A New Kernel-Based Approach for NonlinearSystem Identification

Gianluigi Pillonetto, *Member, IEEE*, Minh Ha Quang, and Alessandro Chiuso, *Senior Member, IEEE*

***Abstract*—We present a novel nonparametric approach for identification of nonlinear systems. Exploiting the framework of Gaussian regression, the unknown nonlinear system is seen as a realization from a Gaussian random field. Its covariance encodes the idea of "fading" memory in the predictor and consists of a mixture of Gaussian kernels parametrized by few hyperparameters describing the interactions among past inputs and outputs. The kernel structure and the unknown hyperparameters are estimated maximizing their marginal likelihood so that the user is not required to define any part of the algorithmic architecture, e.g., the regressors and the model order. Once the kernel is estimated, the nonlinear model is obtained solving a Tikhonov-type variational problem. The Hilbert space the estimator belongs to is characterized. Benchmarks problems taken from the literature show the effectiveness of the new approach, also comparing its performance with a recently proposed algorithm based on direct weight optimization and with parametric approaches with model order estimated by AIC or BIC.**

***Index Terms*—Bayesian estimation, direct weight optimization, Gaussian processes, kernel-based methods, nonlinear system identification, regularization.**

I. Introduction

BLACK-BOX identification approaches are widely used to learn models from observed data. When the system to be modeled can be well approximated by a linear one, many identification techniques are available in the literature [1], [2]. In particular, the classical identification paradigm postulates a set of competitive parametric models and chooses the most suitable one using complexity criteria such as the Bayesian information criterion (BIC) or Akaike's information criterion (AIC) [3]–[5]. The recent work [6] introduced an alternative nonparametric paradigm for linear system identification, based on the framework of regression via Gaussian processes. Instead of postulating finite-dimensional models, the system impulse response is searched for in a suitable infinite-dimensional space. This type of scheme has proved especially effective for model order selection, see also [7] where advantages are described in the context of predictor estimation.

Much recent research has been devoted to extend classical parametric methods for linear system identification to nonlinear systems modeling [8], [9]. In this scenario, NARX/NARMAX models [10] are often employed. If the parametric functional relationship among present output and past input/output data is specified, the resulting estimation problem can be solved by standard optimization algorithms. Even in this ideal scenario in which the parametric structure has been fixed, one is faced with non trivial issue related to non-convexity of the functional to be optimized as a function of the model parameters. One is eventually faced with a non-convex optimization problem, possibly in high dimension, where finding the optimizer is often out of reach.

On top of this, the user is faced with the challenge of determining the optimal structure of the model which is often described by a set of basis functions, see also [11], [12] for approaches relying upon neural and fuzzy networks. A significant difficulty is to handle the large number of potentially relevant regressors, e.g., related to delays and powers of monomials entering a polynomial model. Beyond classical complexity criteria, such as AIC and BIC, step-wise regression algorithms exploiting a validation data set are often used in this context, see [13], [14]. The forward-regression orthogonal estimator provides suboptimal solutions by iteratively incrementing model structure on the basis of the ability of model terms to describe the output variance [15], see also [16] where simulation (in place of prediction) error reduction ratio is exploited. Another class of algorithms relies upon Direct Weight Optimization [17], e.g., equipped with the so called minimal probability approach discussed in [18]. Here, after fixing the model order, the predictor function is computed using a weighted linear combination of the observed outputs in a neighborhood of the target point. The extent of the neighborhood is determined by a parameter that has to be tuned by the user. The reader is also refereed to [9] for an approach for regressor selection that uses ANOVA.

The aim of this paper is to extend the linear identification techniques developed in [6] to the nonlinear context. We will present a novel estimator for nonlinear system identification whose architecture depends on parameters that are all learnt from data. In particular, our numerical procedure does not require the user to select critical variables such as regressors/model order and automatically learn the basis functions from the training set. In addition, if the system is known to have a significant linear component, this information can be included in the model. The new approach is cast in the framework of regression via Gaussian processes, see, e.g., [19], [20]. We convexify the identification problem by placing a suitable

Manuscript received March 16, 2010; revised January 16, 2011; accepted March 14, 2011. This work was supported in part by the PRIN Project "Sviluppo di nuovi metodi e algoritmi per l'identificazione, la stima Bayesiana e il controllo adattativo e distribuito", in part by the Progetto di Ateneo CPDA090135/09 funded by the University of Padova, in part by the European Community's Seventh Framework Programme under agreement FP7-ICT-223866-FeedNetBack, and in part by the German Research Foundation grant DFG:GZ WI 1515/2-1. Date of publication March 24, 2011; date of current version December 07, 2011. Recommended by Associate Editor E. Weyer.

G. Pillonetto is with the Dipartimento di Ingegneria dell'Informazione, Universita' di Padova, Padova 35131, Italy (e-mail: giapi@dei.unipd.it).

M.H. Quang is with the Istituto di Tecnologia Italiano, Genova 16163, Italy (e-mail: minh.haquang@iit.it).

A. Chiuso is with the Dipartimento di Tecnica e Gestione dei Sistemi Industriali, University of Padova, Vicenza 36100, Italy (e-mail: chiuso@dei.unipd.it).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TAC.2011.2131830

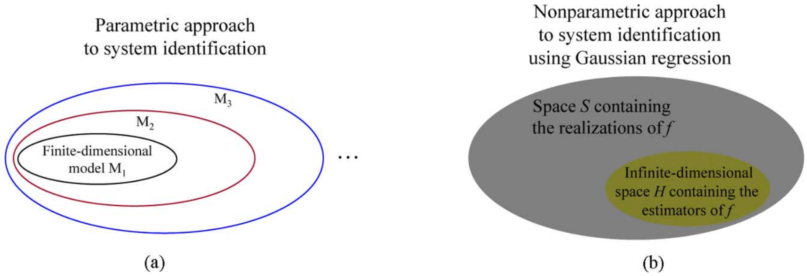
Parametric approach
to system identification
M_3
M_2
Finite-dimensional model M_1
…
Nonparametric approach
to system identification
using Gaussian regression
Space S containing the realizations of f
Infinite-dimensional space H containing the estimators of f
(a)
(b)

Fig. 1. *Left*: Parametric approach to system identification. $\{M_i\}$ denote finite-dimensional spaces of different complexity. Model order is typically chosen by criteria such as AIC or BIC requiring the solution of a nonlinear optimization problem for each postulated model and relying upon likelihood functions which are only asymptotically exact. *Right*: Nonparametric approach for system identification using Gaussian regression. The unknown system is defined by a realization from a zero-mean Gaussian random field f whose covariance (kernel) encodes the stability constraint. Model order selection is replaced by estimation of few hyperparameters entering the kernel, obtained by optimizing a likelihood function that is exact, irrespective of the sample size, and accounts for the uncertainty of f. Once such parameters are determined, the minimum variance estimate of f is available in closed form and belongs to a (generally infinite-dimensional) Reproducing Kernel Hilbert Space $\mathcal{H}$. In the linear scenario described in [6], the realization from f is the unknown system impulse response. In the nonlinear scenario treated in this paper, it models the unknown predictor mapping the past inputs and outputs $\{y_k, u_k\}_{k=-\infty}^{t-1}$ into the predicted output $\hat{y}_{t|t-1}$.

normal prior on an infinite-dimensional function space. In our approach, parametric models are replaced by nonparametric ones, defined via a class of kernels (covariances) specifically suited to nonlinear system identification. Each kernel encodes the idea of "fading" memory in the predictor and consists of a mixture of Gaussian kernels, described by a few hyperparameters which account for the interactions among past outputs inputs; each covariance is associated with an infinite-dimensional space that will contain the estimators. The most suitable kernel is determined by evaluating its posterior probability via a Bayesian model selection paradigm. Estimating the hyperparameters involves a non-convex but low-dimensional optimization problem. Once the kernel is selected, the nonlinear model is obtained by solving a convex Tikhonov-type variational problem defined on a Reproducing Kernel Hilbert Space (RKHS), see [21] and also [22] where regularization in RKHS for solving regression and classification problems is widely discussed. This gives the solution in closed form solving a set of linear equations. This new nonparametric paradigm to system identification is graphically depicted in Fig. 1

The paper is organized as follows. Section II reports the problem statement while in Section III the new prior for nonlinear system identification is formulated. In Section IV the full Bayesian model used to solve the nonlinear identification problem is presented. In Section V the algorithm for nonlinear system identification is provided. Section VI is a bit more technical and provides an in-depth discussion on the RKHS the estimator belongs to. In Section VII the new method is tested on six benchmarks problems taken from the literature and a system close to be linear. In addition, a comparison with the minimal probability estimator described in [18] is also reported. Conclusions are finally offered in Section VIII while proofs are gathered in Appendix.

Notation

In the paper f, g, h or q denote random fields while r or v indicate deterministic functions. In addition, given a column vector or sequence w, w_i is its i-th element. We shall use the notation of finite dimensional linear algebra, such as matrix products, also with (infinite dimensional) sequences $w := [w_1, w_2, \ldots]^T \in \mathbb{R}^\infty$. This is perfectly legitimate provided convergence is guaranteed. The symbol $||w||$ indicates the Euclidean norm if w is a vector or the norm in ℓ^2 (the space of square-summable sequences) if w is a sequence, i.e., $w \in \mathbb{R}^\infty$.

The symbols $\mathbb{E}[\cdot]$ and $\mathbb{E}[\cdot|\cdot]$ denote, respectively, expectation and conditional expectation.

We also use $\mathbb{N}$ to indicate the natural numbers not including 0, $\mathbb{Z}$ indicates the integers and C_0 denotes the space of continuous functions equipped with the classical sup-norm (uniform topology). In addition, L^m, $m \in \mathbb{N}$, denotes the classical Lebesgue spaces while $\mathcal{N}(\mu, \Sigma)$ indicates a Gaussian random vector with mean μ and covariance Σ. The terms kernel and covariance will be hereby used interchangeably. The symbol $\perp$ is also used to indicate statistical independence between random variables.

With some abuse of notation the symbol y_t will both denote a random variable (from the random process $\{y_t\}_{t\in\mathbb{Z}}$) and its sample value. We use $\{y_t\}$ to denote the set of noisy output measurements from a nonlinear dynamic system fed with a measured input $\{u_t\}$. In particular we define the sets of past measurements at time t

$$y^t = [\, y_{t-1} \quad y_{t-2} \ldots]^T, \quad u^t = [\, u_{t-1} \quad u_{t-2} \ldots]^T$$

and the vector

$$y^+ = [\, y_1 \quad y_2 \ldots y_N \,]^T .$$

The shorthand $y^- := y^1 = [\, y_0 \quad y_{-1} \ldots]^T$ is reserved for the past at time $t = 1$.

Given two fixed time instants t and τ we also define, for $i \in \mathbb{N}$, $x_i, z_i \in \mathbb{R}^2$ as

$$\begin{array}{ll} x_1 = [\, y_{t-1} \quad u_{t-1} \,], & z_1 = [\, y_{\tau-1} \quad u_{\tau-1} \,] \\ x_2 = [\, y_{t-2} \quad u_{t-2} \,], & z_2 = [\, y_{\tau-2} \quad u_{\tau-2} \,] \\ \quad \vdots & \quad \vdots \end{array} \tag{1}$$

with $x \in \mathbb{R}^{\infty \times 2}$ and $z \in \mathbb{R}^{\infty \times 2}$ given by

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \end{pmatrix}, \quad z = \begin{pmatrix} z_1 \\ z_2 \\ \vdots \end{pmatrix} \tag{2}$$

II. Problem Statement

Given the scalar[1] input and output time series $\{u_t\}$ and $\{y_t\}$, $t \in \mathbb{Z}$, we assume that the one-step ahead predictor

$$\hat{y}_{t|t-1} = F(y^t, u^t) := \mathbb{E}[y_t | y^t, u^t] \tag{3}$$

[1] The results of this paper can be easily extended to the multi-input, multi-output (MIMO) case, but for ease of exposition we restrict ourselves to the single-input, single-output (SISO) case.

is time invariant, i.e., F does not depend explicitly on time. Above, and in the sequel, it is always assumed that all the conditional expectations are well defined and, for simplicity of exposition, we shall also assume that the predictor $\hat{y}_{t|t-1}$ is strictly causal, i.e., it does not depend upon u_t. It is then perfectly legitimate to consider y_t as the output of the nonlinear dynamic model[2]:

$$y_t = F(y^t, u^t) + e_t, \quad t = 1, \ldots, N \tag{4}$$

where $\{e_t\}$ is the innovation sequence, i.e., the one step ahead prediction error $e_t := y_t - \mathbb{E}[y_t|y^t, u^t]$. The innovation sequence is, by construction, a martingale difference sequence [23] with respect to the sigma algebra $\mathcal{Z}_t$ generated by past measurements (y^t, u^t). We shall also assume e_t is Gaussian and it has constant but unknown conditional variance

$$Var\{e_t\} = Var\{e_t|y^t, u^t\} = Var\{y_t|y^t, u^t\} = \eta^2.$$

Note $e_t = y_t - F(y^t, u^t)$ is contained in the sigma algebra generated by y^{t+1}, u^{t+1}; therefore e_t is uncorrelated with $e_s \, \forall \, t \neq s$. This, together with the Gaussianity and constant variance assumptions, implies that e_t is a stationary stochastic process and $e_t \perp e_s, t \neq s$. Notice that (4) contains as special cases, e.g., NFIR, NARX and NARMAX models. Our problem is to determine F from the available input-output data (training set). To do so we shall adopt a prediction error minimization type of criterion in Gaussian regression framework. The quality of the obtained model will be then assessed in terms of predictive capability on new data (test set).

Remark 1: It is worth recalling that the most common notation for NARMAX model is (see, e.g., [10])

$$\begin{aligned} y_t = g(&y_{t-1}, \ldots, y_{t-n_y}, u_{t-1}, \\ &\ldots, u_{t-n_u}, e_{t-1}, \ldots, e_{t-n_e}) + e_t; \end{aligned} \tag{5}$$

which defines a proper subclass of (4) in which a specific structure of the predictor is postulated. In fact in (5) the "orders" n_y, n_u, n_e need to be specified. In practice they have to be estimated, e.g., using variable selection techniques. Instead, in (4), no specific structure is postulated and the predictor can depend arbitrarily on past input-output data.

III. New Kernel for Nonlinear System Identification

A. Gaussian Kernel Limitations in Linear and Nonlinear System Identification

The perspective adopted in this paper consists in interpreting F in (4) as a realization from a zero-mean Gaussian random field, denoted by f, i.e., $F(\cdot) = f(\cdot, \omega)$ for a certain ω in the sample space contained in the probability space underlying f. Notice also that the Gaussianity assumption implies that, in order to define f, we just need to specify the covariance between the random variables $f(x)$ and $f(z)$ where x and z represent any couple of possible arguments of F.

As an introduction to our modeling problem, let's assume just for a while that the predictor (4) is linear and the problem is formulated in continuous-time. For simplicity, we also neglect the dependence on y^t so that (4) becomes the following output error model:

$$y_t = \int_{-\infty}^{t} F(t-\tau) u_\tau \, d\tau + e_t \tag{6}$$

where, in this linear context, $F : \mathbb{R}^+ \mapsto \mathbb{R}$ is the unknown predictor impulse response. According to the Gaussian regression approach, F is assumed to be a realization from a zero-mean Gaussian process f on $\mathbb{R}^+$, which we assume independent of $\{e_t\}$. The covariance (or kernel) of f has to be specified, possibly based on some prior information on the system to be modeled.

The Gaussian kernel is probably the most commonly used in nonparametric estimation and is defined by

$$\begin{aligned} G(t,\tau) &= cov(f(t), f(\tau)) \\ &= \lambda \exp\left(-\frac{(t-\tau)^2}{\sigma^2}\right), \quad t, \tau \in \mathbb{R} \end{aligned}$$

where the kernel width $\sigma \in \mathbb{R}^+$ and the scale factor $\lambda \in \mathbb{R}^+$ are hyperparameters, i.e., parameters of a prior distribution, that need to be estimated from data. It is a known fact that the mean of f, conditional on the output measurements, belongs to that unique RKHS $\mathcal{H}_G$ associated with G [24]. The next theorem points out an important limitation of the hypothesis space $\mathcal{H}_G$ in the scenario of system identification.

Theorem 1: [25] Let $G(t,\tau) = \exp(-(t-\tau)^2/\sigma^2)$ on $\mathbb{R} \times \mathbb{R}$. Then $\mathcal{H}_G \not\subset L^1$ for any $\sigma > 0$. ■

Thus, even if $\mathcal{H}_G$ is an infinite-order Sobolev space of functions which are smooth, see [26], they are not necessarily absolutely integrable. In other words, the hypothesis space induced by the Gaussian kernel does not include any information on predictor stability. In stochastic terms, this is a consequence of the fact that G models f as a stationary process. Hence, its variance does not decay to zero as time progresses. This drawback typically prevents the impulse response estimator to define a BIBO stable linear system. This is described in [6] where a new kernel, named *stable spline kernel*, that overcomes the limitations of G by including the BIBO stability constraint, has been introduced.

We now move back to the main focus of this paper, i.e., the nonlinear system identification problem (4). Notice that, under the framework of statistical learning theory, input locations[3] now consist of the past inputs and outputs (y^t, u^t) while the training set is $\{(y^t, u^t), y_t\}$, for $t = 1 \ldots, N$. Thus, the dimension of the input space is infinite and the problem amounts to reconstructing the hypersurface $F : \mathbb{R}^{\infty \times 2} \mapsto \mathbb{R}$. In fact, F maps two sequences, made of past outputs and inputs, into the real line. Following the Gaussian regression approach, F is the realization from a Gaussian random field f with kernel

$$G : \mathbb{R}^{\infty \times 2} \times \mathbb{R}^{\infty \times 2} \mapsto \mathbb{R}.$$

In principle, the covariance of f can be still defined by the Gaussian kernel which, due to the nature of the input space, now assumes the following form

$$G(x,z) = \lambda \exp\left(-\frac{\sum_{i=1}^{\infty} \|x_i - z_i\|^2}{\sigma^2}\right). \tag{7}$$

Note that we have used x and z in (2) as two generic arguments of G. This kernel provides information on the smoothness of f, i.e., on the fact that the physical system F is expected to map "similar" input-output pairs into "similar" output values, similarity being measured in terms of Euclidean distance. However,

[2] This is often called prediction or innovation form.

[3] In statistical learning, the term input space indicates the domain of the unknown function which has to be learnt. The generic element of the input space is called input location.

significant limitations of this model can be pointed out also in this nonlinear scenario. First, G in (7) vanishes for most of the inputs due to the infinite-dimensional nature of the input space.[4] Furthermore, such a kernel is unable to discriminate the influence of the input locations on the system output, on the basis of their temporal locations. Instead, it would be desirable that the kernel included the information that, in physical systems, the effect of (y_k, u_k) over y_t decreases as $t-k$ goes to infinity. In the linear context the rate at which this happens is strictly related to BIBO stability of the predictor impulse response. In the non-linear scenario the link with "stability" is a much trickier point which we shall not discuss any further, see [27]. Our aim is to introduce a novel multi-dimensional kernel encoding this idea of "fading" memory in the predictor, extending to the nonlinear scenario the *stable spline kernel* introduced in [6].

B. A New Kernel for Nonlinear System Identification

In order to overcome the limitations discussed above we resort to mixtures of Gaussian kernels. In particular we shall assume that f can be modeled as the sum of independent zero mean Gaussian random fields f_ℓ, i.e., $f = \sum_{\ell=1}^{\infty} f_\ell$, so that

$$K(x,z;p) := \mathbb{E}[f(x)f(z)] = \sum_{\ell=1}^{\infty} K_\ell(x,z;p) \quad p \in \mathbb{N} \tag{8}$$

where we have still used x and z defined in (2) as two generic arguments of K and the kernels $K_\ell(x,z;p)$, $\ell = 1,\ldots,\infty$ are the covariances of the f_ℓ's and are taken of the form

$$\begin{aligned} K_\ell(x,z,p) &:= \mathbb{E}[f_\ell(x)f_\ell(z)] \\ &:= \beta_\ell \exp\left(-\frac{\sum_{i=1}^{p} \|x_{i+\ell-1} - z_{i+\ell-1}\|^2}{\sigma^2}\right) \\ & p \in \mathbb{N}, \sigma \in \mathbb{R}^+ \end{aligned} \tag{9}$$

with

$$\beta_\ell := \lambda_1 e^{-\ell\lambda_2}, \qquad \lambda_1, \lambda_2 \in \mathbb{R}^+. \tag{10}$$

From (8), (9) and (10) we see that the new kernel K depends on the hyperparameters σ, λ_1, λ_2 and p. In particular,

- σ denotes the kernel width;
- λ_1 and λ_2 define β_ℓ that models the influence of the past input-output pairs on y_t, ensuring that this dependence vanishes as ℓ increases[5];
- the integer p parametrizes the class of new kernels by establishing the maximum allowed order of interaction between past input-output data.

As an illustration of the role of p in the model, let us first consider the case $p = 1$; we obtain

$$\begin{aligned} &K(x,z;1) \\ &\quad = \beta_1 \exp\left(-\frac{\|x_1 - z_1\|^2}{\sigma^2}\right) \\ &\quad + \beta_2 \exp\left(-\frac{\|x_2 - z_2\|^2}{\sigma^2}\right) \\ &\quad + \beta_3 \exp\left(-\frac{\|x_3 - z_3\|^2}{\sigma^2}\right) + \cdots. \end{aligned} \tag{11}$$

Then, the random field f with covariance $K(x,z;1)$ admits the representation

$$f(y^t, u^t) = f_1(y_{t-1}, u_{t-1}) + f_2(y_{t-2}, u_{t-2}) + \cdots.$$

This reveals that the model associated with $p = 1$ is able to describe a system where only the nonlinear interactions between inputs and outputs at the same time instants are present. Instead, setting $p = 2$ the complexity of the model increases since

$$\begin{aligned} &K(x,z;2) \\ &\quad = \beta_1 \exp\left(-\frac{\|x_1 - z_1\|^2 + \|x_2 - z_2\|^2}{\sigma^2}\right) \\ &\quad + \beta_2 \exp\left(-\frac{\|x_2 - z_2\|^2 + \|x_3 - z_3\|^2}{\sigma^2}\right) + \cdots. \end{aligned} \tag{12}$$

Thus the random field f can be decomposed as:

$$\begin{aligned} f(y^t, u^t) = {} & f_1(y_{t-1}, u_{t-1}, y_{t-2}, u_{t-2}) \\ & + f_2(y_{t-2}, u_{t-2}, y_{t-3}, u_{t-3}) + \cdots \end{aligned}$$

where $K_\ell(\cdot,\cdot;2)$ is now the covariance of f_ℓ. This shows that $p = 2$ permits also nonlinear interactions between variables contiguous over time.

C. Accounting for Linear Components

Very often, in practical applications, the input output behavior is close to being linear; hence, in such situations, it might be advantageous to explicitly include in the model a linear component[6]. This can be done in the following manner: let us assume that

$$f(y^t, u^t) = \underbrace{q(y^t, u^t)}_{f_{nl}(y^t, u^t)} + \underbrace{(g \otimes y)(t) + (h \otimes u)(t)}_{f_{lin}(y^t, u^t)} \tag{13}$$

where the subscripts nl and lin stand, respectively, for *non linear* and *linear* and

- q is a zero-mean Gaussian random field, with kernel K given by (8), accounting for the nonlinear part of the model;
- g and h are impulse responses of causal systems, modeled as zero-mean Gaussian processes mutually independent and independent of q, describing the linear part of the system.

To fully define our model, we need to specify the kernels for g and h. One suitable choice is to employ the stable spline kernel, enriched with a parametric component (the so called *bias space*), as described in [6]. A more parsimonious choice which requires less unknown hyperparameters in the model, consists of interpreting the components of g and h as independent white noises with variance decaying to zero exponentially. More precisely, for $\xi_1, \xi_2 > 0$ and $\rho_1, \rho_2 > 0$, we have

$$g_k \sim \mathcal{N}(0, \xi_1 e^{-k\xi_2}), \quad g_i \perp g_j \quad i \neq j \tag{14}$$

$$h_k \sim \mathcal{N}(0, \rho_1 e^{-k\rho_2}), \quad h_i \perp h_j \quad i \neq j \tag{15}$$

[4] This problem would not be present if the kernel were defined over a finite-dimensional input space, e.g., in order to handle NARX models.

[5] Notice that, in the scenario of neural networks, automatic relevant determination is often used to detect and penalize inputs having few influence on the prediction using suitable priors on the network parameters, see, e.g., Section VII in [28]. Here, instead, input locations which are less influent on the system output are determined by β_ℓ that models the covariance of the Gaussian random field.

[6] Indeed, in Section VI we will show that the hypothesis space induced by (8) does not contain hyperplanes even if it can approximate them arbitrarily well in the uniform topology. This guarantees that the sum in (13) is direct.

and $h_k \perp g_h, \forall\, h, k \in \mathbb{N}$. Therefore, introducing the covariances (recall the definitions (1) and (2))[7]

$$\begin{aligned} L_1(x,z) &:= \mathbb{E}[(g\otimes y)(t)(g\otimes y)(\tau)] \\ L_2(x,z) &:= \mathbb{E}[(h\otimes u)(t)(h\otimes u)(\tau)] \end{aligned} \tag{16}$$

the covariance of the linear part $f_{lin}(t) = (g\otimes y)(t) + (h\otimes u)(t)$ is given by

$$L(x,z) := \mathbb{E}[f_{lin}(t)f_{lin}(\tau)] = L_1(x,z) + L_2(x,z). \tag{17}$$

Under the assumptions (14) and (15), the kernels L_1 and L_2 can be easily shown to have the form:

$$L_1(x,z) = \sum_{i=1}^{\infty} \xi_1 e^{-i\xi_2} y_{t-i} y_{\tau-i} \tag{18}$$

$$L_2(x,z) = \sum_{i=1}^{\infty} \rho_1 e^{-i\rho_2} u_{t-i} u_{\tau-i}. \tag{19}$$

Then, in view of (13) and our independence assumptions, f is a zero-mean Gaussian random field with covariance

$$\begin{aligned} R(x,z;p) &= \sum_{\ell=1}^{\infty} K_\ell(x,z;p) + L_1(x,z) + L_2(x,z) \\ &= K(x,z) + L(x,z). \end{aligned} \tag{20}$$

With respect to the model developed in the previous subsection, the additional unknown hyperparameters entering R are ξ_1, ξ_2, ρ_1 and ρ_2. Many simplifications are of course possible. For instance, recalling (10), one could set $\lambda_2 = \xi_2 = \rho_2$.

IV. Bayesian Model for Nonlinear System Identification

Let θ denote the vector whose components are the noise standard deviation and all the hyperparameters describing the kernel except for p. More specifically $\theta = [\eta, \sigma, \lambda_1, \lambda_2]$ when the covariance of f is K in (8) while $\theta = [\eta, \sigma, \lambda_1, \lambda_2, \xi_1, \xi_2, \rho_1, \rho_2]$ when the full kernel R in (20) is used. In the sequel, all the densities that will be reported are conditional on the system input, but we omit this dependence to simplify the notation.

Since we adopt a fully Bayesian viewpoint, we also interpret θ as a random vector with mutually independent components. It is assigned an improper prior, that does not depend on p, which has the only purpose of ensuring nonnegativity of its components. The parameter p is also modeled as a random variable, in one-to-one correspondence with equiprobable competitive models. In particular, it is assigned a poorly informative prior on $[1, 2, \ldots, \nu]$ with ν arbitrarily large. Furthermore, θ, p and f are mutually independent. Hereafter, the dependence on the input u is omitted to simplify the notation.

In practice, y^- is never completely available. One solution is to set its unknown components to zero. This is similar in spirit to the methods employed to deal with initial effects when estimating linear parametric predictors, see, e.g., Section 3.2 in [2]. In view of this, it is convenient to approximate $\mathbf{p}(f,\theta,p|y^-)$ with $\mathbf{p}(f,\theta,p)$ so that the imperfect knowledge on y^- does not influence the prior on f, θ and p. In particular, in all the subsequent derivations, the following approximation for the joint density of y^+, f, θ and p

$$\mathbf{p}(y^+, f, \theta, p|y^-) \approx \mathbf{p}(y^+|f,\theta,p,y^-)\mathbf{p}(f|\theta,p)\mathbf{p}(\theta,p) \tag{21}$$

will be thought of as a perfect equality.

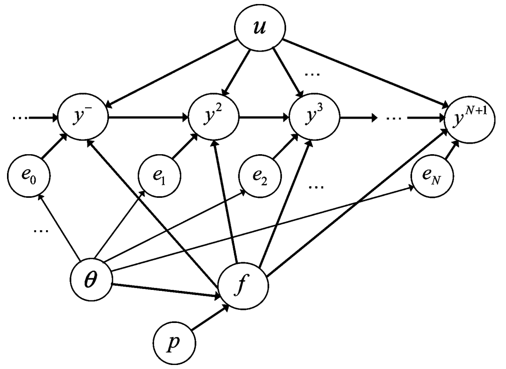

Fig. 2. Bayesian network describing the stochastic model for nonlinear system identification proposed in this paper. In the network, u is the system input, e is the innovations sequence and y^t are the output samples up to time $t-1$. In addition, f is a zero-mean Gaussian random field, the nonlinear map F in (4) being one realization from f. The covariance of f is either the kernel K in (8) or the kernel R in (20) when F is known to be close to linear. Finally, p defines the complexity of the kernel and θ contains the hyperparameters, i.e., the kernel parameters and the innovation variance η^2.

Our stochastic model for nonlinear system identification is graphically described by the Bayesian network in Fig. 2.

Remark 2: It is worth stressing that all the outcomes obtained in the sequel would still hold even if u were thought of as a deterministic signal or if output feedback were present in the system. In Fig. 2, modeling u as a stochastic process independent of $\{e_t\}$ is just a way to simplify the notation needed in the derivation of the results described in Sections V–VIII.

V. Nonparametric Algorithm for Nonlinear System Identification

In what follows, the dependence of certain formulas on y^- is omitted to further simplify the notation. Our purpose here is to obtain an estimator $\hat{f}$ of the nonlinear map in (4) from the data u^N, y^+ and y^-. However, as should be clear from the discussion in the previous paragraphs, the prior distribution for f as well as the likelihood function depend upon the unknown parameters θ and p which, in practice, are unknown. To introduce a numerical scheme to infer them from data, we first observe that two important quantities can be obtained in closed form using the Bayesian model in Fig. 2. The first one is the minimum variance estimator of f for known y^+, θ and p, i.e., $\mathbb{E}[f(x)|y^+,\theta,p]$ with x a generic input location. The second one is the marginal likelihood of the data y^+, i.e., the joint density of f, y^+, θ and p, where f is integrated out. This is illustrated in the following proposition whose proof is reported in Appendix.

Proposition 1: If the approximation (21) holds and the covariance of f is R, then

$$\hat{f}(x) = \mathbb{E}[f(x)|y^+,\theta,p] = \sum_{i=1}^{N} c_i R(x,(y^i,u^i);p) \tag{22}$$

where x is a generic input location, c_i is the i-th component of the vector

$$c = (\Sigma_y(p,\theta))^{-1} y^+ \tag{23}$$

[7] We remind the reader that x and z, and hence also y and u in (16), play the role of "input locations". Therefore y and u are fixed values at which the functions $f_{lin,g}(y^t) = (g\otimes y)(t)$ and $f_{lin,h}(u^t) = (h\otimes u)(t)$ are evaluated; expected values are taken w.r.t. g and h.

and $\Sigma_y \in \mathbb{R}^{N\times N}$ is invertible with (i,j)-entry given by

$$[\Sigma_y]_{i,j} = R((y^i,u^i),(y^j,u^j);p) + \eta^2\delta_{ij} \tag{24}$$

where δ_{ij} is the Kronecker delta. In addition

$$\mathbf{p}(y^+|\theta,p) = \frac{\exp\left(-\frac{1}{2}(y^+)^T(\Sigma_y(p,\theta))^{-1}y^+\right)}{\sqrt{\det(2\pi\Sigma_y(p,\theta))}} \tag{25}$$

A. Estimating Model Structure and Hyper-Parameters

According to the empirical Bayes paradigm, the "optimal" value of p maximizes the model posterior probability. Since $\mathbf{p}(y^+,f,\theta,p) = \mathbf{p}(y^+,f|\theta,p)\mathbf{p}(\theta,p)$ and $\mathbf{p}(\theta,p) = \mathbf{p}(\theta)\mathbf{p}(p)$, using Bayes rule we have

$$\mathbf{p}(p|y^+) \propto \int \mathbf{p}(y^+,f|\theta,p)\mathbf{p}(\theta)\mathbf{p}(p)df\,d\theta. \tag{26}$$

The integration with respect to f, i.e., $\mathbf{p}(y^+|\theta,p)$, was obtained in (25). Instead, integration with respect to θ, although numerically feasible using e.g., stochastic simulation techniques [29], can be computationally expensive. To this aim, define

$$\hat{\theta}_p = \arg\min_\theta J_p(\theta) \tag{27}$$

where

$$J_p(\theta) := -\log\mathbf{p}(y^+|\theta,p). \tag{28}$$

Notice that $J_p(\hat{\theta}_p)$ provides an approximation of the minus log of the objective (26) which neglects only the uncertainty relative to the estimator of θ; in fact, marginalization w.r.t. f allows to account for the effect of different choices of p on the uncertainty in estimating f.

This approximation leads to the following estimator of the kernel structure

$$\hat{p} = \arg\min_p J_p(\hat{\theta}_p) \tag{29}$$

while the estimator of the hyperparameter vector is $\hat{\theta}_{\hat{p}}$.

These estimators are used, according to the empirical Bayes paradigm, to obtain the estimator $\hat{f}(x)$ in Proposition 1 by substituting the unknown θ and p with $\hat{\theta}_{\hat{p}}$ and $\hat{p}$.

B. The Algorithm for Nonlinear System Identification

Our nonparametric algorithm for nonlinear system identification is summarized below.

Algorithm for Nonlinear System Identification: The input to this algorithm includes the available input-output pairs, i.e., u^N together with y^+ and y^-. The output is the estimator of the nonlinear predictor $\hat{f}$. The main steps are as follows:

(i) Determine the kernel structure, i.e., the value of p which minimizes (29).

(ii) Set the hyperparameters to the components of $\hat{\theta}_{\hat{p}}$ according to (27), (28), (29)

(iii) For any input location x, define the prediction at x as

$$\hat{f}(x) = \sum_{i=1}^N c_i R(x,(y^i,u^i);\hat{p}) \tag{30}$$

where kernel hyperparameters are the components of $\hat{\theta}_{\hat{p}}$ and $\{c_i\}$ are computed solving (23), (24).

VI. Characterization of the RKHS Induced by the Mixture of Gaussian Kernels

Using the representer theorem and the correspondence between Gaussian processes and RKHS, see, e.g., [24], one obtains that $\hat{f}(\cdot)$ in (30) is the solution of the following Tikhonov-type variational problem:

$$\arg\min_{r\in\mathcal{H}_R} \sum_{i=1}^N \left(y_i - r(y^i,u^i)\right)^2 + \eta^2\|r\|^2_{\mathcal{H}_R} \tag{31}$$

where $\mathcal{H}_R$ is the RKHS associated with R. Therefore, in order to gain information on the properties of our estimator, it is essential to characterize the space $\mathcal{H}_R$.

Recall that, given a kernel V, $\mathcal{H}_V$ is the Hilbert space of functions which are the completion, w.r.t. the inner product

$$\left\langle \sum_i m_i V(\cdot,t_i), \sum_j n_j V(\cdot,s_j)\right\rangle_{\mathcal{H}_V} = \sum_{i,j} m_i n_j V(t_i,s_j), \tag{32}$$

of the manifolds given by all the finite linear combinations $\sum_{i=1}^l m_i V(\cdot,t_i)$ for all choices of l, $\{m_i\}$ and $\{t_i\}$. Let $\mathcal{H}_R$, $\mathcal{H}_K$ and $\mathcal{H}_L$ be the RKHSs induced by the kernels R, K and L, respectively. Since R is the sum of K and L, functions in $\mathcal{H}_R$ are sums of functions in $\mathcal{H}_K$ and $\mathcal{H}_L$, see pag. 353 in [21], so that $\mathcal{H}_K \subset \mathcal{H}_R$ and $\mathcal{H}_L \subset \mathcal{H}_R$. In view of the definition of L in (17), it is rather simple to see that the elements of $\mathcal{H}_L$ are only linear functionals, i.e., hyperplanes. The aim is now to characterize the hypothesis space $\mathcal{H}_K$ associated with the kernel K in (8). Among other things, in Section VII we will show that $\mathcal{H}_K$ and $\mathcal{H}_L$ have not any function besides zero in common, so that they are complementary closed subspaces in $\mathcal{H}_R$.

A. Characterization of the RKHS $\mathcal{H}_K$

To simplify the exposition, in this section we assume that the kernel $K(x,z)$ is composed by a finite number n of mixtures with $x := [x_1,\ldots,x_n]^T \in \mathbb{R}^n$ and $z := [z_1,\ldots,z_n]^T \in \mathbb{R}^n$. Under this assumption, for $p \in [1,2,\ldots,n]$, we have

$$K(x,z;n,p) = \sum_{j=1}^{n-p+1} \beta_j \exp\left(-\frac{\sum_{i=1}^p (x_{i+j-1} - z_{i+j-1})^2}{\sigma^2}\right). \tag{33}$$

Notice that for $p = n$ we obtain the classical Gaussian kernel on $\mathbb{R}^n$. Let $X \subset \mathbb{R}^n$ with nonempty interior. In this section we will describe the RKHS $\mathcal{H}_K$ of functions on X induced by $K(x,z;n,p)$: we will obtain (i) an explicit orthonormal basis for $\mathcal{H}_K$ and (ii) an explicit expression for the inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}_K}$. This allows to compute explicitly the norm in $\mathcal{H}_K$ which enters in the Tikhonov-type variational problem (31). In order to do so we first need to set up some notation:

Notation 1: Let us define the multi-index $\alpha := (\alpha_1,\ldots,\alpha_n) \in (\mathbb{N}\cup\{0\})^n$ with $|\alpha| := \sum_{j=1}^n \alpha_j$, the monomials $x^\alpha = x_1^{\alpha_1}\ldots x_n^{\alpha_n}$, and the multinomial coefficients $C_\alpha = |\alpha|!/\alpha_1!\ldots\alpha_n!$. We also define $(j) = (j,\ldots,j+p-1)$ and the j-th component kernel as

$$K_j(x,z;n,p) = e^{-\|x_{(j)}-z_{(j)}\|^2/\sigma^2}$$

so that

$$K(x,z;n,p) = \sum_{j=1}^{n-p+1} \beta_j K_j(x,z;n,p)$$

■

We now study in some detail the properties of the RKHSs induced by the kernels $K_j(x,z;n,p)$ providing an explicit orthonormal basis for these spaces as well as an explicit expression for the inner product. The following result can be derived from the description of the RKHS associated with the Gaussian kernel which is given in [30]; see also Section III-C in [25] for a novel and short proof which exploits the concept of the Weyl inner product. It characterizes the RKHS $\mathcal{H}_{K_j}$ induced by K_j whose functions depend on the argument $x_{(j)}$.

Theorem 2: Let $X \subset \mathbb{R}^n$ be any set with non-empty interior. Then, $\dim(\mathcal{H}_{K_j}) = \infty$ and

$$\mathcal{H}_{K_j} = \left\{ r = e^{-\|x_{(j)}\|^2/\sigma^2} \sum_{|\alpha_{(j)}|=0}^{\infty} w_{\alpha_{(j)}} x_{(j)}^{\alpha_{(j)}} : \right.$$
$$\left. \|r\|_{K_j}^2 = \sum_{k=0}^{\infty} \frac{k!}{\left(\frac{2}{\sigma^2}\right)^k} \sum_{|\alpha_{(j)}|=k} \frac{w_{\alpha_{(j)}}^2}{C_{\alpha_{(j)}}} < \infty \right\}.$$

For r, $v \in \mathcal{H}_{K_j}$ given by

$$r(x) = e^{-\|x_{(j)}\|^2/\sigma^2} \sum_{|\alpha_{(j)}|=0}^{\infty} w_{\alpha_{(j)}} x_{(j)}^{\alpha_{(j)}}$$
$$v(x) = e^{-\|x_{(j)}\|^2/\sigma^2} \sum_{|\alpha_{(j)}|=0}^{\infty} c_{\alpha_{(j)}} x_{(j)}^{\alpha_{(j)}}$$

the inner product $\langle \cdot, \cdot \rangle_{H_{K_j}}$ in $\mathcal{H}_{K_j}$ is

$$\langle r, v \rangle_{H_{K_j}} = \sum_{k=0}^{\infty} \frac{k!}{\left(\frac{2}{\sigma^2}\right)^k} \sum_{|\alpha_{(j)}|=k} \frac{w_{\alpha_{(j)}} c_{\alpha_{(j)}}}{C_{\alpha_{(j)}}}. \tag{34}$$

An orthonormal basis for $\mathcal{H}_{K_j}$ is

$$\left\{ \psi_{\alpha_{(j)}}^j(x) = \sqrt{\frac{\left(\frac{2}{\sigma^2}\right)^{|\alpha_{(j)}|} C_{\alpha_{(j)}}}{|\alpha_{(j)}|!}} e^{-\|x_{(j)}\|^2/\sigma^2} x_{(j)}^{\alpha_{(j)}} \right\}_{|\alpha_{(j)}|=0}^{\infty}. \tag{35}$$

■

For our purposes, we first need to obtain a new property of the RKHS induced by the single Gaussian kernel $K(x,z) = \exp(-\|x-z\|^2/\sigma^2)$. This is described in the following theorem, see Appendix for the proof.

Theorem 3: Let $X \subset \mathbb{R}^n$ be any set with non-empty interior. Let $K(x,z) = \exp(-\|x-z\|^2/\sigma^2)$. Then $\mathcal{H}_K$ does not contain any monomial on X, including the nonzero constant function. ■

This result guarantees that $\mathcal{H}_K \cap \mathcal{H}_L = \emptyset$ since hyperplanes are not contained in the space $\mathcal{H}_K$; therefore the sum in (13) is direct.

The next result, whose proof is reported in Appendix, provides the desired characterization of $\mathcal{H}_K$. It shows that the RKHSs $\mathcal{H}_{K_j}$ associated to the components $K_j(x,z;n,p)$ of the mixture (33) satisfy $\mathcal{H}_{K_j} \cap \mathcal{H}_{K_i} = \emptyset$, $i \neq j$. As a consequence $\mathcal{H}_K$ is the direct orthogonal sum of subspaces $\mathcal{H}_{K_j}$. Note that orthogonality follows from the definition of the kernel K as the sum of the kernels K_i [21].

Theorem 4: Let $X \subset \mathbb{R}^n$ be any set with non-empty interior. Let $\mathcal{H}_{K_j}$ be as in Theorem 2. Then the Hilbert space $\mathcal{H}_K$ induced by $K(x,z;n,p)$ is

$$\mathcal{H}_K = \oplus_{j=1}^{n-p+1} \mathcal{H}_{K_j}, \tag{36}$$

that is each $r \in \mathcal{H}_K$ admits a unique orthogonal decomposition $r = \sum_{j=1}^{n-p+1} r^j$, with $r^j \in \mathcal{H}_{K_j}$. For $v = \sum_{j=1}^{n-p+1} v^j$, where $v^j \in \mathcal{H}_{K_j}$, the inner product in $\mathcal{H}_K$ is

$$\langle r, v \rangle_{\mathcal{H}_K} = \sum_{j=1}^{n-p+1} \langle r^j, v^j \rangle_{H_{K_j}},$$

so that the norm in $\mathcal{H}_K$ is

$$\|r\|_{H_K}^2 = \sum_{j=1}^{n-p+1} \|r^j\|_{H_{K_j}}^2 < \infty.$$

An orthonormal basis of $\mathcal{H}_K$ is then

$$\left\{ \psi_{\alpha_{(j)}}^j(x) \right\} \quad j = 1, \ldots, n-p+1 \quad |\alpha_{(j)}| = 0, \ldots, \infty \tag{37}$$

where $\psi_{\alpha_{(j)}}^j(x)$ is defined in (35).

B. Approximation Properties of $\mathcal{H}_K$

The following result, whose proof is reported in Appendix, relies upon the characterization of $\mathcal{H}_K$ reported in the previous subsection and the well known universality[8] of the classical Gaussian kernel, see, e.g., [30]. It clarifies the role played by the kernel hyperparameter p in establishing the complexity of the hypothesis space.

Theorem 5: Denote with X any compact subset of $\mathbb{R}^n$. For any $p < n$, the kernel $K(x,z;n,p)$ is not universal. However, for any function r in the space of continuous functions C_0, there exists $p_0 \in \mathbb{N}$, with $p_0 \leq n$, such that if $p_0 \leq p \leq n$, functions in the space $\mathcal{H}_K$ associated with $K(x,z;n,p)$ can approximate arbitrarily well r under the supremum norm. ■

To gain further insight on the proposed hypothesis space, first note that Theorem 3 and Theorem 4 show that the kernels $\{K_i\}$, L_1 and L_2 in (20) induce mutually orthogonal subspaces. Recall also that our predictor of the system output at a generic input location x is

$$\hat{f}(x) = \sum_{i=1}^{N} c_i R(x, (y^i, u^i); \hat{p})$$
$$= \sum_{i=1}^{N} c_i \left(\sum_{\ell=1}^{\infty} K_\ell(x, (y^i, u^i); \hat{p}) + L(x, (y^i, u^i)) \right).$$

Hence, we can define the components

$$\hat{f}_{K_\ell}(x) := \sum_{i=1}^{N} c_i K_\ell(x, (y^i, u^i); \hat{p}) \tag{38}$$

and

$$\hat{f}_L(x) := \sum_{i=1}^{N} c_i L(x, (y^i, u^i)) \tag{39}$$

[8] A kernel K is said *universal* if, for any $\epsilon > 0$ and $r_1 \in C_0$, there exists one $r_2 \in \mathcal{H}_K$ such that $\|r_1 - r_2\|_{C_0} \leq \epsilon$.

which are, respectively, the orthogonal projections of the estimator $\hat{f}(\cdot) \in \mathcal{H}_R$ onto $\mathcal{H}_{K_\ell}$ and $\mathcal{H}_L$. In particular, $\hat{f}_L$ represents that part of the system output which is due to the sole linear component of the model.

Estimators for the impulse responses g and h in (13) can be also promptly recovered as described in the next proposition whose proof is discussed in Appendix.

Proposition 2: Define

$$Y = [\, y^1 \quad y^2 \quad \dots \quad y^N \,], \quad U = [\, u^1 \quad u^2 \quad \dots \quad u^N \,] \tag{40}$$

and use $\Lambda_g, \Lambda_h \in \mathbb{R}^{\infty \times \infty}$ to denote the covariances of g and h, respectively, that, in view of (14) and (15), are diagonal and defined by

$$[\Lambda_g]_{k,k} = \xi_1 e^{-k\xi_2}, \quad [\Lambda_h]_{k,k} = \rho_1 e^{-k\rho_2}. \tag{41}$$

Then, the estimates of g and h are

$$\hat{g} = \Lambda_g Y c, \quad \hat{h} = \Lambda_h U c \tag{42}$$

where c has been defined in (23). ■

Estimators like that reported in (31) enjoy important consistency properties. In fact, they are able to reconstruct consistently, e.g., in the topology of the RKHS used as hypothesis space, a wide class of functions, dense in the space of continuous functions [31]. Hence, if (13) and some technical assumptions hold, our estimator $\hat{f}$ will converge to the true system F just making the scale factor of the kernel go to zero with a certain rate, as N goes to ∞, see [31] for details. But this, combined with the fact that $\mathcal{H}_L$ and $\{\mathcal{H}_{K_\ell}\}$ do not share any function, allows us also to conclude that $\hat{f}_L$ will reproduce asymptotically the entire contribution to the output due to the linear part of the system. This will permit also to reconstruct consistently g and h by (42). In fact, all the linear part of the model must be captured asymptotically by $\mathcal{H}_L$ since $\mathcal{H}_K$ does not contain any hyperplane. In practice, data set size is always finite. Nevertheless, in real applications we expect this property to greatly help in distinguishing the linear part of the system from the nonlinear one. A numerical example will be illustrated in Section VII-B.

VII. Numerical Experiments

A. Monte Carlo Studies Involving Six Nonlinear Models

The proposed approach is tested on 6 benchmark problems taken from [9], [16], [32] and listed in Table I. For each system in Table I we perform a simulation study of 100 runs as follows: the unknown system has to be reconstructed starting from 200 or 400 data points generated considering the system initially at rest. The input u fed in the systems 4, 5 and 6 is zero mean, unit variance white Gaussian noise.

We test the new nonparametric identification algorithm with the kernel K defined in (33). For computational reasons, we set $n = 20$, a value that has just to be large enough to capture the dynamics of the predictor and does not need to establish any kind of bias-variance trade-off. When considering the time series generated by systems 1, 2 and 3, the dependence of K on u is removed. Analogously, when system 6 is under study, the information that the optimal predictor depends only on past inputs is provided to the estimator. The hyperparameter vector θ and the structure parameter p are then determined from data by the numerical procedure reported in Section V-B.

TABLE I
Monte Carlo Studies (Section VII-A): The Six Nonlinear Models to be Identified

$$\begin{array}{llcl}
(1) & y_t & = & 2e^{-0.1 y_{t-1} y_{t-1}^2} - e^{-0.1 y_{t-2} y_{t-2}^2} + e_t \\
 & e_t & \sim & \mathbf{N}(0,1) \\
(2) & y_t & = & e^{-0.1 y_{t-1}^2} (2y_{t-1} - y_{t-2}) + e_t \\
 & e_t & \sim & \mathbf{N}(0,1) \\
(3) & y_t & = & -2y_{t-1} I(y_{t-1} < 0) + 0.4 y_{t-1} I(y_{t-1} \geq 0) + e_t \\
 & e_t & \sim & \mathbf{N}(0,1) \\
(4) & y_t & = & 0.5 y_{t-1} - 0.05 y_{t-2}^2 + u_{t-1}^2 + 0.8 u_{t-2} + e_t \\
 & e_t & \sim & \mathbf{N}(0, 0.22^2) \\
(5) & y_t & = & 0.8 y_{t-1} + u_{t-1} - 0.3 u_{t-1}^3 + 0.25 u_{t-1} u_{t-2} \\
 & & - & 0.3 u_{t-2} + 0.24 u_{t-2}^3 - 0.2 u_{t-2} u_{t-3} - 0.4 u_{t-3} + e_t \\
 & e_t & \sim & \mathbf{N}(0, 0.14^2) \\
(6) & y_t & = & u_{t-1} + 0.6 u_{t-2} + 0.35 u_{t-3} + 0.9 u_{t-4} + 0.35 u_{t-5} \\
 & & + & 0.2 u_{t-6} + 0.2 u_{t-7} + 0.5 u_{t-1}^2 + 0.25 u_{t-1} u_{t-2} \\
 & & + & 0.5 u_{t-1} u_{t-3} - u_{t-2} u_{t-3} + 0.75 u_{t-3}^3 \\
 & & + & 0.5 u_{t-2} u_{t-4} - 0.25 u_{t-4}^2 + e_t \\
 & e_t & \sim & \mathbf{N}(0,1)
\end{array}$$

The performance measure is prediction capability on test data, generated using zero mean, unit variance white noise as input. In particular, for each Monte Carlo run, after obtaining $\hat{f}$, we generate a test set of 500 new data denoted by $\{y_t^{test}\}_{t=1}^{500}$ and (where needed) $\{u_t^{test}\}_{t=1}^{500}$, respectively. Then, the prediction error at the j-th Monte Carlo run is computed as follows

$$err_j = \sqrt{\frac{\sum_{t=1}^{500} \left(\hat{ya}_t^{test} - y_t^{test} \right)^2}{500}} \tag{43}$$

$$\hat{y}_t^{test} = \hat{f}(y^{t,test}, u^{t,test}) \tag{44}$$

where $(u^{t,test}, y^{t,test})$ is the test set up to time $t-1$.

The six panels of Fig. 3 display the boxplots of $\{err_j\}$ for the six case studies; the dashed line denotes the best achievable expected prediction error, i.e., the innovation standard deviation. It is apparent that in all examples the proposed nonparametric estimator performs reasonably well. The prediction capability on new data is close to that obtainable from the optimal (nominal) predictor also when the training set size is 200. The results reveal that the proposed estimator performs well also when the sole kernel K is used and data are generated by polynomial models. In fact, even if $\mathcal{H}_K$ does not contain any monomial, see Theorem 3, it is however sufficiently rich to approximate every continuous function, see Theorem 5.

B. Identification of a Nonlinear System Using the Kernel With the Linear Component

Consider the following nonlinear system

$$\begin{aligned} y_t &= f_a(u_{t-1}) + f_b(u_{t-2}) + (h \otimes u)(t) + e_t \\ &:= \sin(2u_{t-1}) + \sin(u_{t-2}) + (h \otimes u)(t) + e_t \end{aligned} \tag{45}$$

where innovation variance is 1 and the coefficients of the impulse response h are displayed in the top panel of Fig. 4 (points connected by dashed line); the values h_k are samples from a zero mean white noise with variance $9/k^2$, $k \in \mathbb{N}$. We are interested in estimating the system from 400 input-output measurements

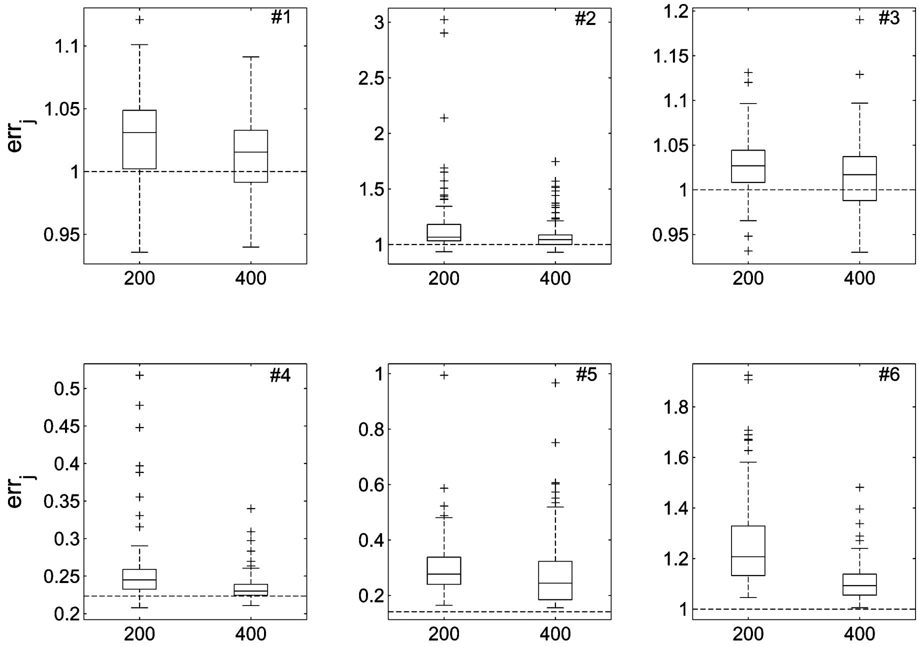

Fig. 3. Monte Carlo studies (Section VII-A): boxplot of prediction errors. In each panel the dashed line denotes the standard deviation of the innovation, i.e., the best achievable expected prediction error.

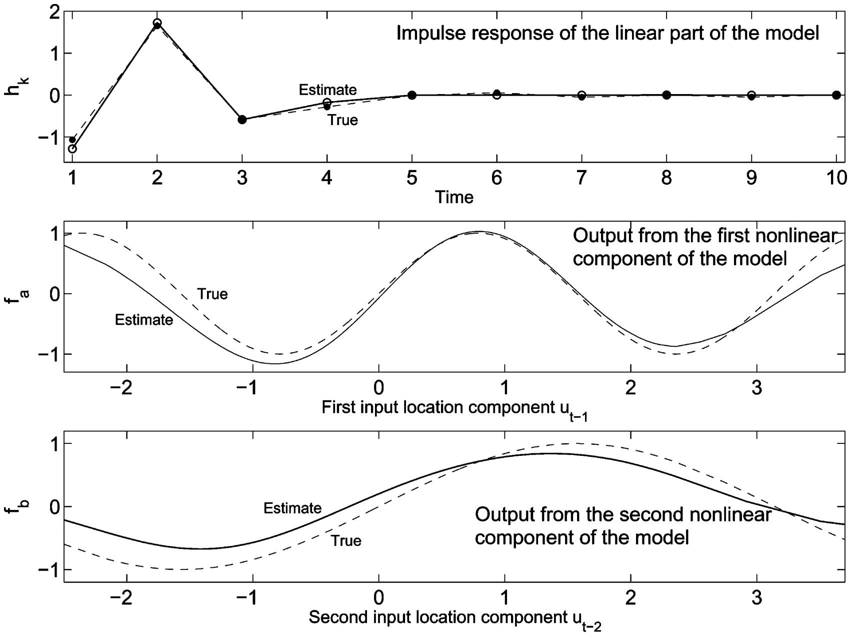

Fig. 4. Simulated case study (Section VII-B). *Top*: true (dashed line) and estimated (solid line) impulse response $\{h_k\}$ (linear part of the system). *Middle and Bottom*: true (dashed) and estimated (solid line) f_a and f_b, i.e., output from the first and second nonlinear part of the system as a function of some input location values, see (45).

obtained by applying to the system at rest a zero mean, unit variance white noise as forcing input. The data points are plotted in Fig. 5 (○ in the top panel).

We use the model (13), with $g = 0$, q known to depend only on past inputs and the covariance of f given by the kernel K in (33) with $n = 20$. The identification procedure described in Section V-B is used. The function $J_p(\hat{\theta}_p)$, defined in (27), (28), (29), which is inversely proportional to the posterior model probability, is displayed in Fig. 6 as a function of the kernel hyperparameter p. Notice that the Bayesian approach suggests that the system is likely to exhibit additive nonlinearities, each depending just on one single component of the input ($\hat{p} = 1$). In Fig. 5 (top panel) the solid line is the output from the estimated one-step ahead predictor. The bottom panel of the same figure depicts the first 100 one-step ahead prediction errors on the training set using the optimal $(+)$ and the estimated $(\times)$ predictor, suggesting that the proposed approach introduces a right amount of regularization in the estimation process. In the three panels of Fig. 4 the solid lines are the estimates of h, computed using (42), and of f_a, f_b, obtained using (38), all of which are close to ground-truth. Denoting by $\hat{h}$ the estimator of h, the relative error $\|h - \hat{h}\|/\|h\|$ turns out to be 0.15.

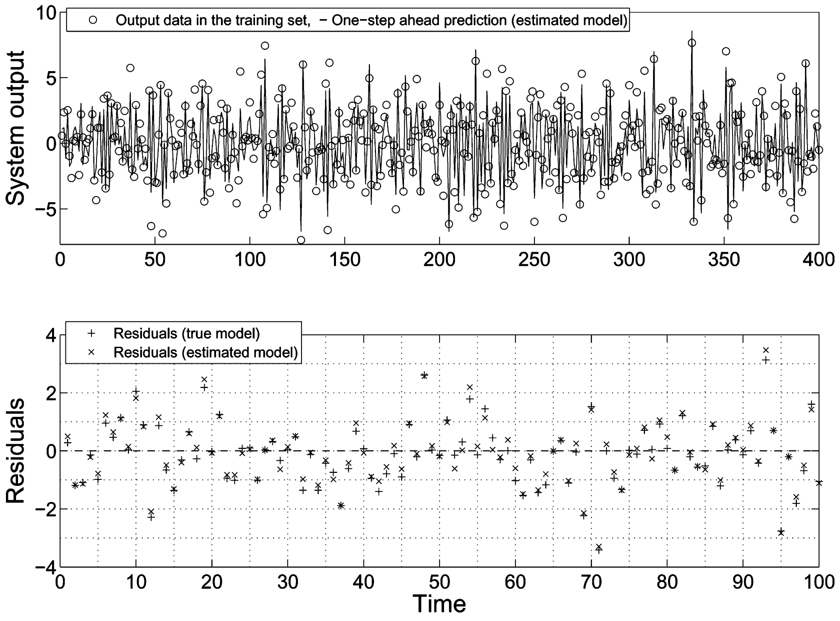

Fig. 5. Simulated case study (Section VII-B). *Top*: system output values contained in the training set (○) and (linearly interpolated) output prediction using the estimated model (solid line). *Bottom*: first 100 residuals (one-step ahead prediction errors) using the optimal (+) and the estimated (×) model.

TABLE II
MONTE CARLO STUDY (SECTION VII-B): PERCENTILES OF THE VALUES $\{err_j\}$, $j = 1, \dots, 100$, COMPUTED VIA (46)

Percentiles	5th	10th	20th	25th	50th	75th	80th	90th	95th
Relative error	0.04	0.063	0.078	0.08	0.15	0.26	0.37	0.75	0.86

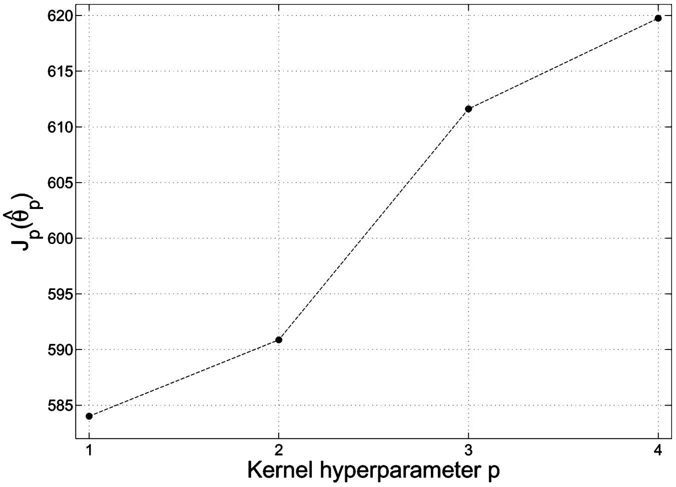

Fig. 6. Simulated case study (Section VII-B): plot of $J_p(\hat{\theta}_p)$ (inversely proportional to the posterior model probability, see (27), (28), (29)) as a function of kernel hyperparameter p.

Next we consider a simulation study of 100 runs in which the same experiment described above is repeated by considering different realizations of h. More precisely, at the j-th run a new impulse response $h^{(j)}$ is generated by drawing its components $h_k^{(j)}$ from a white noise with variance $9/k^2$. The estimator of $h^{(j)}$ and the relative error at the j-th run are indicated, respectively, by $\hat{h}^{(j)}$ and err_j, where

$$err_j = \frac{\|h^{(j)} - \hat{h}^{(j)}\|}{\|h^{(j)}\|} \tag{46}$$

Table II reports some percentiles of the values of $\{err_j\}$. The results show that the estimator equipped with the full kernel R (given by the sum of the kernels L and K) provides good performance in reconstructing $\{h^{(j)}\}$, and thus also in separating the effect of the linear part of the system (captured by $\mathcal{H}_L$) from that due to the nonlinear one (captured by $\mathcal{H}_K$). In particular, the mean and the median of the relative errors are close to 0.25 and 0.15, respectively.

C. Comparison With Direct Weight Optimization

In this subsection we compare the performance of the approach presented in this paper with an algorithm based on Direct Weight Optimization (DWO) [17] and equipped with the minimal probability approach proposed in [18]. First, it is useful to recall that DWO estimates the predictor function at a particular input location using a weighted linear combination of the observed outputs. In particular, the algorithm depends on a parameter, denoted by δ in [18], that establishes which input/output data in a neighborhood of the input location x must be used to estimate $f(x)$. Note that in [18] no specific criteria are suggested to choose δ and to determine the number of regressors to be introduced in the algorithm.

Now, we consider Example 3 in [18]. The nonlinear system is

$$y_t = -0.2y_{t-1}^3 + u_{t-1} + e_t \tag{47}$$

with the variance of e_t equal to 0.0223. The following two case studies are introduced:

- Experiment #1. The experimental conditions are the same as those described in Example 3 in [18], i.e., the training and test sets consist of 1000 and 100 data points, respectively, generated using the following input:

$$u_t = \sin\left(\frac{2\pi t}{10}\right) + \sin\left(\frac{2\pi t}{25}\right), \qquad t = 1, \dots, 1100 \tag{48}$$

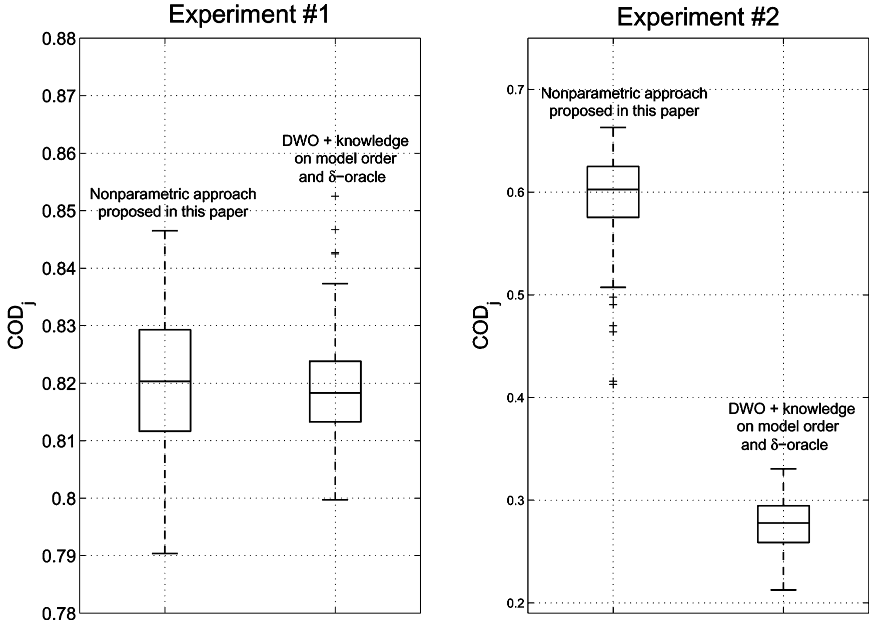

Fig. 7. Monte Carlo studies (Section VII-C): boxplot of $\{COD_j\}_{j=1}^{100}$ using the new nonparametric approach and using DWO with oracle and correct model order. The larger COD_j, the better is the performance of the estimator.

- Experiment #2. The size of the training set is reduced to 200 while that of the test set is 300. In addition, prediction on new data is more difficult since the training and test sets are generated using two different inputs, i.e., a white normal noise of SD 0.2 and 0.4, respectively.

For each experiment, a simulation study of 100 runs is considered. At each run, after generating the test set $\{y_t^{test}\}_{t=1}^{n_t}$ and $\{u_t^{test}\}_{t=1}^{n_t}$, we compute the coefficient of determination as follows:

$$COD_j = 1 - \sqrt{\frac{\sum_{t=1}^{n_t} \left(\hat{y}_t^{test} - y_t^{test}\right)^2}{\sum_{t=1}^{n_t} \left(\bar{y}_t^{test} - y_t^{test}\right)^2}}$$

$$\hat{y}_t^{test} = \hat{f}(y_{t-1}^{test}, u_{t-1}^{test}), \quad \bar{y}_t^{test} = \frac{\sum_{t=1}^{n_t} y_t^{test}}{n_t}. \tag{49}$$

Then, we obtain

$$\overline{COD} = \frac{1}{100} \sum_{j=1}^{100} COD_j. \tag{50}$$

This index quantifies how much of the output variance is captured by an estimator; the larger, the better is the performance of the estimator $\hat{f}$.

At each run, $\hat{f}$ is obtained by adopting the following two estimators:

- K: this is the new nonparametric approach that exploits the kernel K reported in (33) with $n = 10$. All the hyperparameters, including p, are estimated from data via marginal likelihood optimization. Then, $\hat{f}$ is obtained by (22);
- DWO: this is the minimal probability approach described in [18], equipped with the information that $\hat{y}_t$ depends only on y_{t-1} and u_{t-1} and with what we call a δ-oracle. More precisely, at each Monte Carlo run the δ-oracle determines that value of δ that maximizes COD_j using a fine grid of step 0.02 on $[0, 2]$. This results in an ideal tuning that provides the best possible performance obtainable by the minimal probability estimator.

Fig. 7 displays the boxplots of $\{COD_j\}_{j=1}^{100}$ obtained by the two estimators. In Experiment #1 (left panel) $\overline{COD}$ obtained by K and DWO is 0.82 and 0.819, respectively. In Experiment #2 (right panel) $\overline{COD}$ achieved by K and DWO is 0.61 and 0.28, respectively. Thus, remarkably, the proposed nonparametric estimator (whose parameters are all learnt from data) performs either in the same way or much better than DWO equipped with the oracle and the exact model order.

The above results could appear surprising. To better understand them, two comments are now in order:

- the DWO-based minimal probability estimator uses weights that are optimal for any finite number of data points but only in terms of minimizing an upper bound on the prediction error. In other words, the theory underlying DWO and the resulting algorithmic architecture rely upon an inequality on the generalization error that may be conservative. Conversely, the nonparametric scheme presented here does not adopt a conservative point of view. In fact, the bias/variance trade-off is established by hyperparameters estimation using the marginal likelihood (25) that is exact and depends on the specific outputs coming from the system under study. This defines an algorithmic architecture more suited to the specific training set at hand and hence possibly more robust also than DWO $+$ δ-oracle;
- Equation (22) shows that the estimate obtained by K always consists of a linear transformation of all the outputs in the training set, with weights going smoothly to zero as the distance from the target point increases. This may render the proposed estimator less exposed to the curse-of-dimensionality problem. In fact, K can be much more predictive than DWO when the target point falls in those regions of the input space sampled less frequently. These are areas where the local filters used by DWO may be forced to use neighborhoods sparsely populated by training samples. This well

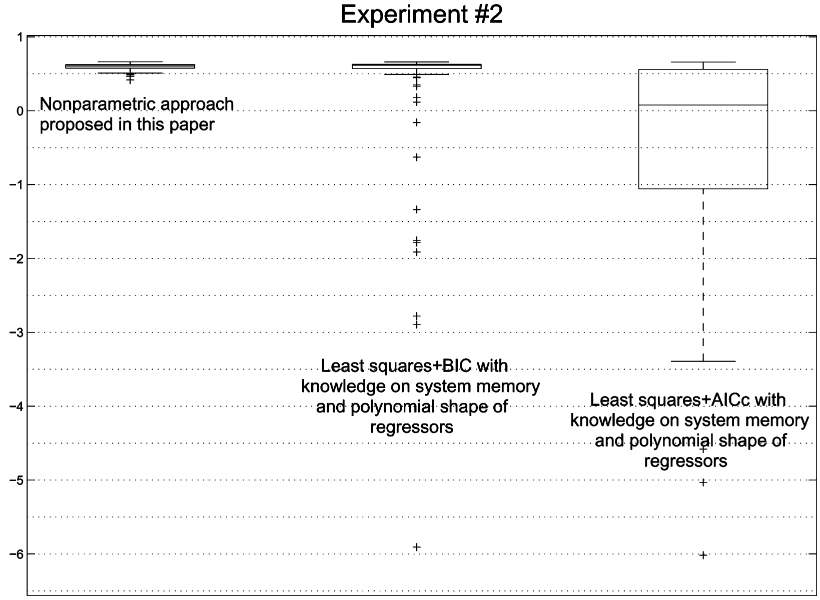

Fig. 8. Monte Carlo study #2 (Section VII-C): boxplot of $\{COD_j\}_{j=1}^{100}$ using the new nonparametric approach (left) and least squares (middle and right). In the latter case, the estimator knows that the optimal prediction at time t depends only on the input and the output at $t-1$ and that the regressors are polynomials, with their number estimated at any run by BIC (middle) or AICc (right). The larger COD_j, the better is the performance of the estimator.

explains the superiority of K over $\text{DWO} + \text{oracle}$ in Example #2, since in this case the test set is obtained using an input with statistics different from those of the input used to generate the training set.

D. Comparison With Parametric Approaches Based on AIC and BIC

To further assess the robustness of the proposed approach, we have also reconsidered Experiment #2 by introducing two additional parametric estimators based either on AIC or BIC. More specifically, we have considered the model

$$y_t = \sum_{i=0}^{3}\sum_{j=0}^{3} \chi_{ij} c_{ij} (y_{t-1})^i (u_{t-1})^j + e_t \tag{51}$$

where each χ_{ij} may assume values 0 or 1 while the $\{c_{ij}\}$ are real scalars. We have repeated the simulation study estimating, at any run, the $\{c_{ij}\}$ via least squares, with the number of χ_{ij} different from zero determined using either the corrected version of Akaike's criterion (AICc), see [5], or BIC. Fig. 8 reports the boxplots of the $\{COD_j\}$ obtained by the new approach proposed in this paper (same as in the right panel of Fig. 7) and by the two new estimators. The performance of the nonparametric approach is superior than that obtained by the parametric estimators. For example, in almost 10% of the cases the COD_j obtained by BIC is close to zero or negative due to the introduction of too many regressors in the model, a problem that is further exacerbated when AICc is employed. This result is remarkable since the nonparametric approach performs better than parametric estimators provided with the information regarding the polynomial shape of the regressors and the fact that the optimal predictor of y_t only depends on y_{t-1} and u_{t-1}. As discussed in [6], this is a consequence of the robustness related to the use of the marginal likelihood for selecting model complexity in comparison with criteria such as AICc and BIC based on an approximation of the likelihood that is only asymptotically exact.

VIII. Conclusion

Following the philosophy developed in [6], a new nonparametric identification algorithm for nonlinear modeling has been proposed. It relies upon regression via Gaussian processes and the design of a new kernel specifically suited to nonlinear system identification. The main features of the new approach can be summarized as follows:

- the user is not required to define any part of the algorithmic architecture, e.g., the regressors and the model order. Basis functions encode the idea of "fading" memory in the predictor and are automatically learnt from the observed data;
- the choice of the hyperparameters of the kernel, whose tuning plays a role similar to model order selection in parametric approaches, is performed by optimizing a marginal likelihood. In this likelihood maximization the uncertainty of the unknown nonlinear system f is accounted for. As a matter of fact, this criterion proves robust in establishing the right trade-off between bias and variance. Benchmarks problems taken from the literature illustrate the potential of the new approach. They also reveal that the generalization capability of the new estimator may be superior than that of well established techniques such as parametric approaches equipped with AICc or BIC and direct weight optimization, even when the latter is combined with an oracle which tunes optimally its parameters.

As far as the computational complexity of the new algorithm is concerned, it depends on the cost of evaluating the minus log of the marginal likelihood (28) that is, in general, an $O(N^3)$ problem. When N is large, to speed up the computation, a simple yet effective strategy consists of estimating the hyperparameters using only a subset of the measurements, subsequently using the entire data set to determine $\hat{f}$ in (28), see also [19] for other more sophisticated strategies.

It has also to be noticed that hyperparameters estimates are obtained by optimizing a non convex objective. Even if the latter is defined in a low-dimensional space and the numerical experiments reported here seem to suggest that local minima are not critical, in the near future it would be interesting to exploit also some new approaches recently developed in [33], [34] to learn the kernel. For example, along this line, it would be interesting to investigate the technique described in [35]. It would permit to tune the hyperparameters of the new kernel proposed in this paper optimizing an objective with no risk of local minima and without resorting to any Gaussian assumption on the innovation sequence which, in our scheme, permits to compute the marginal likelihood.

APPENDIX

Proof of Propositions 1 and 2: In order to streamline notation in this section we omit the dependence on the input u, also when specifying input locations, as well as the dependence on the hyperparameters θ and p.

Our aim is to compute the conditional expectation $\mathbb{E}[f(x)|y^+,y^-]$ (see also Chapter 1 in [24] for the connection between estimation of Gaussian processes and regularization in RKHS). To this purpose it is convenient to define

$$\bar{f} = [\, f(y^1) \quad f(y^2) \quad \dots \quad f(y^N)\,]^T$$

where the input locations y^N, y^{N-1}, etc., are fixed; the autocovariance of $\bar{f}$ is $\bar{R} \in \mathbb{R}^{N\times N}$ defined by

$$[\bar{R}]_{i,j} = R(y^i, y^j). \tag{52}$$

Using the tower property of conditional expectation and (21) we obtain:

$$\begin{aligned}\mathbb{E}[f(x)|y^+,y^-] &= \mathbb{E}[\mathbb{E}[f(x)|\bar{f},y^+,y^-]|y^+,y^-]\\ &= \mathbb{E}[\mathbb{E}[f(x)|\bar{f}]|y^+,y^-].\end{aligned}$$

The inner expected value $\mathbb{E}[f(x)|\bar{f}]$ is given by:

$$\begin{aligned}\mathbb{E}[f(x)|\bar{f}] &= cov(f(x),\bar{f})\bar{R}^{-1}\bar{f}\\ &= [R(x,y^1)\dots R(x,y^N)]\bar{R}^{-1}\bar{f}.\end{aligned}$$

Therefore

$$\begin{aligned}\mathbb{E}[f(x)|y^+,y^-] &= \mathbb{E}[\mathbb{E}[f(x)|\bar{f}]|y^+,y^-]\\ &= cov(f(x),\bar{f})\bar{R}^{-1}\mathbb{E}[\bar{f}|y^+,y^-].\end{aligned} \tag{53}$$

As a last step $\mathbb{E}[\bar{f}|y^+,y^-]$ have to be computed. To this purpose the conditional distribution $\mathbf{p}(\bar{f}|y^+,y^-)$ is needed, which can be obtained as follows. Since $\bar{f}$ is the random field f sampled at the input locations $\{y^N,\dots,y^-\}$, $\mathbf{p}(\bar{f}|y^+,y^-)$ is a marginal density from $\mathbf{p}(f|y^+,y^-)$, i.e., the joint density of all the random field f, conditional on y^+ and y^-, where $f(x)$, $x \notin \{y^N, y^{N-1},\dots,y^1\}$, is integrated out. The conditional density $\mathbf{p}(f|y^+,y^-)$ can be obtained as

$$\mathbf{p}(f|y^+,y^-) = \frac{\mathbf{p}(y^+,f|y^-)}{\mathbf{p}(y^+|y^-)}$$

with $\mathbf{p}(y^+,f|y^-)$ given by

$$\left[\prod_{i=1}^{N}\mathbf{p}(y_i|y^i,f)\right]\mathbf{p}(f|y^-) = \left[\prod_{i=1}^{N}\mathbf{p}(y_i|f(y^i))\right]\mathbf{p}(f)$$

where the first equality uses the chain rule while the last equality uses also the approximation (21).

In order to compute the marginal of $\mathbf{p}(f|y^+,y^-)$ w.r.t. to all $f(x)$, $x \notin \{y^N, y^{N-1},\dots,y^1\}$ it is sufficient to compute the marginal $\mathbf{p}(\bar{f})$ of $\mathbf{p}(f)$ obtaining

$$\mathbf{p}(\bar{f}|y^+,y^-) = \frac{\left[\prod_{i=1}^{N}\mathbf{p}(y_i|f(y^i))\right]\mathbf{p}(\bar{f})}{\mathbf{p}(y^+|y^-)}.$$

From the above equations and using the fact that $\mathbf{p}(\bar{f})$ is a zero mean Gaussian distribution with covariance $\bar{R}$ defined in (52) and $y^+ - \bar{f}$ is white Gaussian noise of variance η^2, independent of $\bar{f}$, we obtain

$$\begin{aligned}Q(y^+,\bar{f}) &:= -\log(\mathbf{p}(\bar{f},y^+|y^-))\\ &= \frac{\|y^+-\bar{f}\|^2}{2\eta^2} + \frac{\bar{f}^T\bar{R}^{-1}\bar{f}}{2}\\ &\quad + \frac{1}{2}\log\det(4\pi^2\eta^2\bar{R}).\end{aligned}$$

Thus, $\bar{f}$ conditional on y^+ remains Gaussian. In addition, the Hessian and the gradient, evaluated at 0, of Q with respect to its second argument are respectively given by

$$\partial^2_{\bar{f}}Q(y^+,\cdot) = \bar{R}^{-1} + \eta^{-2}I_N, \quad \nabla_{\bar{f}}Q(0) = -\eta^{-2}y^+. \tag{54}$$

Hence, the minimizer of $Q(y^+,\cdot)$, that corresponds to the minimum variance estimator of $\bar{f}$, is

$$\begin{aligned}\mathbb{E}[\bar{f}|y^+,y^-] &= -\left[\partial^2_{\bar{f}}Q(y^+,\cdot)\right]^{-1}\nabla_{\bar{f}}Q(0)\\ &= \bar{R}(\bar{R}+\eta^2 I_N)^{-1}y^+ = \bar{R}\Sigma_y^{-1}y^+\end{aligned} \tag{55}$$

where Σ_y was defined in (24). Plugging (56) into (53) proves (22).

As for the second part of the proposition, first recall that given the function r on a finite-dimensional domain, whose Hessian matrix $\partial^2_x r(x)$ is constant and positive definite, Laplace's method provides the following expression for calculating exponential integrals (see [36])

$$\int_{\mathbb{R}^N} e^{-r(x)}dx = \det\left[\frac{\partial^2_x r}{2\pi}\right]^{-1/2} e^{-r(\hat{x})} \tag{56}$$

where $\hat{x}$ minimizes $r(x)$ with respect to x. Then, from (56) we obtain that $-\log[\mathbf{p}(y^+|y^-)]$ is exactly equal to

$$\frac{1}{2}\log\det\left(\frac{\partial^2_{\bar{f}}Q(y^+,\cdot)}{2\pi}\right) + Q(y^+,\mathbb{E}[\bar{f}|y^+]). \tag{57}$$

Using Lemma 19 in [37] one has

$$\begin{aligned}\frac{1}{2}\log\det\left(\frac{\partial^2_{\bar{f}}Q(y^+,\cdot)}{2\pi}\right) &= \frac{1}{2}\log\det(2\pi\Sigma_y)\\ &\quad -\frac{1}{2}\log\det(4\pi^2\eta^2\bar{R}).\end{aligned} \tag{58}$$

Furthermore, after simple computations, we obtain

$$Q(y^+,\bar{R}\Sigma_y^{-1}y^+) = \frac{1}{2}\log\det(4\pi^2\eta^2\bar{R}) + \frac{1}{2}(y^+)^T A y^+$$

with

$$\begin{aligned}A &= \eta^{-2}\left(I_N - \bar{R}\Sigma_y^{-1}\right)\\ &\quad + \left(\Sigma_y^{-1} - \eta^{-2}I_N + \eta^{-2}\Sigma_y^{-1}\bar{R}\right)\bar{R}\Sigma_y^{-1}\\ &= \Sigma_y^{-1} + \left(\bar{R}+\eta^2 I_N\right)^{-1}\bar{R}\Sigma_y^{-1}\\ &\quad - \eta^{-2}\left(I_N - (\eta^2\bar{R}^{-1}+I_N)^{-1}\right)\bar{R}\Sigma_y^{-1} = \Sigma_y^{-1}\end{aligned}$$

where the above equation exploits the following facts

$$\eta^{-2}I_N = \left(\eta^{-2}\bar{R}+I_N\right)\Sigma_y^{-1}$$
$$\left(\eta^{-2}\bar{R}+I_N\right)^{-1} = I_N - \left(\eta^{2}\bar{R}^{-1}+I_N\right)^{-1}$$

with the last equality relying upon the matrix inversion lemma, see, e.g., [38]. This completes the proof of Proposition 1. Now, let

$$\bar{q} = \left[\, q(y^1) \quad q(y^2) \quad \ldots \quad q(y^N)\,\right]^T$$

where q is a zero-mean Gaussian random field of kernel K. Then, in view of (40), one has

$$y^+ = \bar{q} + Y^T g + U^T h + e, \quad e = [e_1, \ldots, e_N]^T. \tag{59}$$

It holds that

$$\begin{aligned}\mathbf{p}(y^+, \bar{q}, g, h|y^-) &= \left[\prod_{i=1}^{N} \mathbf{p}(y_i|y^i, \bar{q}, g, h)\right] \mathbf{p}(\bar{q}, g, h|y^-)\\ &= \left[\prod_{i=1}^{N} \mathbf{p}(y_i|y^i, \bar{q}, g, h)\right] \mathbf{p}(\bar{q}, g, h).\end{aligned}$$

Similarly to the previous case, from the equation above it is easy to see that the log of the a posteriori density of $\bar{q}, g, h$ conditional on y^+ and y^- is quadratic with respect to $\bar{q}, g, h$, i.e., a posteriori $\bar{q}$, g, h remain jointly Gaussian. In view of the structure of the posterior, computing $\mathbb{E}[g|y^+, y^-]$ and $\mathbb{E}[h|y^+, y^-]$ is equivalent to computing the minimum variance estimates of g and h using the model (59) with U and Y thought of as known operators not depending on the measurements. Hence, the expressions for $\hat{g}$ and $\hat{h}$ in (42) are obtained exploiting standard results on estimation of jointly Gaussian processes, see [38].

Proof of Theorem 3: The proof is based on the characterization of $\mathcal{H}_K$ given in Theorem 2; exploiting it we show that monomials do not have finite norm in $\mathcal{H}_K$ and hence they do not belong to $\mathcal{H}_K$. To do this, first the following preliminary lemma is needed.

Preliminary Lemma: For all $k \in \mathbb{N}$

$$\frac{(2k-1)!!}{(2k)!!} > \frac{3}{5}\frac{1}{\sqrt{2k}}.$$

where the definition for double factorial reads as follows:

$$k!! = \begin{cases} k\cdot(k-2)\ldots 5\cdot 3\cdot 1 & k>0 \text{ odd}\\ k\cdot(k-2)\ldots 6\cdot 4\cdot 2 & k>0 \text{ even}\\ 1 & k=0\end{cases}$$

Proof: Stirling's approximation for the Gamma function is (see [39])

$$\begin{aligned}&\Gamma(a+1)\\ &\quad= \sqrt{2\pi a}\left(\frac{a}{e}\right)^a \left[1 + \frac{1}{12a} + \frac{1}{288a^2} - \frac{139}{51840a^3} + \cdots\right]\end{aligned}$$

for $a > 0$. In particular, using formula 5.6.1 in Chapter 5 of [40], one obtains

$$\sqrt{2\pi a}\left(\frac{a}{e}\right)^a < \Gamma(a+1) < \sqrt{2\pi a}\left(\frac{a}{e}\right)^a e^{1/12a}.$$

We have first

$$(2k)!! = 2^k k! < \sqrt{2\pi k}\left(\frac{2k}{e}\right)^k e^{1/12k}.$$

From the formula

$$\Gamma\left(k+\frac{1}{2}\right) = \frac{(2k-1)!!\sqrt{\pi}}{2^k},$$

we obtain

$$\begin{aligned}(2k-1)!! &= \frac{1}{\sqrt{\pi}}2^k\Gamma\left(k+\frac{1}{2}\right) = \frac{1}{\sqrt{\pi}}2^k\Gamma\left(k-\frac{1}{2}+1\right)\\ &> \frac{1}{\sqrt{\pi}}2^k\sqrt{2\pi\left(k-\frac{1}{2}\right)}\left(\frac{k-\frac{1}{2}}{e}\right)^{k-1/2}.\end{aligned}$$

Thus

$$\begin{aligned}\frac{(2k-1)!!}{(2k)!!} &> \frac{\frac{1}{\sqrt{\pi}}2^k\sqrt{2\pi\left(k-\frac{1}{2}\right)}\left(\frac{k-\frac{1}{2}}{e}\right)^{k-1/2}}{\sqrt{2\pi k}\left(\frac{2k}{e}\right)^k e^{1/12k}}\\ &= \sqrt{\frac{e}{\pi}}\frac{1}{\sqrt{k}e^{1/12k}}\left(\frac{k-\frac{1}{2}}{k}\right)^k\\ &= \sqrt{\frac{e}{\pi}}\frac{1}{\sqrt{k}e^{1/12k}}\left(1-\frac{1}{2k}\right)^k.\end{aligned}$$

Consider the function $r(x) = (1-1/x)^x$ for $x \geq 2$. Let $v(x) = \ln r(x) = x\ln(1-1/x)$, then

$$\begin{aligned}v'(x) &= \ln\left(1-\frac{1}{x}\right) + \frac{1}{x-1}\\ &= t - \ln(1+t) \geq 0, \quad t = \frac{1}{x-1}\end{aligned}$$

where we have used the inequality $t \geq \ln(1+t)$ for $t \geq 0$. Thus $v(x)$ is a strictly increasing function on $[2, \infty)$, with minimum $v(2) = -2\ln 2$. Hence r is also a strictly increasing function on $[2, \infty)$, with minimum $r(2) = e^{-2\ln 2} = 1/4$. Thus we have for all $k \in \mathbb{N}$:

$$\left(1-\frac{1}{2k}\right)^{2k} \geq \frac{1}{4}.$$

Finally,

$$\begin{aligned}\frac{(2k-1)!!}{(2k)!!} &> \sqrt{\frac{e}{\pi}}\frac{1}{\sqrt{k}e^{1/12k}}\frac{1}{2} \geq \sqrt{\frac{e}{\pi}}\frac{1}{\sqrt{k}e^{1/12}}\frac{1}{2}\\ &= \sqrt{\frac{e}{2\pi}}\frac{1}{e^{1/12}}\frac{1}{\sqrt{2k}} > \frac{3}{5}\frac{1}{\sqrt{2k}}.\end{aligned}$$

This concludes the proof of the preliminary lemma.

Now, for simplicity, let us consider the case $n = 1$ and the mononomial x^d, $d = 0, 1, \ldots$. Then

$$\begin{aligned}\mathcal{H}_K = \Bigg\{ r = e^{-x^2/\sigma^2}\sum_{k=0}^{\infty} w_k x^k :\\ \|r\|^2_{\mathcal{H}_K} = \sum_{k=0}^{\infty}\frac{\sigma^{2k}k!}{2^k}w_k^2 < \infty\Bigg\}.\end{aligned}$$

The function x^d is then

$$x^d = e^{-x^2/\sigma^2} x^d e^{x^2/\sigma^2} = e^{-x^2/\sigma^2} \sum_{k=0}^{\infty} \frac{x^{2k+d}}{\sigma^{2k} k!}.$$

Thus $w_{2k+d} = 1/\sigma^{2k} k!$ and $w_j = 0$ for $j \neq 2k + d$. Then

$$\begin{aligned}\sum_{k=0}^{\infty} \frac{\sigma^{2k} k!}{2^k} w_k^2 &= \sum_{k=0}^{\infty} \frac{\sigma^{4k+2d}(2k+d)!}{2^{2k+d}} \frac{1}{\sigma^{4k}(k!)^2} \\ &= \frac{\sigma^{2d}}{2^d} \sum_{k=0}^{\infty} \frac{(2k+d)!}{2^{2k}(k!)^2}.\end{aligned}$$

One has $(2k)! = (2k-1)!!(2k)!!$ and $2^k k! = (2k)!!$, thus

$$\begin{aligned}\sum_{k=0}^{\infty} \frac{(2k+d)!}{2^{2k}(k!)^2} \geq \sum_{k=0}^{\infty} \frac{(2k)!}{2^{2k}(k!)^2} &= 1 + \sum_{k=1}^{\infty} \frac{(2k-1)!!}{(2k)!!} \\ &> 1 + \frac{3}{5} \sum_{k=1}^{\infty} \frac{1}{\sqrt{2k}} = \infty, \qquad (60)\end{aligned}$$

where we have used the inequality

$$\frac{(2k-1)!!}{(2k)!!} > \frac{3}{5} \frac{1}{\sqrt{2k}}$$

which follows from Stirling's formula. Hence we have $\sum_{k=0}^{\infty} \sigma^{2k} k!/2^k w_k^2 = \infty$, which shows that x^d does not belong to $\mathcal{H}_K$. The proof then easily extends to the case $n > 1$.

Proof of Theorem 4: We claim that $\mathcal{H}_{K_j} \cap \mathcal{H}_{K_r} = \{0\}$ for $j \neq r$, $1 \leq j, r \leq n - p + 1$. For simplicity, consider the example $K(x, z; 3, 2)$. Then by Theorem 2,

$$\mathcal{H}_{K_1} = \left\{ r = e^{-x_1^2 + x_2^2/\sigma^2} \sum_{\alpha_1+\alpha_2=0}^{\infty} w_{\alpha_1,\alpha_2} x_1^{\alpha_1} x_2^{\alpha_2} : \right. \\ \left. \|r\|_{H_{K_1}}^2 = \sum_{k=0}^{\infty} \frac{k!}{\frac{(2}{\sigma^2)^k}} \sum_{\alpha_1+\alpha_2=k} \frac{w_{\alpha_1,\alpha_2}^2}{C_{\alpha_1,\alpha_2}} < \infty \right\}.$$

with orthonormal basis

$$\left\{ \psi_{\alpha_{(1)}}^1(x) = \sqrt{\frac{2^{|\alpha_{(1)}|} C_{\alpha_1,\alpha_2}}{\sigma^{2|\alpha_{(1)}|} |\alpha_{(1)}|!}} e^{-x_1^2 + x_2^2/\sigma^2} x_1^{\alpha_1} x_2^{\alpha_2} \right\}_{|\alpha_{(1)}|=0}^{\infty}. \qquad (61)$$

From in [21, Section VIII] it comes that $\mathcal{H}_{K_1} = \mathcal{H}_{K_{11}} \otimes \mathcal{H}_{K_{12}}$, where $K_{11}(x,z) = e^{-(x_1-z_1)^2/\sigma^2}$, $K_{12}(x,z) = e^{-(x_2-z_2)^2/\sigma^2}$ and $\otimes$ denotes the direct product between Hilbert spaces. Here we make use of the crucial property that both spaces $\mathcal{H}_{K_{1i}}$, $i = 1, 2$, do not contain the nonzero constant function. Suppose that a function $r(x_1, x_2) \in \mathcal{H}_{K_1}$ can be written in the form $r(x_1, x_2) = v(x_1)$. Then $r(c, x_2) = v(c) \in \mathcal{H}_{K_{12}}$ for any constant c. By the property we just stated, this is only possible if v is the zero function. This means that functions in $\mathcal{H}_{K_1}$ must depend nontrivially on the two arguments x_1 and x_2. Similarly, functions in $\mathcal{H}_{K_2}$ must depend nontrivially on the two arguments x_2 and x_3. Therefore we must have $\mathcal{H}_{K_1} \cap \mathcal{H}_{K_2} = \{0\}$. Then from a result on sums of kernels (see Aronszajn [21], pages 353–354), the space $\mathcal{H}_K$ is the direct sum of the spaces $\mathcal{H}_{K_j}$, which are complementary subspaces in $\mathcal{H}_K$. Thus each $r \in \mathcal{H}_K$ admits a unique decomposition $r = r^1 + r^2$, where $r^i \in \mathcal{H}_{K_i}$ and

$$\|r\|_{\mathcal{H}_K}^2 = \|r^1\|_{\mathcal{H}_{K_1}}^2 + \|r^2\|_{\mathcal{H}_{K_2}}^2.$$

The results stated then follow immediately. The reasoning in the general case is similar.

Proof of Theorem 5: For what regards the first part of the theorem, for simplicity we will provide the proof for the case $n = 2$ and $p = 1$, with $X = [0,1] \times [0,1]$. The reasoning in the general case is entirely similar. We know that

$$\mathcal{H}_K = \mathcal{H}_{K_1} \oplus \mathcal{H}_{K_2} \quad \text{where} \quad K_i(x,z) = e^{-(x_i - z_i)^2/\sigma_i^2}.$$

Thus, each function $r \in \mathcal{H}_K$ is of the form $r(x_1, x_2) = r^1(x_1) + r^2(x_2)$ where $r^i \in \mathcal{H}_{K_i}$. Then, by the universality of the single Gaussian kernel, $\mathcal{H}_K$ can approximate, in the uniform topology any continuous function $v \in C_0(X)$ of the form $v(x_1, x_2) = v^1(x_1) + v^2(x_2)$, where $v^i \in C_0([0,1])$. In other words, the p_0 introduced in the statement of the theorem needs just to be set to 1 to approximate arbitrary well every function of two arguments that decomposes into $v^1(x_1) + v^2(x_2)$. However, $\mathcal{H}_K$ generally can not approximate continuous functions of the form $v^1(x_1)v^2(x_2)$. In fact, consider the function $x_1 x_2$. Suppose that for $\epsilon > 0$ there exists a function $r(x_1, x_2) = r^1(x_1) + r^2(x_2) \in \mathcal{H}_K$ such that:

$$\|x_1 x_2 - (r^1(x_1) + r^2(x_2))\|_{C_0} < \epsilon, \\ (x_1, x_2) \in [0,1] \times [0,1].$$

One has

$$\begin{aligned} &|r^1(0) + r^2(0)| < \epsilon, \qquad &|r^1(0) + r^2(1)| < \epsilon \\ &|r^1(1) + r^2(0)| < \epsilon, \qquad &|1 - (r^1(1) + r^2(1))| < \epsilon. \end{aligned}$$

The last expression implies that $|r^1(1) + r^2(1)| > 1 - \epsilon$. On the other hand, from the first three expressions, we obtain

$$\begin{aligned} |r^1(1) + r^2(1)| &= \\ &= |r^1(1) + r^2(0) + r^1(0) + r^2(1) - (r^1(0) + r^2(0))| \\ &\leq |r^1(1) + r^2(0)| + |r^1(0) + r^2(1)| + |r^1(0) + r^2(0)| \\ &< \epsilon + \epsilon + \epsilon = 3\epsilon. \end{aligned}$$

We obtain a contradiction if $3\epsilon < 1 - \epsilon$, that is if $\epsilon < 1/4$. This completes the first part of our proof. As far as the second part is concerned, it suffices taking, e.g., $p_0 = n$ and exploiting the universality of the classical Gaussian kernel.

Acknowledgment

The authors would like to thank Dr. E. Bai for providing them with the code for the direct weight optimization method.

References

[1] T. Söderström and P. Stoica, *System Identification*. Englewood Cliffs, NJ: Prentice-Hall, 1989.

[2] L. Ljung, *System Identification, Theory for the User*. Englewood Cliffs, NJ: Prentice Hall, 1997.

[3] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Automat. Contr.*, vol. AC-19, pp. 716–723, 1974.

[4] G. Schwarz, “Estimating the dimension of a model,” *Ann. Statist.*, vol. 6, pp. 461–464, 1978.

[5] C. Hurvich and C. Tsai, “Regression and time series model selection in small samples,” *Biometrika*, vol. 76, pp. 297–307, 1989.

[6] G. Pillonetto and G. De Nicolao, “A new kernel-based approach for linear system identification,” *Automatica*, vol. 46, no. 1, pp. 81–93, 2010.

[7] G. Pillonetto, A. Chiuso, and G. De Nicolao, “Prediction error identification of linear systems: A nonparametric Gaussian regression approach,” *Automatica*, vol. 47, no. 2, pp. 291–305, 2011.

[8] J. Sjoberg, Q. Zhang, L. Ljung, A. Benveniste, B. Delyon, P. Glorennec, H. Hjalmarsson, and A. Juditsky, “Nonlinear black-box modeling in system identification: A unified overview,” *Automatica*, vol. 31, pp. 1691–1724, 1995.

[9] I. Lind and L. Ljung, “Regressor and structure selection in NARX models using a structured ANOVA approach,” *Automatica*, vol. 44, pp. 383–395, 2008.

[10] I. Leontaritis and S. Billings, “Input-output parametric models for non-linear systems part II: Stochastic non-linear systems,” *Int. J. Contr.*, vol. 41, no. 2, pp. 329–344, 1985.

[11] T. Lin, B. Horne, P. Tino, and C. Giles, “Learning long-term dependencies in NARX recurrent neural networks,” *IEEE Trans. Neural Netw.*, vol. 7, no. 6, pp. 1329–1338, Nov. 1996.

[12] S. Shun-Feng and F. Yang, “On the dynamical modeling with neural fuzzy networks,” *IEEE Trans. Neural Netw.*, vol. 13, pp. 1548–1553, 2002.

[13] N. Draper and H. Smith, *App. Regression Anal.* New York: Wiley, 1981.

[14] R. Haber and H. Unbehauen, “Structure identification of nonlinear systems-a survey,” *Automatica*, vol. 26, pp. 651–677, 1990.

[15] S. Billings, A. Chen, and M. Korenberg, “Identification of MIMO nonlinear systems using a forward-regression orthogonal algorithm,” *Int. J. Contr.*, vol. 49, pp. 2157–2189, 1989.

[16] W. Spinelli, L. Piroddi, and M. Lovera, “On the role of prefiltering in nonlinear system identification,” *IEEE Trans. Automat. Contr.*, vol. 50, pp. 1597–1602, 2005.

[17] J. Roll, A. Nazin, and L. Ljung, “Nonlinear system identification via direct weight optimization,” *Automatica*, vol. 41, pp. 475–490, 2005.

[18] E. Bai and Y. Liu, “Recursive direct weight optimization in nonlinear system identification: A minimal probability approach,” *IEEE Trans. Automat. Contr.*, vol. 52, no. 7, pp. 1218–1231, 2007.

[19] C. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning*. Cambridge, MA: The MIT Press, 2006.

[20] G. Pillonetto and B. Bell, “Bayes and empirical Bayes semi-blind deconvolution using eigenfunctions of a prior covariance,” *Automatica*, vol. 43, no. 10, pp. 1698–1712, 2007.

[21] N. Aronszajn, “Theory of reproducing kernels,” *Trans. Amer. Math. Soc.*, vol. 68, pp. 337–404, 1950.

[22] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Berlin, Germany: Springer-Verlag, 2008.

[23] E. Hannan and M. Deistler, *The Statistical Theory of Linear Systems*. New York: Wiley, 1988.

[24] G. Wahba, *Spline Models for Observational Data*. Philadelphia, PA: SIAM, 1990.

[25] H. Minh, “Reproducing kernel Hilbert spaces in learning theory,” Ph.D., Brown Univ., Providence, RI, 2006.

[26] S. Saitoh, *Theory of Reproducing Kernels and its Applications*. New York: Longman, 1988.

[27] E. Bai, R. Tempo, and Y. Liu, “Identification of IIR nonlinear systems without prior structural information,” *IEEE Trans. Automat. Contr.*, vol. 52, no. 3, pp. 442–453, Mar. 2007.

[28] D. MacKay, “Probable networks and plausible predictions—A review of practical Bayesian methods for supervised neural networks,” *Network: Computation in Neural Systems*, vol. 6, pp. 469–505, 1995.

[29] W. Gilks, S. Richardson, and D. Spiegelhalter, *Markov Chain Monte Carlo in Practice*. London, U.K.: Chapman and Hall, 1996.

[30] I. Steinwart, D. Hush, and C. Scovel, “An explicit description of the reproducing Kernel Hilbert space of Gaussian RBF kernels,” *IEEE Trans. Inf. Theory*, vol. 52, no. 10, pp. 4635–4643, Oct. 2006.

[31] S. Smale and D. Zhou, “Learning theory estimates via integral operators and their approximations,” *Constructive Approximation*, vol. 26, pp. 153–172, 2007.

[32] L. Piroddi and W. Spinelli, “An identification algorithm for polynomial NARX models based on simulation error minimization,” *Int. J. Contr.*, vol. 76, pp. 1767–1781, 2003.

[33] A. Argyriou, C. A. Micchelli, and M. Pontil, “Learning convex combinations of continuously parameterized basic kernels,” in *Proc. Conf. Learning Theory (COLT’05)*, 2005, pp. 338–352.

[34] C. A. Micchelli and M. Pontil, “Learning the kernel function via regularization,” *J. Machine Learning Res.*, vol. 6, pp. 1099–1125, 2005.

[35] F. Dinuzzo, Kernel Machines With Two Layers and Multiple Kernel Learning 2010 [Online]. Available: http://www-dimat.unipv.it/dinuzzo

[36] N. D. Bruijn, *Asymptotic Methods in Analysis*. Amsterdam, The Netherlands: North-Holland, 1961.

[37] B. Bell and G. Pillonetto, “Estimating parameters and stochastic functions of one variable using nonlinear measurements models,” *Inverse Problems*, vol. 20, no. 3, pp. 627–646, 2004.

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

[39] M. Abramowitz and I. Stegun, *Handbook of Mathematical Functions, With Formulas, Graphs, and Mathematical Tables*. New York: National Bureau of Standards, 1964.

[40] Digital Library of Mathematical Function NIST [Online]. Available: http://dlmf.nist.gov/

Gianluigi Pillonetto (M’03) was born on January 21, 1975, in Montebelluna (TV), Italy. He received the Doctoral degree in computer science engineering *cum laude* from the University of Padova, Padova, Italy, in 1998 and the Ph.D. degree in bioengineering from the Polytechnic of Milan, Milan, Italy, in 2002.

In 2000 and 2002, he was Visiting Scholar and Visiting Scientist, respectively, at the Applied Physics Laboratory, University of Washington, Seattle. From 2002 to 2005, he was Research Associate at the Department of Information Engineering, University of Padova. Since 2005, he is an Assistant Professor of Control and Dynamic Systems at the Department of Information Engineering, University of Padova. His research interests are in the field of system identification, stochastic systems, deconvolution problems, nonparametric regularization techniques, learning theory, and randomized algorithms.

Minh Ha Quang received the B.Sc. degree in mathematics and computer science from Monash University, Victoria, Australia, and the Ph.D. degree in mathematics from Brown University, Providence, RI , in 2006.

He is a Senior Postdoctoral Researcher in Computer Imaging at the Italian Institute of Technology, Genoa, Italy. His research interests include applied and computational functional and harmonic analysis, machine learning, computational statistics, and applications in data analysis, image and signal processing.

Alessandro Chiuso (SM’06) received the “Laurea” degree summa cum laude in Telecommunication Engineering from the University of Padova, Padova, in July 1996 and the Ph.D. degree in System Engineering from the University of Bologna, Bologna, Italy, in 2000.

He is an Associate Professor with the Department of Management and Engineering, University of Padova. He has held visiting positions with Washington University, St. Louis, MO, KTH (Sweden), and the University of California,Los Angeles. His research interest are mainly in estimation, identification theory and applications. "information can be found at the personal web page http://automatica.dei.unipd.it/people/chiuso.html.

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