2011 50th IEEE Conference on Decision and Control and
European Control Conference (CDC-ECC)
Orlando, FL, USA, December 12-15, 2011

Kernel Selection in Linear System Identification

Part I: A Gaussian Process Perspective

Gianluigi Pillonetto and Giuseppe De Nicolao

Abstract— **In some recent works, an alternative nonparametric paradigm to linear model identification has been proposed, where the unknown system impulse response is interpreted as a realization of a Gaussian process. Its autocovariance belongs to the class of so-called stable spline kernels that incorporate the stability constraint. Within this class, the order of the kernel establishes the degree of smoothness of the system impulse response. In this work, first we prove that such statistical models can be derived through Maximum Entropy arguments. Then, we show that the kernel order can be learnt from data via an efficient computational scheme that maximizes the marginal likelihood with respect to only two hyperparameters. Numerical experiments, with data generated by output error models, show the advantages of the new nonparametric estimator over the classical PEM approach that adopts cross validation to perform model order selection. In Part II of the companion papers the same identification problem is addressed in a deterministic framework.**

Index Terms— **linear system identification; output error models; kernel-based regularization; Bayesian estimation; Gaussian processes; maximum entropy**

I. Introduction

The mainstream approach to identification of linear discrete-time models is given by Prediction Error Methods (PEM), see [9], [15]. As a rule the model order is unknown and model-order selection is a key ingredient of the identification process. Models of different order are identified from data and compared resorting either to complexity measures such as FPE and AIC criteria or cross validation, splitting the data into a training and a validation set, see e.g. [1], [17].

Recently, an alternative nonparametric paradigm has been proposed that focuses on the direct identification of the impulse response [13], [12]. Instead of considering a finite dimensional parametrization of the impulse response, its identification is seen as a function learning problem formulated in an infinite-dimensional space. According to the framework of Gaussian regression [14], the unknown impulse response is seen as a realization of a Gaussian process whose autocovariance encodes the available prior knowledge. Of particular interest is a class of autocovariances, named stable spline kernels [13], [11], that encode exponential stability of the system to be identified. More precisely, the associated Gaussian process is the m-fold integration of white noise subject to an exponential time transformation. A derivation of the stable spline kernel of order $m = 1$ via "deterministic" arguments has been also recently obtained in [3]. A definite advantage of the stable spline kernel is that it is characterized by few hyperparameters that are estimated from data, e.g. via likelihood maximization. Once these hyperparameters have been fixed, the impulse reponse estimate is obtained in closed-form. Remarkably, this new paradigm has been shown to be very competitive with respect to established identification methods such as PEM and subspace methods. Within this new nonparametric framework, the scope of this paper is threefold. A first aim is to show that the statistical model underlying the nonparametric paradigm can be derived from Maximum Entropy arguments. Second, a new kernel that describes impulse responses containing high frequency poles is also derived. The third and final contribution has to do with the optimal selection of the kernel. In fact, recent works have shown that the choice of the kernel order m may play a significant role in the estimation process, tuning the degree of smoothness of the function to reconstruct, see [11] and [3]. Since m can be interpreted as a further hyperparameter that can be learnt from data, we propose a novel Bayesian model for system identification where kernels of different order may provide alternative descriptions of the autocovariance of the unknown impulse response. It is shown that the most suitable kernel can be selected via an efficient computational scheme that maximizes the marginal likelihood with respect to only two hyperparameters. We include several numerical experiments involving output error models and Gaussian measurement noise whose variance has to be estimated from the data, illustrating the advantages of the new nonparametric estimator over the classical PEM approach where model order is selected via cross validation.

The paper is organized as follows. In Section II, the statement of the problem is provided. In Section III, the linear system identification problem is given a Bayesian formulation and a new stochastic model is introduced. In Section IV, a numerical algorithm which performs kernel selection and returns the impulse response estimate is worked out. In Section V, simulated data are used to demonstrate the effectiveness of the proposed approach. Conclusions end the paper while the Appendix contains some mathematical details.

In the same session the companion paper by Tianshi Chen, Henrik Ohlsson, Graham C. Goodwin and Lennart Ljung investigates the same identification problem in a deterministic framework.

G. Pillonetto (giapi@dei.unipd.it) is with Dipartimento di Ingegneria dell'Informazione, University of Padova, Padova, Italy.

G. De Nicolao (giuseppe.denicolao@unipv.it) is with Dipartimento di Informatica e Sistemistica, University of Pavia, Pavia, Italy.

This research has been partially supported by the PRIN Projects "Sviluppo di nuovi metodi e algoritmi per l'identificazione, la stima Bayesiana e il controllo adattativo e distribuito" and Artificial Pancreas: physiological models, control algorithms and clinical test, by the Progetto di Ateneo CPDA090135/09 funded by the University of Padova and by the European Community's Seventh Framework Programme under agreement n. FP7-ICT-223866-FeedNetBack.

II. Statement of the problem and notation

For $t \in \mathbb{Z}$, we use $\{y_t\}$ to indicate noisy output data coming from a discrete-time linear dynamic system fed with a known input $\{u_t\}$. The measurements model is

$$y_t = \sum_{k=1}^{\infty} f_k u_{t-k} + v_t \tag{1}$$

where $f = \{f_t\}_{t=1}^{\infty}$ is the unknown impulse response while $\{v_t\}$ is white Gaussian noise of variance σ^2. Our problem is to estimate f from the available input-output data.

A. Notation

For our future developments, it is useful to introduce some additional notation. In particular, we use y and u to indicate the n-dimensional vectors containing the output and the input measurements, respectively. The number of data observed for $t \leq 0$ and for $t > 0$ is denoted, respectively, by n^- and n^+, so that $n = n^- + n^+$. In addition, let

$$y^+ = [y_1 \dots y_{n^+}]^T, \qquad v^+ = [v_1 \dots v_{n^+}]^T$$

be the vectors containing the observed output data and the unknown noise realizations at positive time instants. We also define

$$[U]_{ji} = u_{j-i}, \quad j = 1, \dots, n^+, \quad i \in \mathbb{N}. \tag{2}$$

so that, using notation of ordinary algebra to handle infinite-dimensional objects, with f representing an infinite-dimensional column vector, one has

$$y^+ = Uf + v^+ \tag{3}$$

In practice, the matrix U is never completely known, since only n^- input samples collected at negative time instants are available. However, in what follows we will always think of U as fully specified by setting its unobserved entries to zero.

III. System identification via Gaussian regression

A. A Bayesian framework for system identification

Under the framework of Gaussian regression [14], the impulse response f is interpreted as the realization of a stochastic process [13]. In particular, our Bayesian model is graphically illustrated in Fig. 1 (left). The node σ is deterministic and represents the noise standard deviation. It is connected to y^+ as, together with f, it defines the statistics of the output vector. When output data are available, the value of σ can be obtained using a low-bias parametric ARX description for f, see e.g. [5]. For these reasons, even if σ will be always estimated from data during our numerical experiments, hereafter it will be considered known and will not be included in the unknown hyperparameter vector.
In the network, the node $(\ell, \lambda_\ell, \beta_\ell)$ takes values in $\{1, \dots, L\} \times \mathbb{R}_+^2$ and gathers unknown hyperparameters. It is connected to f as it determines the statistics of the impulse response. To be more specific, each value of ℓ identifies a different Mercer kernel K_ℓ, i.e. a symmetric and positive definite map from $\mathbb{N} \times \mathbb{N}$ into the real line. Then, f is a a nonstationary discrete-time zero-mean Gaussian process on $\mathbb{N}$, with covariance defined by

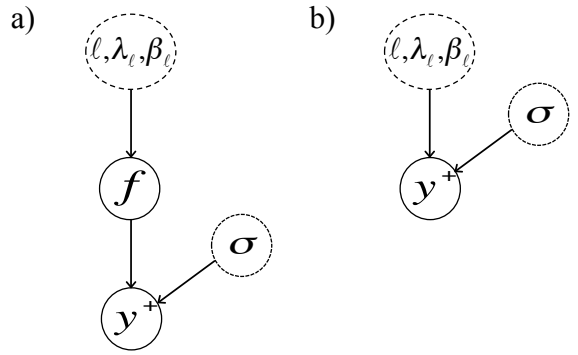

Fig. 1. Bayesian network describing the nonparametric modeling approach to system identification adopted in this paper (case *a*). The same model with f integrated out (the joint distribution of f and y^+ is marginalized with respect to f) is also reported (case *b*). In the network, dotted and solid lines denote deterministic and stochastic variables, respectively.

$$\mathbb{E}[f(t)f(s)] = \lambda_\ell^2 K_\ell(t, s; \beta_\ell) \tag{4}$$

where $t, s \in \mathbb{N}$ are two time instants. From (4), it is apparent that λ_ℓ plays the role of a scale factor while, as it will become clear in the sequel, the hyperparameter β_ℓ is related to the dominant pole of the system. In what follows, with a slight abuse of notation, the notation $K_\ell(\beta_\ell)$ is also used to indicate the matrix whose entry (t, s) is $K_\ell(t, s; \beta_\ell)$. It comes that the set $\{\lambda_\ell^2 K_\ell(\beta_\ell)\}_{\ell=1}^L$ contains different infinite dimensional autocovariances, i.e. different statistical descriptions of f, the "best" of which is to be determined from data.

B. Stable spline kernels for system identification

The quality of the nonparametric estimator exploiting the Bayesian model in Fig. 1 will crucially depend on the kernel chosen to describe f. For convenience, in this subsection we think of f as a continuos-time process on $\mathbb{R}^+$. This is instrumental to the introduction of a class of autocovariances, already discussed in [13], [11], useful also for continuous-time identification. All the derivation naturally extends to the discrete-time context considering the sampled version of the kernels described below, as discussed in the next subsection. In the literature on Gaussian regression, the adopted priors usually reflect the knowledge that the unknown function, and possibly some of its derivatives, are continuous with bounded energy. The most widely used approach models f as the m-fold integral of white Gaussian noise, so that its autocovariance is assumed proportional to

$$W_m(s, t) = \int_0^1 G_m(s, u) G_m(t, u) du \tag{5}$$

where

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

This autocovariance arises also in the Bayesian interpretation of m-th order smoothing splines, see [18] for details.

It is worth noticing that the kernel W_m does not depend on any unknown hyperparameter and does not account for impulse response stability. In fact, the variance of f increases over time. In [13], [11] new kernels, specifically tailored to linear system identification, have been introduced. In the rest of this subsection, we show how such class can be derived from maximum entropy arguments.

For our purposes, it is useful to recall that, in the Bayesian literature, Jaynes proposed a MaxEnt (maximum entropy) approach to derive complete statistical priors distributions from incomplete a priori information [7]. Among all distributions that satisfy some constraints, e.g. in terms of the value taken by a few expectations, the MaxEnt criterion selects the distribution that maximizes the entropy.

The problem of selecting priors for continuous-time Gaussian processes can still be approached using the MaxEnt paradigm by resorting to the notion of differential entropy rate [4]. In particular, let Λ_B be the class of the zero-mean stationary and differentiable Gaussian processes on $[0,1]$ with bandlimited spectrum, i.e. $S(\omega) = 0$ for $|\omega| \leq B$. The following definition is then taken from [4].

Definition 1: The differential entropy rate of $g \in \Lambda_B$ is

$$D(g) := \frac{1}{4\pi} \int_{-\infty}^{+\infty} \log[S(\omega)] d\omega \tag{6}$$

■

Now, we search for the least informative Gaussian prior on $\mathbb{R}^+$ for f including the knowledge on smoothness and exponential BIBO stability. To express the prior knowledge that the variance of the derivative of f goes exponentially to zero as t goes to ∞, it is assumed that there exists a logarithmic time transformation that makes f a stationary stochastic process in Λ_B. To express prior knowledge on the smoothness of f, we introduce the assumption that the first-order derivative $g^{(1)}$ of the process g has finite variance which ensures continuity of its realizations and hence of f. The problem of finding the prior for f is reformulated as the problem of finding a prior on g which is compatible with the constraints induced by prior knowledge. Among the infinite probability distributions that satisfy the constraints, we will look for the MaxEnt prior, i.e. the one maximizing the entropy rate (6). The next proposition shows that, under the available prior knowledge, the MaxEnt prior for g leads to a prior for f which coincides with the stable spline kernel of order 1.

Proposition 2: Let f be a stochastic process on $\mathbb{R}^+$ such that $f(-\log(t)/\beta) = g(t)$, where $g \in \Lambda_B$ with the variance of $g^{(1)}$ finite. Then, as the bandwidth B goes to ∞, the kernel of f induced by the MaxEnt prior for g, conditional on $\lim_{t\to\infty} f(t) = 0$, is

$$\Sigma_1(s,t) := \mathbb{E}(f(s), f(t)) = \max(e^{-\beta s}, e^{-\beta t}) \tag{7}$$

■

The autocovariance Σ_1 obtained above corresponds exactly to the Stable Spline kernel of order 1 introduced in [11] which thus enjoys favorable MaxEnt properties. Interestingly, this type of kernel for system identification was also derived in [3] using a totally different deterministic argument.

One can easily see that $\Sigma_1(s,t) = W_1(e^{-\beta s}, e^{-\beta t})$ from which it comes that the process g becomes the Wiener process. We notice that smoother descriptions of the impulse response can be obtained modeling g as the m-fold integration of white Gaussian noise, with $m > 1$. This argument leads to the following class of kernels Σ_m parametrized by the integer m:

$$\Sigma_m(s,t) = W_m(e^{-\beta s}, e^{-\beta t}), \qquad m = 1,2,\ldots \tag{8}$$

In particular, setting $m = 2$ in (8) the kernel becomes the stable spline kernel originally introduced in [13], i.e.

$$\Sigma_2(s,t) = \frac{e^{-\beta(s+t)} e^{-\beta \max(s,t)}}{2} - \frac{e^{-3\beta \max(s,t)}}{6} \tag{9}$$

It can be shown that modeling f as a Gaussian process of autocovariance Σ_m corresponds to assume that the Bayes estimate of the impulse response belongs to a particular reproducing kernel Hilbert space dense in the space of continuous functions, see [13] for details. In other words, irrespective of the chosen m, the estimator associated with each kernel has a negligible model bias. However, the choice of the kernel order regulates the degree of smoothness of f and is likely to have a significant influence on the estimation bias[1], which increases with m. Indeed, the variance of the estimates decreases because smoother profiles, less influenced by the measurement noise, are preferred.

C. Three kernels for nonparametric system identification

The first two statistical models for the discrete-time impulse response f are the sampled versions of stable spline kernels of order 1 and 2. The associated infinite-dimensional autocovariance matrices are defined for $s,t \in \mathbb{N}$ as follows

$$[K_1(\beta_1)]_{st} = \max(e^{-\beta_1 s}, e^{-\beta_1 t}) \tag{10}$$

$$[K_2(\beta_2)]_{st} = \frac{e^{-\beta_2(s+t+\max(s,t))}}{2} - \frac{e^{-3\beta_2 \max(s,t)}}{6} \tag{11}$$

A third kernel is obtained by the following argument. The model underlying the stable spline kernel of order 1 is a random walk (subject to an exponential time transformation parametrized by β_1), i.e.

$$g_{k+1} = g_k + w_k, \quad k = 1,2,\ldots \tag{12}$$

where $\{w_k\}$ is white noise. However, if the impulse response is rapidly oscillating due to the presence of dominant poles with negative real part, it could be better explained by a model accounting for negative correlation between adjacent samples, i.e.

$$g_{k+1} = -g_k + w_k, \quad k = 1,2,\ldots \tag{13}$$

[1] See subsection 7.3 in [6] for a discussion on model and estimation bias.

Once this model is projected onto $\mathbb{R}^+$ via an exponential transformation parametrized by β_3, the following third competitive model, defined for $s,t \in \mathbb{N}$, is obtained

$$[K_3(\beta_3)]_{st} = \begin{cases} \max(e^{-\beta_3 s}, e^{-\beta_3 t}) & \text{if } s+t \text{ is even} \\ -\max(e^{-\beta_3 s}, e^{-\beta_3 t}) & \text{if } s+t \text{ is odd} \end{cases} \tag{14}$$

In the sequel, this model is referred to as high-frequency (HF) stable spline kernel.

IV. Numerical algorithms

According to (4), the statistics of f are known up to the three hyperparameters ℓ, λ_ℓ and β_ℓ that must be estimated from data. To this aim, a key quantity is the marginal likelihood of y^+, i.e. the marginalization with respect to f of the joint density of y^+ and f. After simple computations, one obtains

$$\mathbf{p}(y^+) = \frac{\exp\left(-\frac{1}{2}(y^+)^T (Var[y^+])^{-1} y^+\right)}{\sqrt{\det(2\pi(Var[y^+]))}} \tag{15}$$

where the autocovariance of y^+ is

$$Var[y^+] = \lambda_\ell^2 U K_\ell(\beta_\ell) U^T + \sigma^2 I_{n^+} \tag{16}$$

where I_{n^+} is the $n^+ \times n^+$ identity matrix. The model, obtained after marginalization, is graphically depicted in Fig. 1b (right).

Hyperparameter estimation proceeds trough sequential optimization of sections of the marginal loglikelihood. First, fix the integer ℓ and let β_ℓ take values on a logarithmically scaled grid in $\mathbb{R}^+$. The elements of such grid are ordered in the vector Ω_β whose dimension is denoted by $|\Omega_\beta|$. Given a value of β_ℓ, the singular value decomposition of $UK_\ell(\beta_\ell)U^T$ is computed, i.e.

$$UK_\ell(\beta_\ell)U^T = P(\beta_\ell) D(\beta_\ell) P(\beta_\ell)^T \tag{17}$$

where $D = diag\{d_i(\beta_\ell)\}$, $i = 1,\dots,n^+$. Letting $z = P^T y^+$, with i-th element given by z_i, one obtains

$$\begin{aligned} J_\ell(\lambda) &:= -\log(\mathbf{p}(y^+)) \\ &= \frac{n^+}{2}\log(2\pi) + \frac{1}{2}\sum_{i=1}^{n^+}\left(\frac{z_i^2}{d_i\lambda^2+\sigma^2} + \log(d_i\lambda^2+\sigma^2)\right) \end{aligned} \tag{18}$$

Notice that both z_i and d_i depend on β_ℓ, but we omit this dependence to simplify the notation. It is easy to prove that the global minimum of J_ℓ must fall in the compact $[0, \max_i(z_i^2 - \sigma^2)/d_i]$. In addition, a good starting point for the optimizer is given by the mean of $\{z_i^2/d_i\}_{i=1}^{n^+}$, that corresponds to the minimizer of J_β once σ is set to 0. Since the pointwise evaluation of the objective $J_\ell(\lambda)$ requires only $O(n^+)$ operations, a grid method can be efficiently employed to perform optimization avoiding the risk of local minima.

For a given value of ℓ, associated with the kernel function K_ℓ, the numerical procedure returning the estimates of λ_ℓ and β_ℓ is summarized below.

Algorithm 1: The input to this algorithm includes a kernel function K_ℓ, the input-output data contained in the n-dimensional vectors u and y, the values of n^+ and n^-, the variance σ^2 and the grid Ω_β. The outputs of this algorithm are the estimates of λ_ℓ and β_ℓ and the optimized value of the marginal loglikelihood. The steps are as follows:

- Compute U using (2), setting its unobserved entries to zero.
- For $i = 1$ to $|\Omega_\beta|$, perform the following operations:
  - Letting β_ℓ^i denote the i-th element of Ω_β, compute the SVD of $UK_\ell(\beta_\ell^i)U^T$, i.e.

    $$UK_\ell(\beta_\ell^i)U^T = P(\beta_\ell^i) D(\beta_\ell^i) P(\beta_\ell^i)^T, \qquad D = diag\{d_i\}$$

  - Compute λ_ℓ^i as

    $$\begin{aligned} \arg\min_{\lambda} \quad & \frac{n^+}{2}\log(2\pi) \\ + \quad & \frac{1}{2}\sum_{i=1}^{n^+}\left(\frac{z_i^2}{d_i\lambda^2+\sigma^2} + \log(d_i\lambda^2+\sigma^2)\right) \end{aligned}$$

    and set J_ℓ^i to the minimum of the above objective
- Return $\hat{J}_\ell := \min_i J_\ell^i$ as well as the optimal pair $\{\hat{\lambda}_\ell, \hat{\beta}_\ell\}$.

After Algorithm 1 is repeated L times using each kernel K_ℓ, the optimizers $\{\hat{\lambda}_\ell, \hat{\beta}_\ell\}_{\ell=1}^L$ and the corresponding objective values $\{\hat{J}_\ell\}_{\ell=1}^L$ become available. Then, the index ℓ that minimizes $\hat{J}_\ell$ provides the "optimal" statistical model for f. Such index defines the "best" kernel and the corresponding hyperparameters, denoted, respectively, by $\hat{K}$ and $\{\hat{\lambda}_\ell, \hat{\beta}_\ell\}$. Once kernel selection has been completed, the minimum variance estimator of f admits a simple expression. In fact, f and y^+, are jointly Gaussian. Hence, exploiting a well known property of multivariate Gaussians [2], one obtains

$$\mathbb{E}[f|y^+] = \hat{\lambda}^2 \hat{K}(\hat{\beta}) U^T (Var[y^+])^{-1} y^+ \tag{19}$$

We are now in a position to summarize the numerical procedure that returns the estimate of the system impulse response.

Algorithm 2: The input to this algorithm includes the candidate kernel functions $\{K_\ell\}_{\ell=1}^L$, the input-output data contained in the n-dimensional vectors u and y, the values of n^+ and n^-, the variance σ^2 and the grid Ω_β. The output of this algorithm is the estimate of the system impulse response. The steps are as follows:

- For each value of ℓ, associated with the kernel function K_ℓ, execute Algorithm 1 and store the optimizers $\{\hat{\lambda}_\ell, \hat{\beta}_\ell\}_{\ell=1}^L$ and the corresponding optimal values $\{\hat{J}_\ell\}_{\ell=1}^L$
- Determine the best kernel minimizing $\{\hat{J}_\ell\}_{\ell=1}^L$ with respect to ℓ. Let such optimal kernel and the corresponding hyperparameters be denoted by $\hat{K}$ and $\{\hat{\lambda}_\ell, \hat{\beta}_\ell\}$.
- According to (16) and (19), return the estimate of the system impulse response as

$$\hat{f} = \mathbb{E}[f|\hat{\lambda}, \hat{\beta}] = \hat{\lambda}^2 \hat{K} U^T (\hat{\lambda}^2 U \hat{K}(\hat{\beta}) U^T + \sigma^2 I_n)^{-1} y$$

Remark 3: The proposed kernel selection scheme admits a Bayesian interpretation. First of all, regard the hyperparameters $\{\lambda_\ell, \beta_\ell\}$ of each K_ℓ as known and equal to their estimates $\{\hat{\lambda}_\ell, \hat{\beta}_\ell\}$. Then, assume that the candidate autocovariance K of f is randomly drawn with equiprobable outcomes $\{K_\ell\}_{\ell=1}^{L}$. Then, the a posteriori probability of the event $K = K_\ell$ is

$$\begin{aligned} \mathbf{P}(K = K_\ell | y^+) &= \frac{\int \mathbf{p}(y^+, f | K = K_\ell)\mathbf{P}(K_\ell) df}{\mathbf{p}(y^+)} \\ &= \frac{\mathbf{p}(y^+ | K = K_\ell)}{L\mathbf{p}(y^+)} = \frac{\exp(-\hat{J}_\ell)}{L\mathbf{p}(y^+)} \end{aligned} \quad (20)$$

Hence, our kernel selection procedure can be seen as a Bayesian model selection scheme, e.g. see [8], where one maximizes the a posteriori probability of K just neglecting the uncertainty relative to the hyperparameters λ_ℓ and β_ℓ.

V. Numerical experiments

A. Set up of four Monte Carlo experiments

The performance of the proposed approach was evaluated by 4 Monte Carlo studies, each consisting of 1000 runs. The four experiments differ each other in terms of the four features listed below.

- *Measurement noise.* Data are collected after getting rid of initial conditions and are corrupted by white Gaussian noise. In the first three experiments $\sigma = 1$ while in the last one σ is $1/3$ of the sample standard deviation of the noiseless output.
- *System input u.* In the first two Monte Carlo studies u is white noise with unit variance (WN). In the other two experiments it is a Low-Pass random Gaussian signal (LP), generated by the idinput.m Matlab function with band $[0, 0.8]$, where 0 and 0.8 are the lower and upper limits of the passband, expressed in fractions of the Nyquist frequency.
- *Data set size n.* It is 250 in all the experiments except in the second one where the size of input and output vectors is 500.
- *System generators.* Two different random generators of systems are employed. The first type is used in the first three experiments and exploits the MATLAB function drmodel.m to generate at any run a random stable 30-th order model. System poles are restricted to be inside the circle of radius 0.95 while the ℓ_2 norm of the impulse response lies between 0.5 and 10 (drmodel.m is repeatedly called at any run until such requirements are fulfilled). In this way, since σ is always equal to 1, at any run the ratio between the standard deviation of the noiseless output and that of the measurement noise is a random variable uniformly distributed in $[0.5, 10]$. The first type of system generator leads to a great variety of challenging systems. They may well contain poles at high frequency, leading to oscillating impulse responses, see the top panel of Fig. 2 where five impulse response realizations are shown. The second system generator is instead derived from the bioengineering literature. It exploits prior information coming from real data to generate smooth impulse responses whose spectrum is concentrated at low frequencies. The impulse response is the sum of three exponentials describing the kinetics of C-peptide in humans, a hormone related to insulin secretion. In particular, at each run the six system parameters are drawn from a truncated multivariate Gaussian distribution, derived from population studies and reported in the Appendix of [16]. The impulse response is then sampled with a unit sampling step. Five realizations are visible in the bottom panel of Fig. 2.

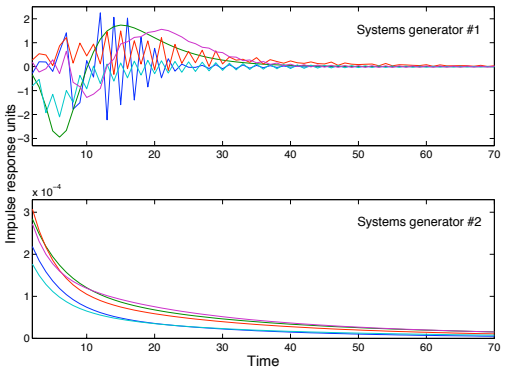

Fig. 2. Five impulse response realizations drawn from the systems generator #1 (top panel) and #2 (bottom panel).

Experiment	#1	#2	#3	#4
System input	WN	WN	LP	LP
System generator	1	1	1	2
Measurement noise	WN	WN	WN	WN
Data set size	250	500	250	250

TABLE I
FEATURES CHARACTERIZING THE 4 MONTE CARLO STUDIES.

All information relevant to the 4 experiments is summarized in Table I.

B. Performance index

The performance index regards the quality of the estimated system impulse response. In particular, we use f^j and $\hat{f}^j$ to denote, respectively, the true impulse response, randomly generated at the j-th run, and the corresponding estimate. Then, the error is computed as

$$err_j = \frac{\|\hat{f}^j - f^j\|_2}{\|f^j\|_2}, \qquad j = 1, 2, \ldots, 1000 \quad (21)$$

In (21), $\|\cdot\|_2$ is the ℓ_2 norm that is approximated numerically by considering only the first n samples of f^j and $\hat{f}^j$ ($n = 250$ or 500, depending on the considered experiment).

C. The competing estimators

During the Monte Carlo simulations, the following estimators are used:

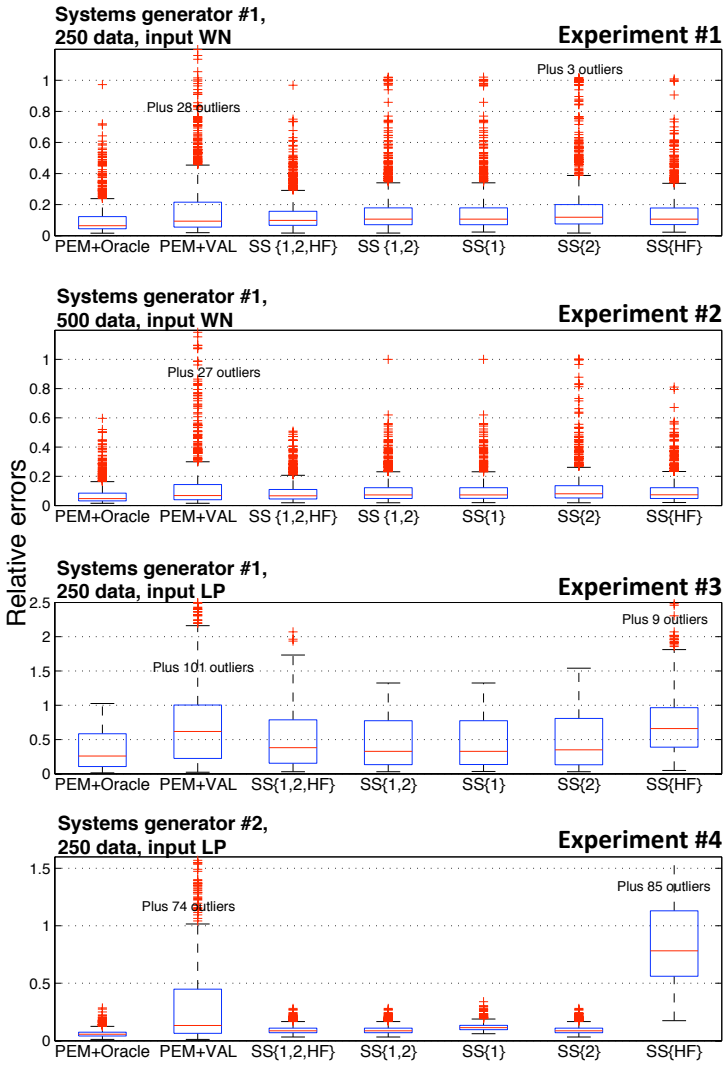

Fig. 3. Boxplot of the relative errors $\{err_j\}_{j=1}^{1000}$ defined in (21) obtained by the 7 estimators during the four Monte Carlo studies. The number of possible outliers contained in $\{err_j\}$, but not displayed in the panel, is also reported near each boxplot.

- *SS{1,2,HF}*. This is the nonparametric estimator described in the previous section that considers three candidate autocovariances: the stable spline kernels of order 1 and 2 and the new HF kernel. The value of the noise variance σ^2 is obtained at each Monte Carlo run using a low-bias parametric ARX description for f, see e.g. [5]. In the experiments #1, 3, 4, where data set size is 250, the stable spline estimator uses the first 100 input-output pairs in the training set just to obtain 100 past inputs entering U in (2). In this way, $n^- = 100$ while the dimension n^+ of y^+ in (3) is 150. In the experiment #2, where $n = 500$, we instead set $n^- = 150$ and $n^+ = 350$. Finally, the grid Ω_β contains 27 elements, being defined by

$$-\log([0.01, 0.02, 0.03, 0.04, 0.05, 0.1, 0.15, 0.2, \ldots, 0.95, 0.96, 0.97, 0.98, 0.99])$$

- *SS{1,2},SS{1},SS{2},SS{HF}*. These other four stable spline estimators are defined as above except that they use either two candidate autocovariances, excluding the HF kernel, or only one of the three candidate kernels.
- *PEM+oracle*. Classical PEM approach, as implemented in the oe.m function of the MATLAB System Identification Toolbox [10], equipped with an oracle. This is an ideal tuning, not implementable in practice, since at each run it requires the knowledge of the true impulse response f^j. Model selection is restricted to transfer functions whose numerator and denominator polynomi-

als have the same order. Then, at every run j, the oracle provides a bound on the best achievable performance of PEM selecting the model order minimizing err_j in (21).

- *PEM+VAL.* The same as above except that the model order is selected using cross validation. In particular, data are split into a training and a validation data set of equal size[2]. For every model order ranging from 1 to 30, first system identification is performed through the oe.m function fed with the training set. Then, the prediction error on the validation set is computed using the predict.m function fed with all the input data contained in the training and test set. Finally, the estimate of the impulse response is obtained fixing the model order to the one leading the best prediction on the validation data set and using the oe.m function fed with all the available measurements (the union of the training and of the validation data sets).

The system input delay is assumed known and its value is provided to all the estimators described above.

D. Results

The four panels of Fig. 3 report boxplots of the relative errors $\{err_j\}_{j=1}^{1000}$ defined in (21) obtained by the 7 different estimators in the four experiments. The performance reference is represented by *PEM+oracle*.

We start comparing the results obtained by *PEM+oracle*, *PEM+VAL* and $SS\{1,2,HF\}$. In all the four case studies, the nonparametric estimator $SS\{1,2,HF\}$ provides results comparable to those of *PEM+oracle* and outperforms *PEM+VAL*. Notice also that, differently from $SS\{1,2,HF\}$, the boxplots produced by *PEM+VAL* contain many outliers, not all displayed in Fig. 3. This is particularly true in the last two experiments where system identification is made more difficult by the LP input (less exciting than WN).

We now compare the results obtained by $SS\{1\}$, $SS\{2\}$ and $SS\{HF\}$. In the first two experiments (first two panels of Fig. 3) all the three nonparametric estimators perform well, with the performance of $SS\{1\}$ and $SS\{HF\}$ slightly better than that of $SS\{2\}$. This derives from the fact that in the first two case studies the input is WN and the systems to identify may well possess oscillating poles that are better captured by the HF and the first-order stable spline kernel. In the third experiment (third panel of Fig. 3), the adopted system generator is the same but the input is LP. One can see that the performance of $SS\{1\}$ and $SS\{2\}$ is much similar while the quality of the results achieved by $SS\{HF\}$ deteriorates, possibly due to the lack of high frequency content in the training signal. On the other hand, $SS\{1\}$ and $SS\{2\}$ can deal also with high-frequency impulse responses thanks to their built-in regularization properties that keep estimation variance low.

In the last experiment (fourth panel of Fig. 3) the estimator $SS\{2\}$ outperforms $SS\{1\}$ and $SS\{HF\}$. This is not surprising since the frequency content of the transfer functions to reconstruct is now mostly located at low frequencies. Hence, the impulse responses are better described as realizations of smooth processes.

In view of these results, the robustness of the estimator $SS\{1,2,HF\}$ is remarkable. In fact, Fig. 3 reveals that the proposed kernel selection procedure very often leads to the kernel guaranteeing the minimum estimation error. In other words, the estimator adapts well to the different experimental conditions. In fact, the boxplot of the errors coming from $SS\{1,2,HF\}$ always appears as a suitable synthesis of those associated with $SS\{1\}$, $SS\{2\}$ and $SS\{HF\}$. For instance, in the last experiment it turns out that the HF kernel, which leads to large estimation errors, is almost always discarded while the the second-order stable spline kernel is selected in almost all of the 1000 runs.

E. Computational considerations

A simple analysis of Algorithm 2 developed in Section IV reveals that, for large values of the data set size n, the complexity of the nonparametric scheme proposed in this paper is $O\left(L|\Omega_\beta|(n^+)^3\right)$, where $|\Omega_\beta|$ is the grid size for β, while L is the number of candidate autocovariances. A possible strategy to reduce the number of operations is to use a reduced data set to perform kernel selection and then employ the entire data set to estimate the impulse response via (19). As an example, we have repeated Experiment #2 forcing the estimator $SS\{1,2,HF\}$ to exploit only 250 data, in place of the overall 500, for kernel selection. More specifically, we split the 250 data available for hyperparameter estimation setting $n^- = 150$ and $n^+ = 100$. Then, we set $n^- = 150$ and $n^+ = 350$ to obtain the impulse response estimate from the 500 measurements. In this way, using a Pentium 3GHz, very few seconds were needed at every Monte Carlo run to complete system identification. Define the average identification error at the j-th run as

$$\overline{err}_j = \frac{\sum_{k=1}^{j} err_k}{j} \tag{22}$$

Then, Fig. 4 plots $\overline{err}_j$ as a function of the Monte Carlo run, using the entire and the reduced data set for selecting the kernel. One can see that the quality of the estimates is very similar, further corroborating the robustness of the proposed approach.

VI. Conclusions

In this paper, we have extended a recently proposed nonparametric approach to system identification where the unknown impulse response is modeled as a Gaussian process with autocovariance defined by stable spline kernels. First, we have shown how the class of autocovariances introduced in [13], [11], [3] can be derived via Maximum Entropy arguments. In addition, we have also proposed a new kernel suited to systems containing high frequency poles. This analysis leads to the definition of a new Bayesian model for system identification where different competitive kernels are introduced. All the kernels include the stability constraint, but describe impulse responses with different degrees of

[2]We have also tried a different partition where the split is 2/3 for training and 1/3 for validation, as e.g. suggested in Chapter 7 of [6], finding that this does not lead to improved results.

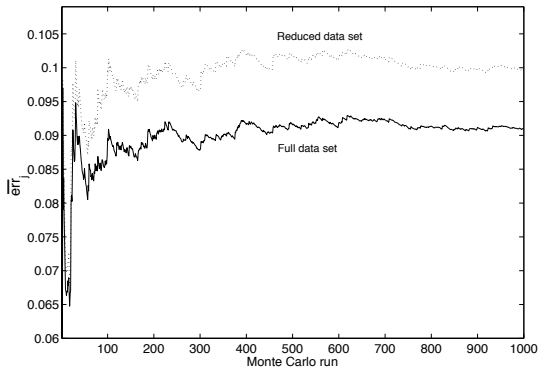

Fig. 4. Monte Carlo experiment #2. Average reconstruction error $\overline{err}_j$, see (22), as a function of the Monte Carlo run, obtained by $SS\{1,2,HF\}$ using the full and reduced data set for hyperparameter estimation.

smoothness. An efficient computational scheme that performs kernel selection and impulse response estimation has been worked out.

Numerical experiments involving output error models show that the proposed technique outperforms the classical PEM approach that uses cross validation for model selection. The main drawback of PEM+VAL has to do with the possible selection of wrong-order models that perform well on the validation set but provide unreliable estimates of the impulse response. In fact, when some frequency ranges are dominant in the validation set, cross validation may select models that perform badly in other frequency ranges. This does not happen in the SS approach due to its inherent regularization properties.

APPENDIX: PROOF OF PROPOSITION 2

The proof is an adaption of the proof of the main result in [4] where the issue of stability is not addressed. We introduce the following constraints on g

$$Var[g^{(k)}] = \lambda_k^2, \qquad k = 0,1 \tag{23}$$

Notice that the constraint on the variance of g (k=0) will be subsequently relaxed by letting $Var[g]$ go to ∞. Now, the MaxEnt spectrum of g solves the equality constrained optimization problem whose Lagrangian is

$$\begin{aligned} L[S,\varsigma] &= \frac{1}{4\pi}\int_{-B}^{B}\log[S(\omega)]d\omega \\ &+ \sum_{k=0}^{1}\varsigma_k\left(\lambda_k^2 - \frac{1}{2\pi}\int_{-B}^{B}w^{2k}S(\omega)d\omega\right) \end{aligned}$$

where $\varsigma = [\varsigma_0 \quad \varsigma_1]$ is the vector of Lagrange multipliers. We have

$$\frac{\partial L}{\partial S} = \frac{1}{2\pi}\int_{-B}^{B}\left(\frac{1}{2}\frac{1}{S(\omega)} - \varsigma_0 - \varsigma_1\omega^2\right)d\omega$$

Hence

$$S(\omega) = \begin{cases} \frac{1}{2(\varsigma_0+\varsigma_1\omega^2)} & \text{if } -B < \omega < B \\ 0 & \text{otherwise} \end{cases}$$

where $\{\varsigma_k\}$ must satisfy the following conditions coming from the equality constraints:

$$\int_{-B}^{B}\frac{\omega^{2j}}{\sum_{k=0}^{1}\varsigma_k\omega^{2k}}d\omega = 2\pi\lambda_j^2, \quad j = 0,1 \tag{24}$$

If λ_0^2 tends to ∞, then, from (24) the corresponding Lagrange multiplier ς_0 goes to zero. Thus, the MaxEnt spectrum of g becomes

$$S(\omega) = \frac{1}{\varsigma_1\omega^2}$$

Roughly speaking, for B tending to ∞, this corresponds to the spectrum of the integrated continuous-time white noise. The additional constraint $\lim_{t\to\infty} f(t) = 0$ implies $\lim_{t\to 0} g(t) = 0$ so that, under the logarithmic time-transformation, g becomes a Brownian motion. This corresponds to $\mathbb{E}[g(s),g(t)] = \min(s,t)$ for $s,t \in [0,1]$ from which, using the equation $f(t) = g(e^{-\beta t})$, the stable spline kernel (7) is immediately obtained.

REFERENCES

[1] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723, 1974.

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

[3] T. Chen, H. Ohlsson, and L. Ljung. On the estimation of transfer functions, regularization and Gaussian processes - Revisited. In *Proceedings of the 2011 IFAC World Congress, Milan (submitted)*, 2011.

[4] G. De Nicolao, G. Ferrari Trecate, and A. Lecchini. MaxEnt priors for stochastic filtering problems. In *Proc. Symp. on Math. Theory of Networks and Systems (MTNS'98)*, pages 755–758, Padova, Italy, 1998.

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

[6] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning*. Springer, 2008.

[7] E.T. Jaynes. On the rationale of maximum-entropy methods. *Proceedings of IEEE*, 70:939–952, 1982.

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

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

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

[11] G. Pillonetto, A. Chiuso, and G. De Nicolao. Regularized estimation of sums of exponentials in spaces generated by stable spline kernels. In *Proceedings of the 2010 American Control Conference, Baltimora*, 2010.

[12] G. Pillonetto, A. Chiuso, and G. De Nicolao. Prediction error identification of linear systems: A nonparametric Gaussian regression approach. *Automatica*, 47:291–305, 2011.

[13] G. Pillonetto and G. De Nicolao. A new kernel-based approach for linear system identification. *Automatica*, 46:81–93, 2010.

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

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

[16] G. Sparacino, C. Tombolato, and C. Cobelli. Maximum-likelihood versus maximum a posteriori parameter estimation of physiological system models: the C-peptide impulse response case study. *IEEE Transactions on Biomedical Engineering*, 47, 2000.

[17] M. Stone. An asymptotic equivalence of choice of model by cross-validation and Akaike criterion. *Journal of the Royal Statistical Society*, 39, 1977.

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