Proceedings of the
47th IEEE Conference on Decision and Control
Cancun, Mexico, Dec. 9-11, 2008

Predictor estimation via Gaussian regression

Gianluigi Pillonetto, Alessandro Chiuso and Giuseppe De Nicolao

Abstract— **A novel nonparametric paradigm to model identification has been recently proposed where, in place of postulating finite-dimensional models of the system transfer function, the system impulse response is searched for within an infinite-dimensional space. In this paper, we extend such nonparametric approach to the design of optimal predictors by interpreting the predictor coefficients as realizations of Gaussian processes. Numerical experiments, where data are generated by ARMAX models, are used to show advantages of the new approach in terms of both predictive capability on new data and accuracy in reconstruction of predictor coefficients. In a companion paper, it is also shown how this new approach to predictor design may greatly enhance performance of subspace identification methods.**

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

I. Introduction

The most widespread approach to optimal prediction of discrete-time systems relies on Prediction Error Methods (PEM) for which a large corpus of theoretical results is available [1], [2]. Within this paradigm, a key point is the selection of the most adequate model structure, which is usually carried out by resorting to complexity measures such as FPE and AIC criteria [1], [2].

Recently, an alternative paradigm to model identification has been proposed which relies on nonparametric estimation of impulse responses [3]. Rather than postulating finite-dimensional structures for the system transfer function, e.g. ARX, ARMAX or Laguerre [4], [5], the system impulse response is searched for within an infinite-dimensional space. In order to circumvent the intrinsical ill-posed nature of the problem, regularization methods, admitting a Bayesian interpretation, are employed [6], [7], [8]. Within this nonparametric paradigm, a breakthrough has been the design of a prior distribution on the impulse response such that the realizations are almost surely BIBO stable [3]. This method has been shown to compare very favorably with respect to established parametric approaches.

Along this line, it is of interest to extend the nonparametric paradigm to the design of optimal predictors. By the way, predictor estimation, beyond being of interest on its own, is the preliminary step of subspace identification methods [9], [10], [11], [12]. Therefore, improving predictor design may enhance performance of subspace identification methods as well. As a matter of fact, this topic is investigated in a companion paper [13]. In this paper, without loss of generality, analysis will be restricted to SISO systems.

In the nonparametric approach to predictor estimation, the main point is to see the predictor as a system with two inputs (past outputs and inputs) and one output (output predictions). Therefore, predictor design amounts to estimating two impulse responses. In the proposed method, the impulse responses are assumed to be the realizations of a Gaussian process [14], [15]. In particular, they are the convolution of an infinite-dimensional nonparametric component and a low-order finite-dimensional one. The latter is used to capture high-frequency oscillations, e.g. poles with negative real part. The overall scheme for predictor estimation relies on an empirical Bayesian paradigm. First, the vector of unknown hyperparameters characterizing the priors is estimated via marginal likelihood maximization. In the second and final step the hyperparameters are set to their estimates and minimum variance estimates of the impulse responses are computed. Numerical experiments, with data generated by ARMAX models of different order, show that the proposed approach provides substantial improvement over existing methods in terms of both predictive capability on new data and accuracy in the reconstruction of predictor coefficients. Further elements in favor of this new technique will be found in the companion paper [13], where benefits for subspace identification of state-space models will be demonstrated.

The paper is organized as follows. In Section II, the statement of the problem is provided. In Section III, a new Gaussian prior for predictor estimation is proposed by introducing suitable autocovariances (kernels). In Section IV, a numerical algorithm which determines both the unknown hyperparameters present in the prior and the predictor structure is worked out. Further, expressions of estimates of predictor impulse responses are obtained by resorting to the theory of Reproducing Kernel Hilbert Spaces (RKHS). In Section V, simulated data are used to demonstrate the effectiveness of the proposed approach. Conclusions end the paper.

II. Statement of the problem

In the sequel, $\mathscr{B}$ denotes the Banach space of impulse responses $\{f_k\}_{k=0}^{+\infty}$ of BIBO stable discrete-time causal systems. In addition, $\mathbb{N}$ is the set of natural numbers.

We are given a finite set of noisy output data $\{y_k\}$ from a discrete-time linear dynamic system fed with a known input

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

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

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

$\{u_k\}$. The measurements model is

$$y_t = \sum_{k=1}^{\infty} q_k u_{t-k} + \sum_{k=0}^{\infty} w_k v_{t-k} \qquad (1)$$
$$\{q_k\}, \{w_k\} \in \mathscr{B}$$

where $\{v_k\}$ is white Gaussian noise. Our problem is to estimate the one-step-ahead predictor for (1) starting from $\{u_k\}$ and the output data available from $\{y_k\}$.

III. Defining the prior for predictor coefficients

In the sequel, let $\hat{y}(t)$ denote the one-step-ahead prediction for (1) at instant t. A classical approach to estimate predictor coefficients considers $\hat{y}(t)$ parametrized by a finite-dimensional vector $\theta \in \Re^p$:

$$\hat{y}(t;\theta) = \sum_{k=1}^{\infty} a_k(\theta) y_{t-k} + \sum_{k=1}^{\infty} b_k(\theta) u_{t-k} \qquad (2)$$

where $a : \Re^p \mapsto \mathscr{B}$ and $b : \Re^p \mapsto \mathscr{B}$ while $a_k(\cdot)$ and $b_k(\cdot)$ denote the impulse responses defining the predictor evaluated at instant k. In contrast with (2), we let the predictor impulse responses belong to infinite-dimensional function spaces. Moreover, a Bayesian paradigm is adopted so that statistical priors on the predictor coefficients are introduced. To be more specific, the predictor now takes the form

$$\hat{y}(t;\zeta) = \sum_{k=1}^{\infty} F_k(\zeta) y_{t-k} + \sum_{k=1}^{\infty} G_k(\zeta) u_{t-k} \qquad (3)$$

where

$$F_t(\zeta) = \sum_{k=1}^{\infty} a_k(\zeta) f_{t-k} \qquad G_t(\zeta) = \sum_{k=1}^{\infty} b_k(\zeta) g_{t-k} \qquad (4)$$

and ζ is a vector of unknown hyperparameters to be better specified in the following. In (4), $f = \{f_k\}_{k=1}^{\infty}$ and $g = \{g_k\}_{k=1}^{\infty}$ indicate zero-mean Gaussian processes, mutually independent and independent of $\{v_k\}$. Their autocovariances (kernels) are denoted by $cov(f_i, f_j) = \lambda_1^2 K_1(i,j)$ and $cov(g_i, g_j) = \lambda_2^2 K_2(i,j)$, where K_1 and K_2 map $\mathbb{N}\times\mathbb{N}$ into $\Re$, while $\{\lambda_i\}$ are unknown hyperparameters contained in ζ. Further, $a(\zeta)$ and $b(\zeta)$ represent finite-dimensional components of the model. Their choice, together with the choice of the kernels, is discussed in the next two subsections.

A. Choice of the kernels

As far as the choice of K_1 and K_2 is concerned, we will formulate a prior on $\mathbb{N}$ incorporating the BIBO-stability constraint and information on the regularity of f and g. To this aim, it is useful to recall that the most popular approach to model a continuous-time signal h which is just known to be smooth consists of describing it as an integrated Wiener process. Assuming zero initial condition at time zero, the autocovariance of the integrated Wiener process is (see e.g. [16])

$$W(s,\tau) = cov(h(s), h(\tau)) = \begin{cases} \frac{s^2}{2}\left(\tau - \frac{s}{3}\right) & s \le \tau \\ \frac{\tau^2}{2}\left(s - \frac{\tau}{3}\right) & s > \tau \end{cases} \qquad (5)$$

However, this autocovariance does not include information on BIBO-stablity because the variance of the process increases over time. Following [3], BIBO stability can be guaranteed by performing an exponential time-transformation

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

which maps the unit interval S of the real line into the positive real axis, and defining the new kernel

$$K(s,t;\beta) = W(e^{-\beta s}, e^{-\beta t}) \qquad (s,t) \in \Re^+ \times \Re^+ \qquad (6)$$

We model the discrete-time functions f and g in (4) by exploiting the sampled versions of the kernel (6), i.e.

$$cov(f_k, f_j) = \lambda_1^2 K(k,j;\beta_1) \qquad (7)$$
$$cov(g_k, g_j) = \lambda_2^2 K(k,j;\beta_2) \qquad (8)$$
$$cov(f_k, g_j) = 0, \qquad \forall k,j \in \mathbb{N} \qquad (9)$$

Notice that additional hyperparameters $\{\beta_i\}$ are included in the prior. They represent the asymptotic exponential decay rates of the variance of $\{f_k\}$ and $\{g_k\}$ which will be tuned from data together with the scale factors $\{\lambda_i\}$.
The following result provides information on the trajectories of the processes $\{f_k\}$ and $\{g_k\}$. For the proof, the reader is referred to [3] where a spectral characterization of the kernel (6) can also be found.

Proposition 1: Assuming that $\{f_k\}$ and $\{g_k\}$ are Gaussian processes with autocovariances specified by (7), (8) and (9), their realizations belong to $\mathscr{B}$ almost surely.

B. Choice of the finite-dimensional components

Maps a and b in (4) represent the finite-dimensional components of $F(\zeta)$ and $G(\zeta)$ which can be used to enhance flexibility of the predictor. In particular, they can be exploited to capture dynamics that are hardly represented by the smooth processes f and g, e.g. high-frequency oscillating poles. A significant example, also discussed in the numerical experiments section, is provided by ARMAX models, in which case it is convenient to set $a_k(\zeta) = b_k(\zeta)$, $\forall \zeta, k$, and let this part of the model describe poles with negative real part.

IV. Estimating hyper-parameters and predictor coefficients

In real applications, the parameters $\{\beta_i\}$, $\{\lambda_i\}$ and those entering in a and b, as well as the variance σ^2 of the innovation, have to be estimated from data together with the predictor coefficients. In addition, the complexity of a and b, e.g. the number of negative poles to be introduced in the prior, may not be known in advance. For these reasons, it is useful to introduce the vector $\zeta^{\mathscr{M}}$ which gathers all the unknown parameters of the nonparametric estimator once a certain structure $\mathscr{M}$ for a and b is postulated.

A. Estimates of the predictor coefficients given $\zeta^{\mathscr{M}}$

We start considering a situation where $\zeta^{\mathscr{M}}$ is perfectly known. To simplify the notation, dependencies on $\zeta^{\mathscr{M}}$ are often omitted. In addition, let $A \in \Re^{n\times\infty}$ and $B \in \Re^{n\times\infty}$ where

$$\begin{aligned}[A]_{ji} &= \sum_{k=1}^{\infty} a_k y_{j-k-i+1},\\ [B]_{ji} &= \sum_{k=1}^{\infty} b_k u_{j-k-i+1}, \quad j=1,2,\ldots,n \quad i\in\mathbb{N}\end{aligned}$$

In the sequel, notation of ordinary finite-dimensional algebra is used to handle infinite-dimensional objects. In particular, A and B will represent operators mapping $\mathscr{B}$ into $\Re^n$, e.g. the j-th element of Af is given by $\sum_{i=1}^{\infty}[A]_{ji}f_i$. So, in view of (3) and (4), it holds that

$$y = A(y,y^-)f + B(u)g + e \tag{10}$$

where u is the input sequence while

$$\begin{aligned} y &= [y_1 \quad y_2 \ldots y_n]^T\\ y^- &= [y_0 \quad y_{-1}\ldots]^T\\ e &= [e_1 \quad e_2 \ldots e_n]^T\end{aligned}$$

with $\{e_k\}$ being the sequence of innovations having variance σ^2. In practice y^- is not (at least completely) known. A solution is to set its unknown components to zero, introducing an error which goes to zero as n increases, see e.g. Section 3.2 in [1]. Letting y_a^- denote the available (observed) components of y^-, in the sequel perfect knowledge of $A(y,y^-)$ is assumed, i.e. $A(y,y^-) = A(y,y_a^-)$. Further, the following approximation for the joint density of y, f and g is assumed to hold

$$p(y,f,g) \approx p(y|y_a^-,f,g)p(f)p(g) \tag{11}$$

so that components of y_a^- are interpreted just as known parameters A depends on. To simplify the notation, dependence on y_a^- is omitted in the sequel as well as the dependence of A and B on y and u.

Recall from [17], that given a symmetric and positive-definite kernel defined on a metric space X, the RKHS associated with K is the Hilbert spaces of functions on X which are the completion of the manifolds given by all the finite linear combinations

$$\sum_{i=1}^{l} m_i K(\cdot,t_i) \tag{12}$$

for all choices of l, $\{m_i\}$ and $\{t_i\}$, with the inner product being defined by

$$<\sum_i m_i K(\cdot,t_i), \sum_j n_j K(\cdot,s_j)> = \sum_{i,j} m_i n_j K(t_i,s_j) \tag{13}$$

In the sequel, let $\mathscr{H}_1$ and $\mathscr{H}_2$ be the RKHS on $\mathbb{N}$ associated with K_1 and K_2 with norms denoted by $\|\cdot\|_{\mathscr{H}_1}$ and $\|\cdot\|_{\mathscr{H}_2}$, respectively.

Assumption 2: The linear operators $A : \mathscr{H}_1 \mapsto \Re^n$ and $B : \mathscr{H}_2 \mapsto \Re^n$ are continuous in the topologies of $\mathscr{H}_1$ and $\mathscr{H}_2$, respectively.

■

For a given model structure $\mathscr{M}$, let f^{MV} denote the minimum variance estimate of f, i.e. $f^{MV} = E[f|y,\zeta^{\mathscr{M}}]$. The minimum variance estimate g^{MV} is defined in the same way. The following result exploits the correspondence between Gaussian processes and RKHS [18].

Proposition 3: Consider (10), where f and g are Gaussian processes with distribution as specified in Section III. If the approximation (11) holds, we have

$$\begin{aligned}(f^{MV},g^{MV}) &= \arg\min_{h_1\in\mathscr{H}_1, h_2\in\mathscr{H}_2} \|y - Ah_1 - Bh_2\|^2\\ &+ \gamma_1\|h_1\|^2_{\mathscr{H}_1} + \gamma_2\|h_2\|^2_{\mathscr{H}_2}\end{aligned} \tag{14}$$

where $\|.\|$ is the Euclidean norm, $\gamma_1 = \sigma^2/\lambda_1^2$ and $\gamma_2 = \sigma^2/\lambda_2^2$.

■

With a slight abuse of notation, in the following equations we think of K_1 and K_2 as elements of $\Re^{\infty\times\infty}$, where the i-th column of K_1 and K_2 are the sequences $K(\cdot,i;\beta_1)$ and $K(\cdot,i;\beta_2)$, for $i\in\mathbb{N}$, respectively. The following result provides the solution of (14) and shows that f^{MV} and g^{MV} admit the structure of a regularization network [19].

Proposition 4: The solutions of (14) are given by

$$f^{MV} = \lambda_1^2 K_1 A^T c \qquad g^{MV} = \lambda_2^2 K_2 B^T c \tag{15}$$

where

$$c = \left(\lambda_1^2 A K_1 A^T + \lambda_1^2 B K_2 B^T + \sigma^2 I_n\right)^{-1} y \tag{16}$$

with I_n being the $n\times n$ identity matrix.

B. Estimating hyper-parameters and the structure of the predictor

In many practical situations we can assume $F_k = G_k = 0$ for $k > q$. It is worth stressing that q does not have to establish any kind of trade-off between bias and variance. It is just a value large enough to capture the dynamics of the predictor. Let $\breve{A}\in\Re^{n\times q}$ and $\breve{B}\in\Re^{n\times q}$ be matrices obtained from A and B by retaining only the first q columns while $\breve{K}_1\in\Re^{q\times q}$ and $\breve{K}_2\in\Re^{q\times q}$ are obtained by retaining only the first p rows and columns of K_1 and K_2, respectively. For the next developments, it is also useful to introduce the notations $\breve{f}$ and $\breve{g}$ indicating q-dimensional random vectors which are in one-to-one correspondence with f and g subject to the constraints $f_k = g_k = 0$ for $k > q$. When such constraints hold, we have

$$y = \breve{A}\breve{f} + \breve{B}\breve{g} + e \tag{17}$$

Given a predictor structure $\mathscr{M}$, the hyperparameter vector $\zeta^{\mathscr{M}}$ can be determined by maximizing the marginal likelihood of y, i.e. the total probability of y, f and g where f and g are integrated out, as described in the next proposition.

Proposition 5: If the approximation (11) holds, the maximum marginal likelihood estimate of $\zeta^{\mathscr{M}}$ is the solution of the optimization problem

$$\hat{\zeta}^{\mathscr{M}} = \arg\min_{\zeta^{\mathscr{M}}} J(y;\zeta^{\mathscr{M}}) \tag{18}$$

where J is the opposite of the log-marginal likelihood of y given by

$$J(y;\zeta^{\mathscr{M}}) = \frac{1}{2}\ln\left(\det[2\pi V(\zeta^{\mathscr{M}})]\right) + \frac{1}{2}y^T V^{-1}(\zeta^{\mathscr{M}})y \quad (19)$$

with

$$V[\zeta^{\mathscr{M}}] = \lambda_1^2 A K_1 A^T + \lambda_2^2 B K_2 B^T + \sigma^2 I_n \quad (20)$$

If also (17) holds, we have

$$V[\zeta^{\mathscr{M}}] = \lambda_1^2 \breve{A}\breve{K}_1\breve{A}^T + \lambda_2^2 \breve{B}\breve{K}_2\breve{B}^T + \sigma^2 I_n \quad (21)$$

■

Among the possible nonparametric estimators identified by the choice of $\mathscr{M}$, model selection is performed according to Akaike criterion, that is by minimizing

$$\mathrm{AIC}(\mathscr{M}) = 2J(y;\hat{\zeta}^{\mathscr{M}}) + 2d^{\mathscr{M}} \quad (22)$$

where $d^{\mathscr{M}}$ denotes the dimension of $\zeta^{\mathscr{M}}$. We are now in a position to summarize the entire numerical procedure for predictor estimation.

Algorithm 6: The input to this algorithm includes the input and output sequences $\{u_k\}$ and $\{y_k\}$ together with a set of competitive structures $\{\mathscr{M}_i\}$ defining a and b in (4). The outputs of this algorithm are the sequences $\{\hat{F}_k\}$ and $\{\hat{G}_k\}$ which define the predictor coefficients in (4).

- Choose the model $\hat{\mathscr{M}}_i$ which minimizes (22).
- Conditional on $\hat{\mathscr{M}}_i$, define $\hat{\zeta}^{\hat{\mathscr{M}}_i}$ as in (18).
- According to the empirical Bayes approach, determine the estimates f^{MV} and g^{MV} using (15), with hyperparameters set to $\hat{\zeta}^{\hat{\mathscr{M}}_i}$.
- Compute the sequences $\{\hat{F}_t\}$ and $\{\hat{G}_t\}$ as follows

$$\begin{aligned}\hat{F}_t(\hat{\zeta}^{\hat{\mathscr{M}}_i}) &= \sum_{k=1}^{\infty} a_k(\hat{\zeta}^{\hat{\mathscr{M}}_i}) f_{t-k}^{MV} \\ \hat{G}_t(\hat{\zeta}^{\hat{\mathscr{M}}_i}) &= \sum_{k=1}^{\infty} b_k(\hat{\zeta}^{\hat{\mathscr{M}}_i}) g_{t-k}^{MV}\end{aligned}$$

V. Numerical experiments

The performance of the proposed approach is evaluated by numerical experiments where output data are generated by three ARMAX models of order 2,4 and 6. In the z-transform domain, it holds that

$$Y(z) = \frac{P_1(z)}{P_3(z)}U(z) + \frac{P_2(z)}{P_3(z)}V(z) \quad (23)$$

where $\{u_k\}$ and $\{v_k\}$ are mutually independent white noises of unit variance. Polynomials $\{P_i\}$ defining the three models are specified below. Poles of the predictor transfer functions defining the optimal one-step-ahead predictor (PTF poles) are also reported (they coincide with the zeros of $P_2(z)$).

ARMAX models

1)

$$\begin{aligned}P_1(z) &= 0.5578z - 0.2420 \\ P_2(z) &= z^2 + 0.4z - 0.21 \\ P_3(z) &= z^2 - 0.7z - 0.18\end{aligned}$$

PTF poles = $\{-0.7, 0.3\}$

2)

$$\begin{aligned}P_1(z) &= 1.5723z^3 - 7.7367z^2 - 1.7896z - 0.9056 \\ P_2(z) &= z^4 + 0.8z^3 + 0.8z^2 + 0.256z - 0.1785 \\ P_3(z) &= z^4 - 1.1z^3 + 0.95z^2 - 0.523z - 0.153\end{aligned}$$

PTF poles = $\{-0.2 + 0.9j, -0.2 - 0.9j, -0.7, 0.3\}$

3)

$$\begin{aligned}P_1(z) &= 0.5578z^5 - 0.242z^4 + 0.2z^3 - 0.1z^2 \\ &+ 0.05z - 0.02 \\ P_2(z) &= z^6 - 0.9z^5 + 0.38z^4 + 0.22z^3 - 0.5416z^2 \\ &+ 0.2678z - 0.0392 \\ P_3(z) &= z^6 - 1.4z^5 + 1.01z^4 - 0.408z^3 - 0.1932z^2 \\ &+ 0.1851z - 0.0326\end{aligned}$$

PTF poles = $\{-0.8, 0.2 + 0.8j, 0.2 - 0.8j, 0.3, 0.4, 0.6\}$

Our aim is to estimate the one-step-ahead predictors for the three models starting from 150 noisy output data.

We compare different estimators by resorting to Monte Carlo simulations and using two measures of performance. The first one regards prediction capability on new data. In particular, at any Monte Carlo run the estimates $\{\hat{F}_k\}$ and $\{\hat{G}_k\}$ of predictor coefficients are first obtained. Then, we generate a test set of 500 new output and input data, denoted respectively by $\{y_k^{new}\}_{k=1}^{500}$ and $\{u_k^{new}\}_{k=1}^{500}$. The one-step-ahead prediction $\hat{y}_t^{new}$ is then computed and the generalization error at the j-th Monte Carlo run is

$$err_{j1} = \sqrt{\frac{\sum_{t=1}^{500}(\hat{y}_t^{new} - y_t^{new})^2}{500}} \quad (24)$$

The other measure of performance regards quality of predictor coefficients reconstruction and is characterized by the following two quantities

$$err_{j2} = \sqrt{\sum_{t=1}^{\infty}\left(\hat{F}_t - F_t\right)^2} \qquad err_{j3} = \sqrt{\sum_{t=1}^{\infty}\left(\hat{G}_t - G_t\right)^2} \quad (25)$$

During Monte Carlo simulations, 4 different predictors are designed. The first one relies upon the classical PEM approach where competitive ARMAX models of order ranging from 1 to 10 are postulated and the best one is selected according to the AIC criterion (as implemented in the MATLAB System Identification Toolbox [20]).[1]

The second predictor design method is the nonparametric one described in the previous section. In particular, let $P^j(z)$ denote a generic polynomial of order j whose roots belong to the open left semidisk of unit radius in the complex plane. Define also $\mathscr{B}_j^l$ as the impulse responses $\{h_k\} \in \mathscr{B}$ admitting the following representation in the z-transform domain

$$H^j(z) = \frac{z^j}{P^j(z)}$$

[1]We do not allow the order of the polynomials to be different from each other. However, it has been observed that introducing further competitive models would not improve the results presented in the sequel. This will also hold in the sequel when dealing with ARX modeling.

Then, we set $a_k(\zeta^{\mathscr{M}_j}) = b_k(\zeta^{\mathscr{M}_j})$, $\forall \zeta^{\mathscr{M}_j}, k$ and we define

$$a : \zeta^{\mathscr{M}_j} \mapsto \mathscr{B}_j^l \qquad b : \zeta^{\mathscr{M}_j} \mapsto \mathscr{B}_j^l$$

In this way, the finite-dimensional component of the prior in (4) describes the poles with negative real part of the predictor impulse responses. As far as the prior on f and g is concerned, we set $\beta_1 = \beta_2$ and hyperparameters $\{\lambda_1, \lambda_2, \beta_1, \beta_2, \sigma\}$, as well as a and b, are determined from data according to Algorithm 6. As far as the issue of initial effects is concerned, we set $q = 50$ in (17) and define y_a^- and y in (10,11) as the vector containing the first 50 and the last 100 available output samples, respectively. Finally, the marginal likelihood was evaluated by using the expressions reported in (19) and (21). The last two predictors rely upon ARX modeling so that predictor coefficients are estimated via least squares. The difference lies in the way model order (bounded by 40 during the simulations) is selected. To be specific, the third predictor design scheme uses the AIC criterion while the fourth one uses an "oracle", in which case model order is the one which minimizes the generalization error defined in (24). This is an ideal situation which provides a lower bound on ARX modeling performance.

Top panels of Fig. 1 report boxplots of $\{err_{j1}\}$ for the four predictors. It is apparent that the nonparametric estimator performs significantly better than PEM and ARX. Remarkably, the mean of the prediction error associated with the nonparametric approach is close to 1, the lower bound achievable by means of the optimal predictor. In addition, the performance of the nonparametric estimator is always very close to (or also better than) that of the oracle-based ARX predictor. In middle and bottom panels of Fig. 1 we display boxplots regarding $\{err_{j2}\}$ and $\{err_{j3}\}$. The superiority of the nonparametric approach is even more apparent. These outcomes are remarkable also in view of the fact that the PEM approach exploits a finite number of competitive models among which the true model is present. The nonparametric approach, instead, searches the estimate within a much larger and infinite-dimensional space. However, the AIC criterion, applied to the marginal likelihood (19), proves to be remarkably robust, while, using PEM, it is far more difficult to select model complexity since the joint likelihood, associated with a much richer parametric structure, has to be handled.

Finally, we consider an even more probing situation where the predictor is trained using an input whose nature is different from that used for generating the test set. In this way, prediction on new input data is made more difficult. In particular, the input $\{u_k\}$, used for system identification, is a square wave of period 20, which alternates between levels 0 and 1, while the input $\{u_k^{new}\}$, used to generate the test set, remains white noise. In Fig. 2 we display boxplots of $\{err_{j1}\}$ for the four predictors obtained in such a situation. Compared with the results reported in top panels of Fig. 1, one can see that the nonparametric approach still performs well whereas there is a significant degradation of the quality of the results obtained via PEM and ARX modeling.

VI. Conclusions

Approaches which are currently used for predictor design postulate finite-dimensional models which are identified by standard techniques such as least-squares and PEM. In this paper, we have extended a recently proposed nonparametric paradigm to identify the predictor within infinite-dimensional spaces of candidate models. In particular, predictor coefficients are modeled as the convolution between a Gaussian process which incorporates information on BIBO-stability and a low-order finite-dimensional model which is used to capture high-frequency oscillations. The predictor structure as well as unknown hyperparameters contained in the prior are estimated from data. Then, according to an empirical Bayes paradigm, minimum variance estimates of the predictor impulse responses are obtained. Numerical experiments involving ARMAX models show that the approach may greatly improve predictor design in terms of both prediction capability on new data and accuracy in reconstruction of predictor coefficients. In a companion paper, benefits related to the use of such technique for subspace identification of state-space models will be discussed.

VII. Acknowledgments

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

References

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

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

[3] G. D. Nicolao and G. Pillonetto, "A new kernel-based approach for system identification," in *Proceedings of 2008 American Control Conf., Seattle, WA, USA*, 2008.

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

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

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

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

[8] D. Barry, "Nonparametric bayesian regression," *The Annals of Statistics*, vol. 14, pp. 934–953, 1986.

[9] M. Verhaegen, "Identification of the deterministic part of MIMO state space models given in innovations form from input-output data," *Automatica*, vol. 30, pp. 61–74, 1994.

[10] D. Bauer, "Asymptotic properties of subspace estimators," *Automatica*, vol. 41, pp. 359–376, 2005.

[11] A. Chiuso, "The role of Vector AutoRegressive modeling in predictor based subspace identification," *Automatica*, vol. 43, no. 6, pp. 1034–1048, June 2007.

[12] P. Van Overschee and B. De Moor, *Subspace Identification for Linear Systems*. Kluwer Academic Publications, 1996.

[13] A. Chiuso, G. Pillonetto, and G. De Nicolao, "Subspace identification using predictor estimation via Gaussian regression," in *Proceedings of 2008 IEEE Conference on Decision and Control, Cancun, Mexico*, 2008.

[14] A. J. Smola and B. Schölkopf, "Bayesian kernel methods," in *Machine Learning, Proceedings of the Summer School, Australian National University*, S. Mendelson and A. J. Smola, Eds. Berlin, Germany: Springer-Verlag, 2003, pp. 65–117.

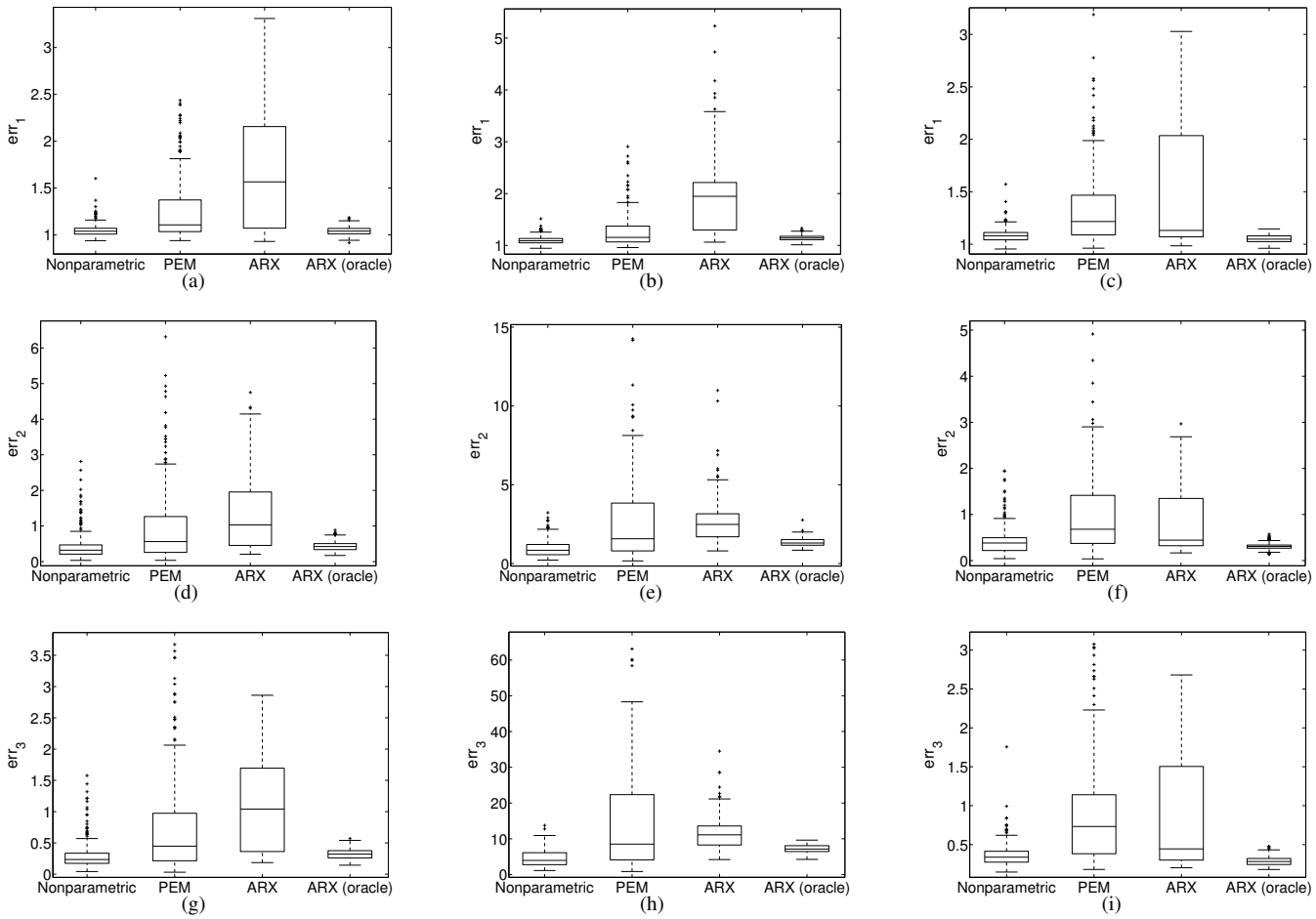

Fig. 1. Monte Carlo simulation with white noise as input for identification. Results relative to the four predictors with data generated by ARMAX models of order 2 (left), 4 (middle) and 6 (right) *Top* Prediction capability on new data: boxplot of prediction errors $\{err_{j1}\}$. *Middle and bottom* Predictor coefficients reconstruction: boxplot of errors $\{err_{j2}\}$ (middle) and $\{err_{j3}\}$ (bottom).

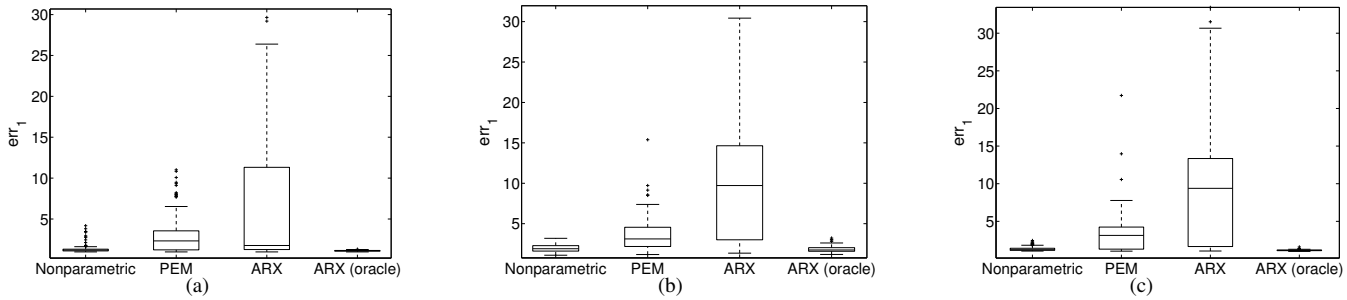

Fig. 2. Monte Carlo simulation with square wave as input for identification. Prediction capability on new data: boxplot of prediction errors $\{err_{j1}\}$ relative to the four predictors with data generated by ARMAX models of order 2 (left), 4 (middle) and 6 (right).

[15] C. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

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

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

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

[19] T. Poggio and F. Girosi, "Networks for approximation and learning," *Proceedings of the IEEE*, vol. 78, pp. 1481–1497, 1990.

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