] ——, *Spline models for observational data*. SIAM, Philadelphia, 1990.

[14] M. Lukic and J. Beder, "Stochastic processes with sample paths in reproducing kernel Hilbert spaces," *Trans. Amer. Math. Soc.*, vol. 353, pp. 3945–3969, 2001.

[15] M. Neve, G. De Nicolao, and L. Marchesi, "Nonparametric identification of population models via Gaussian processes," *Automatica*, vol. 97, no. 7, pp. 1134–1144, 2007.

[16] M. Neve, G. D. Nicolao, and L. Marchesi, "Nonparametric identification of population models via gaussian processes," *Automatica*, vol. 43, pp. 1134–1144, 2007.

[17] G. Pillonetto and B. Bell, "Bayes and empirical bayes semi-blind deconvolution using eigenfunctions of a prior covariance," *Automatica*, 2007.

[18] H. Sun, "Mercer theorem for RKHS on noncompact sets," *Journal of Complexity*, vol. 21, pp. 337–349, 2005.

[19] T. Poggio and F. Girosi, "Networks and the best approximation property," *Biological Cybernetics*, vol. 63, pp. 169–176, 1990.

[20] A. Lecchini and M. Gevers, "Explicit expression of the parameter bias in identification of laguerre models from step responses," *Systems and Control Letters*, vol. 52, pp. 149–165, 2004.