A Bayesian approach to sparse dynamic network identification

Alessandro Chiuso [a], Gianluigi Pillonetto [b]

[a] *Dipartimento di Tecnica e Gestione dei Sistemi Industriali*
University of Padova, Vicenza, (Italy)

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

Abstract

Modeling and identification for high dimensional (i.e. signals with many components) data sets poses severe challenges to off-the-shelf techniques for system identification. This is particularly so when relatively small data sets, as compared to the number signal components, have to be used. It is often the case that each component of the measured signal can be described in terms of few other measured variables and these dependence can be encoded in a graphical way via so called "Dynamic Bayesian Networks". Finding the interconnection structure as well as the dynamic models can be posed as a system identification problem which involves variables selection. While this variable selection could be performed via standard selection techniques, computational complexity may however be a critical issue, being combinatorial in the number of inputs and outputs. Parametric estimation techniques which result in sparse models have nowadays become very popular and include, among others, the well known Lasso, LAR and their "grouped" versions Group Lasso and Group LAR. In this paper we introduce two new nonparametric techniques which borrow ideas from a recently introduced Kernel estimator called "stable-spline" as well as from sparsity inducing priors which use ℓ_1-type penalties. Numerical experiments regarding estimation of large scale sparse (ARMAX) models show that this technique provides a definite advantage over a group LAR algorithm and state-of-the-art parametric identification techniques based on prediction error minimization.

Key words: linear system identification; sparsity inducing priors; kernel-based methods; Bayesian estimation; regularization; Gaussian processes

1 Introduction

Black-box identification approaches are widely used to learn dynamic models from a finite set of input/output data [32,49]. In particular, in this paper we focus on the identification of *large scale* linear systems that involve a wide amount of variables and find important applications in many different domains such as chemical engineering, econometrics/finance, computer vision, systems biology, social networks and so on [7,39,30].

In engineering applications, when data are collected from a physical plant, it is often the case that there is an underlying interconnection structure; for instance the overall plant could be the interconnection via cascade, parallel, feedback and combinations thereof, of many dynamical systems. In this scenario any given variable may be directly related to only a few other variables.

In the static Gaussian case, the "relation" is expressed in terms of conditional independence conditions between subsets of variables, see e.g. [14]. Estimation of sparse graphical models have been the subject of intense research which is impossible to survey in this paper; we only point the reader to the early paper [37] which propose using the Lasso to this purpose, and to the more recent technical report [21] which suggests new directions introducing symmetric procedures as well as grouping strategies to reduce the number of nodes, providing also comparisons between different methods.

In the dynamic case, i.e. when observed data are trajectories of (possibly stationary) stochastic processes, one may consider several notions of conditional independence which can be encoded via the so-called time series correlation (TSC) graphs, Granger causality graphs and "partial correlation" graphs, see [13] for details.

When the number of measured variables is very large and possibly larger than the number of data available (i.e. the

[1] This paper was not presented at any IFAC meeting. Corresponding author Alessandro Chiuso Ph. +390498277709

Email addresses: alessandro.chiuso@unipd.it (Alessandro Chiuso), giapi@dei.unipd.it (Gianluigi Pillonetto).

number of "samples" available for statistical inference), even though there is no "physical" underlying network, constructing meaningful models which are useful for prediction/monitoring/intepretation requires trading off model complexity vs. fit. In a parametric setup this complexity depends on the number of parameters which is related to both the complexity of each "subsystem" (e.g. measured via its order) as well as to their number (i.e. the number of dynamical systems which are "non zero").

Problems of this sort have been recently studied in the literature, see for instance [52,40,35,36] and references therein. In the paper [52] coupled nonlinear oscillators (Kuramoto type) are considered where the coupling strengths are to be estimated; in [40] nonlinear dynamics are allowed and the attention is restricted to the linear term [1] in the state update equation, equivalent to a vector autoregressive (VAR) model of order one. In both cases it is assumed that the entire state space is measurable and an ℓ_1-penalized regression problem is solved for estimating the coupling strenghts/linear approximations. Sparse models under "smoothing" conditional independence relations, encoded by "partial correlation" graphs or equivalently via zeros in the inverse spectrum [8], have been recently studied in the literature. For instance, in [50] considers VAR models and ℓ_1-type penalized regression while in [35,36] a methodology based on smoothing *a la* Wiener is proposed, where interconnections are found by putting a threshold on the estimated transfer functions.

In this work we shall focus on stationary stochastic processes described via Granger causality graphs, where conditional independence conditions encode the fact that the prediction of (the future of) one variable (which we shall call "output variable") may require only the past history of few other variables (which we shall call "inputs") plus possibly its own past. This can be represented with a graph where nodes are variables and (directed) edges are (non zero) transfer functions, self-loops encoding dependence on the "output" own past [2] . In general both the dynamical systems and the interconnection structure is unknown and have to be inferred from data. Without loss of generality we shall address the problem of modeling the relation between one node in this graph (the "output" variable) and all the other measured variables (the "inputs") in a "prediction error" framework. Beyond linearity, we shall not make any assumption on each subsystem (e.g. no knowledge of system orders). Our focus is both on finding the underlying connection structure (if any) as well as obtaining reliable and easily interpretable models which can be used, e.g. for prediction/monitoring etc. Of course, the problem of modeling an "output" y as a function of certain inputs u is meaningful *per se*, and one may not be interested at all in building a complete "network of dependences" for the joint process (u,y) but just to perform variable selection in linear system identification when many "exogenous" variables are present.

In this scenario a key point is that the identification procedure should be sparsity-favoring, i.e. able to extract from the large number of subsystems entering the system description just that subset which influences significantly the system output. Such sparsity principle permeates many well known techniques in machine learning and signal processing such as feature selection, selective shrinkage and compressed sensing [27,16].
In the classical identification scenario, Prediction Error Methods (PEM) represent the most used approaches to optimal prediction of discrete-time systems [32]. The statistical properties of PEM (and Maximum Likelihood) methods are well understood when the model structure is assumed to be known. However, in real applications, first a set of competitive parametric models has to be postulated. Then, a key point is the selection of the most adequate model structure, usually performed by AIC and BIC criteria [1,47]. Not surprisingly, the resulting prediction performance, when tested on experimental data, may be distant from that predicted by "standard" (i.e. without model selection) statistical theory, which suggests that PEM should be asymptotically efficient for Gaussian innovations. If this drawback may affect standard identification problems, a fortiori it renders difficult the study of large scale systems where the elevated number of parameters, as compared to the number of data available, may undermine the applicability of the theory underlying e.g. AIC and BIC.
Some novel estimation techniques inducing sparse models have been recently proposed. They include the well known Lasso [51] and Least Angle Regression (LAR) [17] where variable selection is performed exploiting the ℓ_1 norm. This type of penalty term encodes the so called bi-separation feature, i.e. it favors solutions with many zero entries at the expense of few large components. Consistency properties of this method are discussed e.g. in [63,64]. Extensions of this procedure for group selection include Group Lasso and Group LAR (GLAR) [61] where the sum of the Euclidean norms of each group (in place of the absolute value of the single components) is used. Theoretical analysis of these approaches such and connections with the multiple kernel learning problem can be found in [5,38] while a discussion on the advantages of Group Lasso over Lasso is discussed in [29]. We warn the reader that one should not take "sparse" estimators as *panacea*; it is for instance shown in [31] that sparse estimators which possess some sort of "Oracle property" [19] have unbounded (normalized) maximal risk as the sample size increases.

Most of the work available in the literature addresses the "static" scenario while very little, with some exception [57,28], can be found regarding the identification of dynamic systems.
In this paper we adopt a Bayesian point of view to prediction and identification of sparse linear systems. Our starting point is the new identification paradigm developed in [45] that relies on nonparametric estimation of impulse

[1] Thinking of a first order Taylor expansion around the trajectory

[2] In the language of classical System Identification, dependence of the predictor on the past outputs will result in ARMAX models, lack of dependence in Output Error (OE) models.

responses (see also [44] for extensions to predictor estimation). Rather than postulating finite-dimensional structures for the system transfer function, e.g. ARX, ARMAX or Laguerre [32], the system impulse response is searched for within an infinite-dimensional space. The intrinsical ill-posed nature of the problem is circumvented using Bayesian regularization methods. In particular, working under the framework of Gaussian regression [46], in [45] the system impulse response is modeled as a Gaussian process whose autocovariance is the so called *stable spline kernel* that includes the BIBO stability constraint.
In this paper, expanding on our recent works [12,11], we extend this nonparametric paradigm to the design of optimal linear predictors for sparse systems. Without loss of generality, analysis is restricted to MISO systems, where the variable to be predicted is called "output variable" and all the other (say $m-1$) available variables are called "inputs". In this way we interpret the predictor as a system with m inputs (given by the past outputs and inputs) and one output (output predictions). Thus, predictor design amounts to estimating m impulse responses modeled as realizations of Gaussian processes. We set their autocovariances to stable spline kernels with unknown scale factors.

We consider two approaches: the first, which we shall call *Stable-Spline GLAR* (SSGLAR), is based in the GLAR algorithm in [61] and can be seen as a variation of the so-called "elastic net" [65]; the second, which we shall call *Stable-Spline Exponential Hyperprior* (SSEH) uses a hierarchical prior which assigns exponential hyperpriors having a common hypervariance to the scale factors. This second approach has connections with the so-called *Relevance Vector Machine* in [53]; see also the discussion on scale-mixture distributions in [24]. In this way, while SSGLAR uses the sum of the ℓ_1 norms of the single impulse responses, the hyerarchical hyperprior favors sparsity through an ℓ_1 penalty on kernel hyperparameters. Inducing sparsity by hyperpriors is an important feature of our second approach. In fact, this permits to obtain the marginal posterior of the hyperparameters in closed form and hence also their estimates in a robust way. Once the kernels are selected, the impulse responses are obtained by a convex Tikhonov-type variational problem.

As we shall see, however, SSEH requires solving a nonlinear optimization problem which may benefit from a "good" initialization. We shall argue that a forward-selection type of procedure which is in some sense related to SSGLAR provides a robust and computationally attractive way of initializing SSEH.

Numerical experiments involving sparse ARMAX systems show that this approach provides a definite advantage over both the standard GLAR (applied to ARX models) and PEM (equipped with AIC or BIC) in terms of predictive capability on new output data while also effectively capturing the "structural" properties of the dynamic network, i.e. being able to identify correctly, with high probability, the absence of dynamic links between certain variables.

The paper is organized as follows: Section 2 contains the problem formulation while Section 3 contains some background material including the nonparametric approach to system identification introduced in [45] as well as standard approaches to sparsification. Section 4 formulates the input selection as a "group-sparsity" problem also formulating the predictor estimation problem in our Bayesian framework. Sections 5 and 6 describe the two algorithms we introduce while Section 7 reports some simulation results. Conclusions end the paper.

Notation

The symbols $\mathbb{E}[\cdot]$ denotes expectation while $\hat{\mathbb{E}}[\cdot|\cdot]$ denotes the best linear estimator (conditional expectation in the Gaussian case). In addition for $A \in \mathbb{R}^{n\times m}$, $A^{[ij]}$ will denote the element of A in position (i,j). If A is a vector the notation $A^{[i]}$ will be used in place of $A^{[i1]}$ or $A^{[1i]}$; in addition $A^{[-i]}$ denotes the vector A with the $i-th$ component suppressed. The symbol I denotes the identity matrix of suitable dimensions, $A^\top$ is the transpose of the matrix A and $\|x\|_p$ is the $p-$norm of the vector x. The symbol $\ell_1(\mathbb{Z}^+)$ will denote the space of real infinite sequences (indexed by $\mathbb{Z}^+$) having finite ℓ_1 norm, i.e. the infinite column vector $g := [g_1, g_2, .., g_k, ..]^\top \in \ell_1(\mathbb{Z}^+)$ iff $\sum_{i=1}^{\infty} |g_i| < \infty$.

2 Statement of the problem and notation

Let $\{z_t\}_{t\in\mathbb{Z}}$, $z_t \in \mathbb{R}^m$ be a stationary stochastic processes which models the joint time evolution of some variables of interests. With some abuse of notation the symbol z_t will both denote a random variable (from the random process $\{z_t\}_{t\in\mathbb{Z}}$) and its sample value. We can think of each component of the vector process $\{z_t\}$ as being attached to the node of a network. Our purpose is to build linear dynamical models which describe dynamically each of the components of $\{z_t\}$ as a function of the others. To this purpose we define $y_t := z_t^{[i]}$ (the $i-th$ component of z_t) as "output" and all the others $u_t := z_t^{[-i]} \in \mathbb{R}^{m-1}$ as "inputs". Of course the argument can be repeated for $i = 1, .., m$ thus obtaining a description of all the variables in z_t as a function of the others. Throughout the paper we shall make a specific choice of i which, w.l.o.g., can be taken equal to 1 so that

$$z_t := \begin{bmatrix} y_t \\ u_t \end{bmatrix} \tag{1}$$

This sort of notation is standard in modeling feedback interconnections (see e.g. [22,20,10]) where one concentrates on one variable viewing the others as "inputs", with the assumption that the overall interconnection is such that the joint process is stationary. Also the absence of direct feedthrough terms (i.e. $f_0 = 0$ in (2)) makes life a bit easier (see e.g.

[54]) in that under mild excitation conditions it guarantees identifiability.

From stationarity of $\{z_t\}_{t\in\mathbb{Z}}$ it follows that $\{y_t\}_{t\in\mathbb{Z}}$ and $\{u_t\}_{t\in\mathbb{Z}}$ are jointly stationary stochastic processes which can be thought of, respectively, as the output and input of an unknown time-invariant dynamical system [3]:

$$y_t = \sum_{k=1}^{\infty} f_k u_{t-k} + \sum_{k=0}^{\infty} g_k e_{t-k} \tag{2}$$

were $f_k \in \mathbb{R}^{1\times m}$ and $g_k \in \mathbb{R}$ are (matrix) coefficients of the unknown impulse responses and e_t is the innovation sequence, i.e. the one step ahead linear prediction error

$$\begin{aligned} e_t &:= y_t - \hat{y}_{t|t-1} \\ &:= y_t - \hat{\mathbb{E}}[y_t|y_{t-1}, y_{t-2}, \ldots, u_{t-1}, u_{t-2}, \ldots] \end{aligned} \tag{3}$$

where

$$\begin{aligned} &\hat{\mathbb{E}}[y_t|y_{t-1}, y_{t-2}, \ldots, u_{t-1}, u_{t-2}, \ldots] \\ &:= \textstyle\sum_{j=1}^{m-1} \left[\sum_{k=1}^{\infty} h_k^{[j]} u_{t-k}^{[j]}\right] + \sum_{k=1}^{\infty} h_k^{[m]} y_{t-k}. \end{aligned}$$

The sequences $h_k := [h_k^{[1]}, .., h_k^{[m-1]}, h_k^{[m]}] \in \mathbb{R}^{1\times m}$, $k \in \mathbb{Z}^+$ are the predictor impulse response coefficients and are required to describe (BIBO) stable systems, i.e. $h^{[m]} \in \ell_1(\mathbb{Z}^+)$.

In the prediction error minimization (PEM) framework identification of the dynamical system in (2) can be framed as estimation of the predictor impulse responses h_k in (3) from a finite set of input-output data. We specifically address situations in which m is large as compared to the number of available data and only few variables are in fact needed to predict y_t. Mathematically this means that $h_k^{[i]} = 0$, $\forall k \in \mathbb{Z}^+$. In a graphical representation there will be a directed link from the node representing $u_k^{[i]}$ to that representing y_k if and only if $\exists k \in \mathbb{Z}^+ :\ h_k^{[i]} \neq 0$, $i = 1, .., m-1$; in addition there is a self loop if and only if $\exists k \in \mathbb{Z}^+ :\ h_k^{[m]} \neq 0$. For instance for the network represented in Figure 1, $h_k^{[5]} = h_k^{[1]} = 0$, $\forall k \in \mathbb{Z}^+$ while $h_k^{[2]}, h_k^{[3]}, h_k^{[4]}$ and $h_k^{[6]}$ are not identically zero, meaning that for prediction of y_t one needs (only) the past of $u^{[2]}, u^{[3]}, u^{[4]}$ and of y itself.

In practice one does not know whether a measured signal is significant for prediction of y_t. Standard PEM methods [32,49] do not attempt to perform input selection and estimate a "full" model which uses all inputs. As we shall see this may yield poor results when the number of inputs becomes large as compared to the data available. Variable selection methods has been subject of intense research; classical methods can be found in the books [58,26] while we refer to the survey [25] for a more recent overview.

[3] In order to streamline notation we shall assume one delay from u_t to y_t. If this is true for all possible decompositions $y_t = z_t^{[i]}$, $u_t = z_t^{[-i]}$, $i = 1, .., m$, it can be shown that the interconnection is well posed. Of course to achieve stationarity further restrictions have to be imposed.

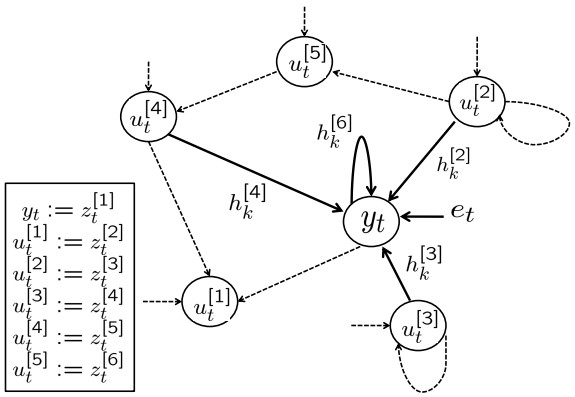

Fig. 1. A dynamical network representing the interaction between $m = 6$ variables. The solid edges represent the links related to the dynamical model for node $y_t := z_t^{[1]}$ given all the others. With reference to equation (3), absence of links from $u_t^{[i]} = z_t^{[i+1]}$, $i = 1,5$ to $y_t := z_t^{[1]}$ means that $h_k^{[1]} = h_k^{[5]} = 0,\ \forall k \in \mathbb{Z}^+$. The node containing y_t has an "entering" arrow which represents the influence of e_t (the one step ahead prediction error of y_t). The dotted edges refer to other decompositions of the form (1) where $y_t = z_t^{[j]}$, $u_t = z_t^{[-j]}$ for $j \neq 1$.

In this paper we shall be specifically concerned with methodologies which, favoring sparsity, will be able to capture the structure of a dynamical network, like the one Figure 1, and at the same time estimate all the (non-zero) impulse responses $h_k^{[i]}$ in (3).

3 Preliminaries: kernels for system identification and sparsity inducing priors

3.1 Bayesian estimation and Kernel-based regularization

Consider the problem of reconstructing a single unknown function h from indirect noisy measurements. In the framework of Gaussian regression, the key point is to interpret h as (a realization of) a zero-mean Gaussian process whose covariance (also called kernel) encodes the available prior knowledge.

Just for a while, it is now useful to think of h as a continuous-time signal. Often, the only available prior knowledge is the fact that h, and possibly some of its derivatives, are continuous with bounded energy. Hence, one often models h as the p-fold integral of white noise. If the white noise has unit intensity, the autocorrelation of the process h, with

domain restricted to the unit interval, is W_p where

$$W_p(s,t) = \int_0^1 G_p(s,u)G_p(t,u)du, \quad s,t \in [0,1] \tag{4}$$

$$G_p(r,u) = \frac{(r-u)_+^{p-1}}{(p-1)!}, \quad (u)_+ = \begin{cases} u \text{ if } u \geq 0 \\ 0 \text{ if } u < 0 \end{cases} \tag{5}$$

The autocovariance W_p is associated with the Bayesian interpretation of the p-th order smoothing splines [55]. In particular, when $p = 2$, one obtains the cubic spline kernel.

Now, it is useful to recall that, when data become available, the Bayes estimate of h belongs to a reproducing kernel Hilbert space (RKHS) $\mathscr{H}$ defined by the covariance of h [4], Such space is equipped with a norm that, as also illustrated in the sequel, controls the complexity of the function to reconstruct, regularizing the estimation process. For instance, as described in [55], the cubic spline kernel is associated with a particular Sobolev space equipped with the norm

$$\|h\|_{\mathscr{H}}^2 = \int_0^1 (h^{(2)}(s))^2 ds \tag{6}$$

Thus, the Gaussian prior associated with W_2 introduces information on the smoothness of h via a regularization term given by the energy of the second-order derivative of h.

3.2 Stable spline kernels

In the system identification scenario, the main drawback of the covariances (4) is that they do not account for impulse response stability, as illustrated in Fig. 2 (left) which displays 100 realizations using an autocovariance proportional to W_2. In fact, if the autocovariance of h is W_p, the variance of $h(t)$ is zero at $t = 0$ and tends to ∞ as t increases. However, if f represents a stable impulse response, one should let it have a finite variance at $t = 0$ which goes exponentially to zero as t tends to ∞. Following [45], this property can be ensured by modeling h via stable spline kernels defined by

$$K_p(s,t) = W_p(e^{-\beta s}, e^{-\beta t}), \quad s,t \in \mathbb{R}^+ \tag{7}$$

where β is a positive scalar governing the decay rate of the variance [45,43]. In real applications, β will be unknown so that, in what follows, it is treated as a hyperparameter to be estimated from data.
When $p = 2$ the autocovariance becomes the Stable Spline kernel introduced in [45]:

$$K_2(t,\tau) = \frac{e^{-\beta(t+\tau)}e^{-\beta\max(t,\tau)}}{2} - \frac{e^{-3\beta\max(t,\tau)}}{6} \tag{8}$$

and the following result holds.

Proposition 1 *[45] Let h be zero-mean Gaussian with autocovariance K_2. Then, with probability one, the realizations of h are continuous impulse responses of BIBO stable dynamic systems.*

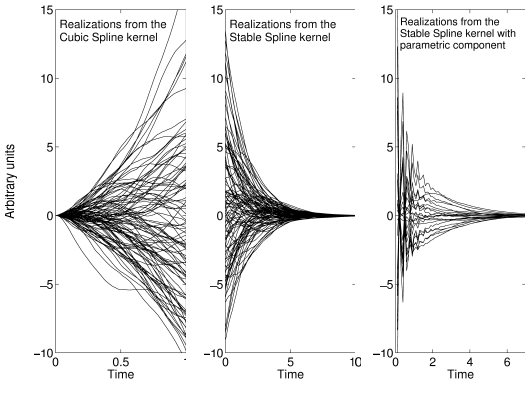

Fig. 2. Realizations of a stochastic process h with autocovariance proportional to the standard Cubic Spline kernel (left), the new Stable Spline kernel (middle) and its sampled version enriched by a parametric component defined by the poles $-0.5 \pm 0.6\sqrt{-1}$ (right).

The effect of the stability constraint is now illustrated in Fig. 2 (middle) which displays 100 realizations drawn from a zero-mean Gaussian process whose autocovariance is proportional to K_2 with $\beta = 0.4$.

A full characterization of the RKHS induced by the stable spline kernels can be found in [43]. Here, we just recall that the kernel K_2 induces the following norm

$$\|h\|_{\mathscr{H}}^2 = \int_0^\infty (h^{(2)}(s) + \beta h^{(1)}(s))^2 \frac{e^{3\beta s}}{\beta^3} ds \tag{9}$$

In this way, the Gaussian prior defined by K_2 defines a penalty term on h that not only forces the energy of the derivatives to be bounded, but also requires them to decay to zero at least exponentially. Hence, information on both smoothness and exponential stability of h are introduced in the stochastic model.

3.3 Prior for predictor impulse responses

Coming back to our original problem, instead of one unknown function, our aim is to estimate the set $\{h^{[i]}\}$ of discrete-time impulse responses. Then, we model them as sampled versions of continuous-time and independent zero-mean Gaussian processes. As in [44], their autocovariances are defined by an "enriched" version of K_2 and share the same hyperparameters, apart from the scale factors $\{\lambda_i^2\}$. More specifically, each discrete-time impulse response $h^{[i]}$ is the convolution of a zero-mean Gaussian process, with autocovariance given by the sampled version of $\lambda_i^2 K_2$, with a parametric impulse response r used to capture dynamics hardly represented by a smooth process such as

high-frequency oscillations. The zeta-transform $R(z)$ of r is parametrized as follows

$$R(z) = \frac{z^2}{P_\theta(z)}, \qquad P_\theta(z) = z^2 + \theta_1 z + \theta_2, \qquad \theta \in \Theta \subset \mathbb{R}^2 \tag{10}$$

where the feasible region Θ constraints the two roots of $P_\theta(z)$ to belong to the open left unit semicircle in the complex plane. The role of the finite-dimensional component of the model is illustrated in Fig. 2 (right panel). Here, we display some realizations (with samples linearly interpolated) drawn from a discrete-time zero-mean normal process with autocovariance given by K_2 and enriched using $P_\theta(z) = z^2 + z + 0.61$ in (10). This corresponds to introducing high-frequency dynamics in the realizations by enriching the Stable Spline kernel with the poles $-0.5 \pm 0.6\sqrt{-1}$. The autocovariance of each predictor impulse response $h^{[i]}$, defined by (8) and (10), is denoted by $K : \mathbb{N} \times \mathbb{N} \mapsto \mathbb{R}$ so that one has

$$\mathbb{E}[h_\ell^{[i]} h_k^{[i]}] = \lambda_i^2 K(\ell, k; \theta, \beta), \quad i = 1, \ldots, m, \quad \ell, k \in \mathbb{N} \tag{11}$$

3.4 Sparsity inducing priors

Differently from the previous subsections, we now discuss a finite-dimensional estimation problem where the goal is to reconstruct the parameter $\phi \in \mathbb{R}^m$ in the linear model

$$y_i = X_i^\top \phi + e_i, \qquad i = 1, \ldots, T \tag{12}$$

In (12), $\{X_i \in \mathbb{R}^m\}$ are the T "regression vectors" while $\{e_i\}$ is zero-mean Gaussian noise of variance σ^2.
When the number m of regressors is very large, e.g. as compared to the number T of data available, obtaining accurate and stable predictors and easily interpretable models becomes a challenging issue which has been quite extensively addressed in the statistical literature in the last decade, see e.g. [51,6,53,26,17,19,9] and references therein.
A pioneering work in this direction has been the so called Lasso (Least Absolute Shrinkage and Selection Operator) [51] that performs regressor selection solving a problem of the form

$$\hat{\phi} := \arg\min_\phi \sum_{i=1}^{T} (y_i - X_i^\top \phi)^2 + \gamma \|\phi\|_1. \tag{13}$$

where the positive scalar γ is the so called regularization parameter. Notice that the problem is finite-dimensional and the penalty term involves the ℓ_1 norm in place of the squared norm in a RKHS. In a Bayesian framework, this difference stems from the fact that ϕ is no longer modeled as a Gaussian process as in the previous section. In particular, $\hat{\phi}$ can be seen as the Maximum a Posteriori (MAP) estimator once the random vector ϕ is independent of the measurement noise and is assigned a double exponential-type prior

$$\mathbf{p}(\phi) \propto e^{-\xi \|\phi\|_1}, \tag{14}$$

yielding

$$\begin{aligned} \hat{\phi} &:= \arg\max_\phi \mathbf{p}(\{y_i\}|\phi)\mathbf{p}(\phi) \\ &= \arg\max_\phi e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{T}(y_i - X_i^\top \phi)^2} e^{-\xi\|\phi\|_1} \end{aligned} \tag{15}$$

that is equivalent to problem (13) once γ is set to $2\xi\sigma^2$. Despite its nice properties, it has been argued that Lasso had not had a significant impact in statistical practice due to its relative computational inefficiency, see [34]. The Least Angle Regression (LAR) algorithm [17] has provided a new approach to regressor selection and, with minor modifications (the "Lasso modification", [17]), also an efficient implementation of the Lasso.
Recently the Lasso has been proposed for estimation of regression models with autoregressive noise [57] and for Vector Autoregressive with eXogenous inputs (VARX) models [28]. This is a rather straightforward application once the regressor vectors $\{X_i\}$ in (13) are formed with past inputs and outputs and ϕ contains the parameters of the finite memory predictors (ARX models).
Another avenue which has been put forward in the statistics literature adopts a Bayesian point of view by modeling the components of ϕ as independent Gaussian random variables $\mathbf{p}(\phi^{[i]}|\lambda_i) = \mathcal{N}(\phi^{[i]}; 0, \lambda_i^2)$ where

$$\mathcal{N}(\phi; m, \lambda^2) = \frac{1}{\sqrt{2\pi\lambda^2}} e^{-\frac{1}{2}\frac{(\phi-m)^2}{\lambda^2}}.$$

A second layer is then added to the model by assuming that also the λ_i's are random variables with a certain density $\mathbf{p}(\lambda_i)$. It follows that

$$\mathbf{p}(\phi) = \prod_i \int \mathbf{p}(\phi^{[i]}|\lambda_i)\mathbf{p}(\lambda_i) d\lambda_i \tag{16}$$

which is a so-called "scale-mixture" distribution [3,59,41,24]. It is well known [3,59,41] that, if λ_i^2 has an exponential distribution itself, then $\mathbf{p}(\phi)$ in (16) has the "double exponential" form (14). This is also related to the so called "Relevance Vector Machine" introduced in [53] which, however, uses a Gamma-type of prior on λ_i^{-2}.

3.5 Concluding remarks of the section

We have introduced two different signal priors that lead to two different regularization terms. The first one derives from Gaussian assumptions and corresponds to a penalty on the signal given by the squared norm associated with the kernel K in (11), hereby denoted by $\|\cdot\|^2_{\mathscr{H}_K}$. This penalty term is suited for identification of infinite-dimensional discrete-time impulse responses. The other one has been introduced in a finite-dimensional context and derives from a double exponential-type prior which leads to the ℓ_1 norm underlying the LASSO.

In the rest of the paper we shall be concerned with a version of the problem (12) where these two types of norms may interact. In particular, each of the components $\phi^{[i]}$ of ϕ will become an unknown impulse response $h^{[i]}$ modeled as a zero-mean Gaussian Process with covariance K. Hence, the Bayes estimate of each $h^{[i]}$ will belong to an infinite-dimensional RKHS denoted by $\mathscr{H}_K$. The regressor vectors $\{X_i\}$ will become linear operators whose representation has infinitely many columns and will contain the past histories of u and y. Details are found in the next section.

4 Variable selection as group sparsity and the sparse identification problem

4.1 The sparse identification problem

In this section we shall see how variable selection can be posed as the problem of obtaining sparse solutions of a linear problem similar to (12) discussed in Section 3.4. There are, however, a few notable differences which makes this, in our opinion, a non-trivial extension of previous results. In particular:

(a) Since we are interested in performing variable selection, we would like that certain impulse responses to be identically zero. This is a sort of "group" problem, similar to those discussed in [61]; however our "groups" are the impulse responses $h^{[i]}$. In a parametric scenario (i.e. when the impulse response are modeled in finite dimensional model classes, see e.g. [32,49]) each group of parameters would describe one impulse response. If we restrict our interest to ARX/FIR models this naturally yields to an algorithm for variable selection which we shall call "ARX-GLAR" since we shall exploit the "group LAR" algorithm (as an alternative, one could also use the "group Lasso" [61]). In general however, the parametrization is non-linear and, in addition, a further model selection problem would have to be faced related to the complexity (order) of the parametric class describing each impulse response. We prefer to work in the nonparametric scenario described in Section 3 so that the "groups" live in an infinite dimensional space.

(b) The unknown "parameters" are the (infinite dimensional) impulse responses modeled as Gaussian Processes. This follows the framework developed in the first subsections of Section 3 and yields to a problem formulation similar to multiple kernel learning [5].

For our purposes, it is useful to set up some notation. Let us define

$$y_t^- := [y_{t-1}, y_{t-2}, y_{t-3}....], \quad u_t^- := [u_{t-1}, u_{t-2}, u_{t-3}....]$$
$$y_t^+ := \begin{bmatrix} y_t \\ \vdots \\ y_{t+T-1} \end{bmatrix}, \quad e_t^+ := \begin{bmatrix} e_t \\ \vdots \\ e_{t+T-1} \end{bmatrix}, \quad h := \begin{bmatrix} h^{[1]} \\ \vdots \\ h^{[m]} \end{bmatrix}; \tag{17}$$

where $h^{[i]}$, $i = 1,..,m$, are impulse responses of stable systems. We also define $A_{ti} \in \mathbb{R}^{T \times \infty}$, $i = 1,..,m$, where

$$\begin{aligned} A_{ti}^{[jk]} &:= u_{t-j-k}^{[i]}, \qquad i = 1,..,m-1 \\ A_{tm}^{[jk]} &:= y_{t-j-k}, \qquad j,k \in \mathbb{Z}^+ \end{aligned} \tag{18}$$

In practice, the operators $\{A_{ti}\}$ above are never completely known since we assume that the measurements y_t, u_t are only taken in an interval of the form $t \in [1,N]$. However, we will think of each A_{ti} as known setting to zero the unobserved entries. We also let the positive integer t_0 denote a positive instant sufficiently large to capture the dynamics of the predictor and define

$$y^+ := y_{t_0}^+, \qquad e_{t_0}^+ := e^+, \qquad A_i := A_{t_0 i} \tag{19}$$

In this way, the predictor in (3) can be rewritten [4] as:

$$y^+ = \underbrace{\begin{bmatrix} A_1 & \dots & A_m \end{bmatrix}}_{:=A} h + e^+ \tag{20}$$

Now, our identification problem corresponds to estimating h in (20), subject to the stability constraints $h^{[i]} \in \ell_1(\mathbb{Z}^+)$, $i = 1,..,m$. Recall that we are interested in estimators which automatically select, among $u^{[1]},..,u^{[m-1]},y$, the variables which are useful for predicting y and which are not. In other words, certain impulse responses $\hat{h}^{[i]}$ are expected to be exactly zero. As said, solving this problem entails estimation in "grouped" variables [61,60] but a peculiarity here is that each "group" lives in an infinite dimensional space.

4.2 A (non sparse) predictor estimator using Gaussian Regression

Under the framework developed in Section 3.3, we assume that the impulse responses $h^{[i]}$ are zero-mean Gaussian processes with covariance function K in (11) [5]. Hence, the prob-

[4] The product of semi-infinite matrices should be intended as the limit of finite sequences. However, given the assumption $h^{[i]} \in \ell_1(\mathbb{Z}^+)$ the limit operation is well posed and, as such, we can formally work with the limiting expressions (see [44]).

[5] When not needed, in order to simplify notations we shall omit the explicit dependence on θ and β.

lem of estimating the impulse responses $h^{[i]}$ from measured data $\{y_t, u_t\}$ can be formulated as the minimum variance estimator

$$\hat{h}^{[i]} = \mathbb{E}[h^{[i]} | \{y_t, u_t\}], \quad i = 1, .., m;$$

Equivalently, one can assume that $h^{[i]}$ are functions in $\mathscr{H}_K$, the reproducing Kernel Hilbert space associated to the sampled Kernel K. Under suitable hypotheses discussed in [44] and also later on in Section 6, one has

$$\{\hat{h}^{[i]}\}_{i=1}^m = \underset{\{h^{[i]} \in \mathscr{H}_K\}_{i=1}^m}{\arg\min} \quad \|y^+ - \sum_{i=1}^m A_i h^{[i]}\|^2 + \sigma^2 \sum_{i=1}^m \frac{\|h^{[i]}\|^2_{\mathscr{H}_K}}{\lambda_i^2} \tag{21}$$

where $\|\cdot\|$ is the Euclidean norm. Notice that in (21), each σ^2/λ_i^2 represents a regularization parameter that trades fit $y_t - \hat{y}_{t|t-1}$ vs. regularity of $h^{[i]}$.
In [44], the estimator (21) has been shown to be very competitive with respect to established identification methods such as PEM and subspace methods. However, in the context of the present paper, a limitation of this estimator is that it does not induce sparse solutions since it exploits quadratic criteria to define both the loss and the penalty terms.

4.3 Sparsifying the predictor estimator

Motivated from the above discussion, the aim now is to introduce two different approaches to sparsify the estimator (21). They are:

(i) SS-GLAR: a "group version" [61] of (13) extended to a non-parametric setup where the "groups" $h^{[i]}$ are modeled as Gaussian Processes with autocovariance equal to the stable spline kernel (8); including a "Laplace-type" prior which enforces sparsity (see Section 3.4) leads us to a mixed $\ell_1 - \ell_2$ regularization problem which can be seen as a "group" version of the so-called "elastic-net" [65]. It is well known that the ℓ_2 penalty in the elastic net helps in selecting groups of correlated variables [65]. Details will be given in Section 5.

(ii) SSEH: a hierarchical model where $h^{[i]}$ is a Gaussian Process with covariance $\lambda_i^2 K(s,t)$ and the hyperparameters $\{\lambda_i\}$ have an exponential distribution. Differently from the previous approach, notice that this will favor sparsity on the space of scale factors. As mentioned in [11], this is also related to multiple kernel learning, see also [15]. This second technique will actually allow us to introduce more flexibility in the Kernels. In fact, we will model each $h^{[i]}$ using K in (11) that corresponds to the stable spline kernel (8) enriched with the parametric component (10); as argued in [44] this may be advantageous in situations where the impulse responses contain "fast" dynamics penalized by the regularization term, see also [43]. Details will be given in Section 6.

5 Stable Splines Group LAR (SSGLAR) algorithm

5.1 Enforcing sparsity using the GLAR algorithm

In this section we shall discuss how to modify (21) to enforce sparsity on the groups $h^{[i]}$ using the GLAR algorithm [61]. In order to do so we shall have to assume the parameter θ in (10) has been fixed (without any prior information it will be fixed equal to zero) and that all the kernel scale factors are equal each other, i.e. $\lambda = \lambda_i$ for $i = 1, \ldots, m$. In addition, it is worth recalling that the norm $\|h^{[i]}\|^2_{\mathscr{H}_K}$ admits a "matrix" representation of the form

$$\|h^{[i]}\|^2_{\mathscr{H}_K} = \left[h^{[i]}\right]^\top \Lambda h^{[i]} \tag{22}$$

where $\Lambda \in \mathbb{R}^{\infty \times \infty}$ can be thought of as the "inverse" of the matrix representation of the Kernel $K \in \mathbb{R}^{\infty \times \infty}$. The matrix Λ is symmetric an positive definite, thus admitting a square root $\Lambda^{1/2}$ such that $\Lambda = \Lambda^{1/2}\Lambda^{1/2}$. [6]
Now, in place of (20), our measurements model is modified as follows

$$\bar{y}^+ = \sum_{i=1}^m \bar{A}_i h^{[i]} + e^+ \tag{23}$$

where

$$\begin{aligned} \bar{y}^+ &:= \begin{bmatrix} y^+ \\ 0_{1\times(\infty\cdot m)} \end{bmatrix} \\ \bar{A}_i &:= \begin{bmatrix} A_i & \chi_i \otimes \sqrt{\gamma}\Lambda^{1/2} \end{bmatrix}, \qquad \gamma = \frac{\sigma^2}{\lambda^2} \\ \chi_i &:= [\underbrace{0 \ \ldots \ 0}_{i-1} \ 1 \ \underbrace{0 \ \ldots \ 0}_{m-i}]^\top \end{aligned} \tag{24}$$

Note that the measurement model (23) is designed so as to include the ℓ_2-type regularization term in (21), which can be written as in equation (22).

Performing input selection can be tackled, as discussed in Section 3.4, via the Group Least Angle Regression algorithm in [17] applied to the regression problem (23). We shall call SS-GLAR (Stable Spline Group Least Angle Regression) the resulting algorithm which we now summarize:

Algorithm: Stable Spline Group Least Angle Regression (SS-GLAR)

(1) fix the parameter β in (11);
(2) fix the parameter γ in (24); form the regressor $\bar{A}_i$ in (23) as described in formulas (18), (19), (24) ;
(3) estimate $h^{[i]}$ applying the GLAR algorithm to problem (23);

[6] Given a nondegenerate Borel measure ν on $\mathbb{N}$, such operator $\Lambda^{1/2}$ is always well defined and corresponds to the matrix form of the square root of the operator L_K mapping $h \in \mathscr{H}_K$ into the function $\int_{\mathbb{N}} K(s,t)h(t)d\nu(t)$, see Section 1 in [48] for details.

5.2 Estimation of the hyper-parameters

Note that, in order to run the previous algorithm, the following parameters have to be chosen:

(a) the scale factor γ of the ℓ_2 penalty in (24) (regularity of $h^{[i]}$ in the space $\mathscr{H}_K$)
(b) the parameter β in (11) (decay rate of the Kernel)
(c) the number of non-zero blocks estimated via the GLAR algorithm.

These can be estimated using a validation based approach as follows: Let $\{y_t, u_t\}_{t=1,..,N}$ be the available data. We split the data set in two parts. We call *identification data set* $\{y_t, u_t\}_{t=1,..,\lfloor 2N/3 \rfloor}$ and *validation data set* $\{y_t, u_t\}_{t=\lceil 2N/3 \rceil,..,N}$. We run the identification algorithms on the identification data set fixing the hyperparameters and computing the entire "GLARS path" [61] which consists, for each choice of hyperparameters, of m models differing by the number of non-zero blocks. We grid the hyperparameter space ($\beta \in \mathbb{R}^+$, $\gamma \in \mathbb{R}^+$) so that only a finite (and possibly small) number of alternatives is tested [7].

The "best" hyperameters and level of sparsity is then selected testing all these models on the validation data set, performance being measured by the root-mean-squared error in one-step-ahead prediction error RMS_1, where RMS_k, $k = 1, 2, ...$ is defined as:

$$RMS_k := \sqrt{\frac{3}{N} \sum_{t=\lfloor \frac{2N}{3} \rfloor + 1}^{N} (y_t - \hat{y}_{t|t-k})^2} \tag{25}$$

Then the hyperparameter vector and the level of sparsity are fixed and the model is re-estimated with all data $\{y_t, u_t\}_{t=1,..,N}$.

6 Stable Splines with Exponential Hyperprior (SSEH) Algorithm

Recall that the estimator (21) is known up to the following parameters:

- the noise variance σ^2;
- the scale factors λ_i (in fact recall that $h^{[i]}$ are Gaussian processes with covariance $\lambda_i^2 K(t,s)$);
- β that enters the kernel K and is related to the dominant pole of the predictor;
- θ that represents the parametric part of the model, as defined in (10).

[7] We have chosen a logarithmically spaced grid with 11 values for β and 5 for γ. Experimental evidence shows that the results are not very sensitive to choice of hyperparameters, and finer grids did not yield any significant improvement

In this section we will show how the estimator (21) can be "sparsified" by interpreting all the parameters listed above as random vectors and assigning suitable hyperpriors.

6.1 Hyperprior for the hyperparameters and the full Bayesian model

Our Bayesian model for sparse identification is defined as follows:

- the noise variance σ^2 will always be estimated via a preliminary step using a low-bias ARX model, as described in [23]. Thus, this parameter, even if always determined from data during our numerical experiments, will be assumed known in the description of our Bayesian model;
- the hyperparameters β, θ and $\{\lambda_i\}$ are described as mutually independent random vectors;
- β is given a non informative probability density on $\mathbb{R}^+$;
- θ has a uniform distribution on the feasible region Θ that constraints the two roots of $P_\theta(z)$ to belong to the open left unit semicircle in the complex plane, see (10);
- each λ_i is an exponential random variable with inverse of the mean (and standard deviation) $\xi \in \mathbb{R}^+$, i.e.

$$\mathbf{p}(\lambda_i) = \xi \exp(-\xi \lambda_i)\, \chi(\lambda_i \geq 0), \qquad i = 1, \ldots, m \tag{26}$$

with χ the indicator function. We also interpret ξ as a random variable with a non informative prior on $\mathbb{R}^+$. Notice that, differently from the approach described in the previous section, the parameters λ_i are now allowed to be all different thus increasing the flexibility of our model.

In what follows, ζ indicates the hyperparameter random vector, i.e. $\zeta := [\lambda_1, \ldots, \lambda_m, \theta_1, \theta_2, \beta, \xi]$.

To simplify the notation, we define $y^- := [y_{t_0}, y_{t_0-1}, y_{t_0-2}, \ldots]^T$ and $u^- := [u_{t_0}, u_{t_0-1}, u_{t_0-2}, \ldots]^T$ where the unobserved entries are set to zero. In addition, $u^+ := u_{t_0}^+$ and recall that $y^+ := y_{t_0}^+$. Further, the following approximation is exploited:

$$\begin{aligned} &\mathbf{p}(y^+, \{h^{[i]}\}, y^-, u|\zeta) = \\ &\propto \left[\prod_{t=t_0}^N \mathbf{p}(y_t|\{h^{[i]}\}, y_t^-, \zeta, u_t^-)\right] \mathbf{p}(y^-, \{h^{[i]}\}, u^-|\zeta) \\ &\approx \left[\prod_{t=t_0}^N \mathbf{p}(y_t|\{h^{[i]}\}, y_t^-, \zeta, u_t^-)\right] \mathbf{p}(\{h^{[i]}\}|\zeta)\mathbf{p}(y^-, u^-) \end{aligned} \tag{27}$$

The first $\propto$ stems from the fact that the predictor of u_t given the past u_t^- and y_{t+1}^- is assumed not to depend on ζ. The last approximated equality follows from the assumption that the past y^-, u^- does not carry information on the predictor impulse responses and the hyperparameters. Our stochastic model is described by the Bayesian network in Fig. 3 (left side).

6.2 Estimation of the hyper-parameters

To simplify the notation, the dependence on y^- and u is omitted in the sequel, so that all the probability densities are now thought of as implicitly conditional on y^- and u.

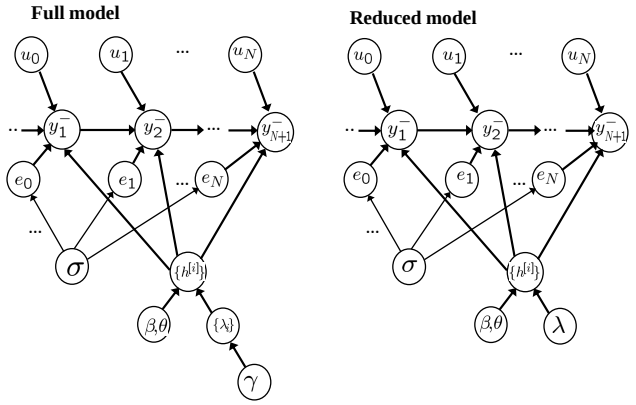

Fig. 3. Bayesian network describing the new nonparametric model for identification of sparse linear systems where $y_l^- := [y_{l-1}, y_{l-2}, \ldots]$ and, in the reduced model, $\lambda := \lambda_1 = \ldots = \lambda_m$.

We start reporting a preliminary lemma, whose proof can be found in [44], which will be needed in propositions 3 and 4.

Lemma 2 *Let the roots of P_θ in* (10) *be stable. Then, if $\{y_t\}$ and $\{u_t\}$ are zero mean, finite variance stationary stochastic processes, each operator $\{A_i\}$ is almost surely (a.s.) continuous in $\mathscr{H}_K$.*

We estimate the hyperparameter vector ζ by optimizing its marginal posterior, i.e. the joint density of y^+, ζ and $\{h^{[i]}\}$ where all the $\{h^{[i]}\}$ are integrated out. This is described in the next proposition that derives from simple manipulations of probability densities whose well-posedness is guaranteed by lemma 2. Below, I_N is the $N \times N$ identity matrix while, with a slight abuse of notation, K is now seen as an element of $\mathbb{R}^{\infty\times\infty}$, i.e. its i-th column is the sequence $K(\cdot, i)$, $i \in \mathbb{N}$. The proof of this proposition follows the same lines as that of Proposition 3 in [42] with minor modifications allowing for the presence of feedback and is therefore omitted.

Proposition 3 *Let $\{y_t\}$ and $\{u_t\}$ be zero mean, finite variance stationary stochastic processes. Then, under the approximation (27), the maximum a posteriori estimate of ζ given y^+ is*

$$\hat{\zeta} = \arg\min_\zeta J(y^+;\zeta) \quad s.t. \quad \theta \in \Theta, \quad \xi, \beta > 0, \quad \lambda_i \geq 0 \quad (i = 1, \ldots, m) \tag{28}$$

where J is almost surely well defined pointwise and, using also (27)*, given by*

$$\begin{aligned} J(y^+;\zeta) &:= \log\left[\int \mathbf{p}(y^+, \{h^{[i]}\}, y^-, u|\zeta)\, dh^{[1]} \ldots dh^{[m]}\right] \\ &\approx \tfrac{1}{2}\log\left(\det[2\pi V[y^+]]\right) + \tfrac{1}{2}(y^+)^T (V[y^+])^{-1} y^+ + \\ &\quad + \xi \textstyle\sum_{i=1}^m \lambda_i - \log(\xi) + const \end{aligned} \tag{29}$$

with $V[y^+] = \sigma^2 I_N + \sum_{i=1}^m \lambda_i^2 A_i K A_i^T$.

Notice that the first term $\frac{1}{2}\log\left(\det[2\pi V[y^+]]\right)$ in the objective (29) penalizes the complexity of the model, in fact it increases as the $\{\lambda_i\}$ get larger. The second term $\frac{1}{2}(y^+)^T (V[y^+])^{-1} y^+$ accounts for adherence of experimental data and decreases as the $\{\lambda_i\}$ augment. The third term is a consequence of the hyperprior (26) whose effect is to include an additional ℓ_1 penalty on $\{\lambda_i\}$. Finally, the last term $\log(\xi)$ derives from the same hyperprior and controls the weight of the ℓ_1 norm which is estimated jointly with the other hyperparameters. Overall, our objective can be interpreted as a Bayesian modified version of that connected with multiple kernel learning, see Section 3 in [15].

An important issue for the practical use of our numerical scheme is the availability of a good starting point for the optimizer. Below, we describe a scheme that achieves a suboptimal solution just solving an optimization problem in $\mathbb{R}^4$ related to the reduced Bayesian model of Fig. 3 (right side). Our main idea is to optimize the objective under the constraint $\lambda_i = \lambda$, for $i = 1, \ldots, m$, and removing the ℓ_1 penalty on $\{\lambda_i\}$. The resulting estimate of λ is used to obtain an estimate of ξ which is then exploited to sparsify the solution. This is described below.

i) Obtain $\{\hat{\lambda}_i\}$, $\hat{\theta}$ and $\hat{\beta}$ solving the following modified version of problem (28)

$$\begin{aligned} &\arg\min_\zeta \left[J(y^+;\zeta) - \xi \textstyle\sum_{i=1}^m \lambda_i + \log(\xi)\right] \\ &\text{s.t. } \theta \in \Theta, \quad \beta > 0, \quad \lambda_1 = \ldots = \lambda_m \geq 0 \end{aligned}$$

ii) Set $\hat{\xi} = 1/\hat{\lambda}_1$ and $\hat{\zeta} = [\hat{\lambda}_1, \ldots, \hat{\lambda}_m, \hat{\theta}, \hat{\beta}, \hat{\xi}]$.
For $i = 1, \ldots, m$ do:
set $\bar{\zeta} = \hat{\zeta}$ except that the i-th component of $\bar{\zeta}$ is set to 0;
if $J(y^+;\bar{\zeta}) \leq J(y^+;\hat{\zeta})$, set $\hat{\zeta} = \bar{\zeta}$.

The procedure we have outlined in step ii) may suffer when the inputs are highly correlated, possibly making it sensitive to the order in which the components of $\hat{\zeta}$ are set to zero. In order to circumvent this difficulty, an alternative consists of using the following "Bayesian forward-selection" type of algorithm where, at every step, the next variable to be included in the model is that leading to the largest objective's improvement. This is obtained substituting item ii) above with:

ii') Let us denote with I the index set of "selected" variables, define $\tilde{\zeta}_I = [\tilde{\lambda}_1, \ldots, \tilde{\lambda}_m, \hat{\theta}, \hat{\beta}, \hat{\xi}]$ where $\tilde{\lambda}_i = \hat{\lambda}_i$ if $i \in I$ and $\tilde{\lambda}_i = 0$ otherwise. Initialize $I := \emptyset$ and repeat the following procedure:
(a) *for* $j \in \{1, .., m\} \setminus I$, define $I_j' := I \cup j$ and compute $J(y^+; \bar{\zeta}_{I_j'})$.

(b) select

$$\bar{j} := \underset{j\in\{1,..,m\}\setminus I}{arg\,max}\ J(y^+;\bar{\zeta}_{I'_j}) - J(y^+;\bar{\zeta}_I)$$

(c) *if* $J(y^+;\bar{\zeta}_{I'_{\bar{j}}}) - J(y^+;\bar{\zeta}_I) > 0$

set $I := I'_{\bar{j}}$ and go back to (a)

else

finish.

The set I contains the indexes of selected variables and $\bar{\zeta}_I$ is used as a starting point for the optimization problem (29).

Remark 1 *Note that more elaborated procedures for variable selection have been proposed which combine forward and backward (addition and elimination) steps such as those introduced and analyzed in [2,62]. Our main focus here is not however on this specific step and further comparative analysis with the literature is postponed to future work. Let us also stress that the results in [56] provide support for the use of forward selection procedures for screening purposes.*

6.3 Estimation of the predictor impulse responses for known ζ

Once all the unknown parameters are learnt from data following the procedure outlined in the previous subsection, the estimator (21) becomes completely known. Hence, the following result, that comes from the representer theorem whose applicability is guaranteed by lemma 2 (see [44] for details), can be utilized to achieve the unknown predictor impulse responses.

Proposition 4 *Under the same assumptions of Proposition 3, almost surely we have*

$$\{\hat{h}^{[i]}\}_{i=1}^m = \underset{\{h^{[i]}\in\mathscr{H}_K\}_{i=1}^m}{\arg\min} \quad \|y^+ - \sum_{i=1}^m A_i h^{[i]}\|^2 + \sigma^2 \sum_{i=1}^m \frac{\|h^{[i]}\|^2_{\mathscr{H}_K}}{\lambda_i^2}$$

where $\|\cdot\|$ is the Euclidean norm. Moreover, almost surely we also have for $k = 1,\ldots,m$

$$\hat{h}^{[i]} = \lambda_i^2 K A_i^T c, \qquad c = \left(\sigma^2 I_N + \sum_{i=1}^m \lambda_i^2 A_i K A_i^T\right)^{-1} y^+ \tag{30}$$

After obtaining the estimates of the $\{h^{[i]}\}$, simple formulas can then be used to derive the system impulse responses f and g in (2) and hence also the k-step ahead predictors, see [32] for details.

7 Simulation results

We consider three Monte Carlo studies of 300 runs where at any run an ARMAX linear system with 15 inputs is generated as follows

- the number of $h^{[i]}$ different from zero is randomly drawn from the set $\{1,2,..,10\}$.
- Then, the order of the ARMAX model is randomly chosen in $[1,30]$ and the model is generated by the MATLAB function `drmodel.m`. The system and the predictor poles are restricted to have modulus less than 0.95 with the ℓ_2 norm of each $h^{[i]}$ bounded by 10.

For each run in the Monte Carlo experiments an identification data set of size 500 and a test set of size 1000 are generated.

In the first experimental setup a white noise input with uncorrelated components is used. In the second one the input still has uncorrelated components each being generated via the MATLAB function `idinput.m` as a realization from a random Gaussian signal with band [8] $[0,0.8]$ for the identification data and $[0,0.9]$ for the validation data; this clearly makes prediction on new data more challenging. In the third Monte Carlo experiment the inputs used for identification are white but are allowed to be correlated, being generated according to the following model:

$$u_k^{[i+1]} = u_k^{[i]} + v_k^{[i]} \quad i = 1,..,m-2$$

where $\{u_k^{[1]}\}$ is unit variance white noise sequence while $\{v_k^{[i]}\}$ is a white noise sequence, independent of $\{u_k^{[1]}\},\{v_k^{[j]}\}, j<i$ with variance $\varepsilon^2 = 0.04$. With this choice, the correlation coefficient

$$\rho_i := \frac{\mathbb{E}\left(u_k^{[i]}u_k^{[i+1]}\right)}{\sqrt{\mathbb{E}\left(u_k^{[i]}u_k^{[i]}\right)\mathbb{E}\left(u_k^{[i+1]}u_k^{[i+1]}\right)}} = \frac{Var\{u_k^{[i]}\}}{\sqrt{Var\{u_k^{[i]}\}Var\{u_k^{[i+1]}\}}}$$

satisfies

$$\rho_i \in [0.9806, 0.9871] \quad i \in [1,14];$$

Note that correlated inputs renders the input selection problem more challenging. The test set, instead is generated using independent zero mean, unit variance white noises as inputs.

We compare the following estimators:

(1) GLAR: this is the GLAR algorithm described in [61] applied to ARX models; the order (between 1 and 30) and the level of sparsity (i.e. the number of null $h^{[i]}$) is determined using the first $2/3$ of the 500 available data as training set and the remaining part as validation data (the use of C_p statistics does not provide better results in this case).

(2) PEM+Oracle: this is the classical PEM approach, as implemented in the `pem.m` function of the MATLAB

[8] The boundaries specify the lower and upper limits of the passband, expressed as fractions of the Nyquist frequency.

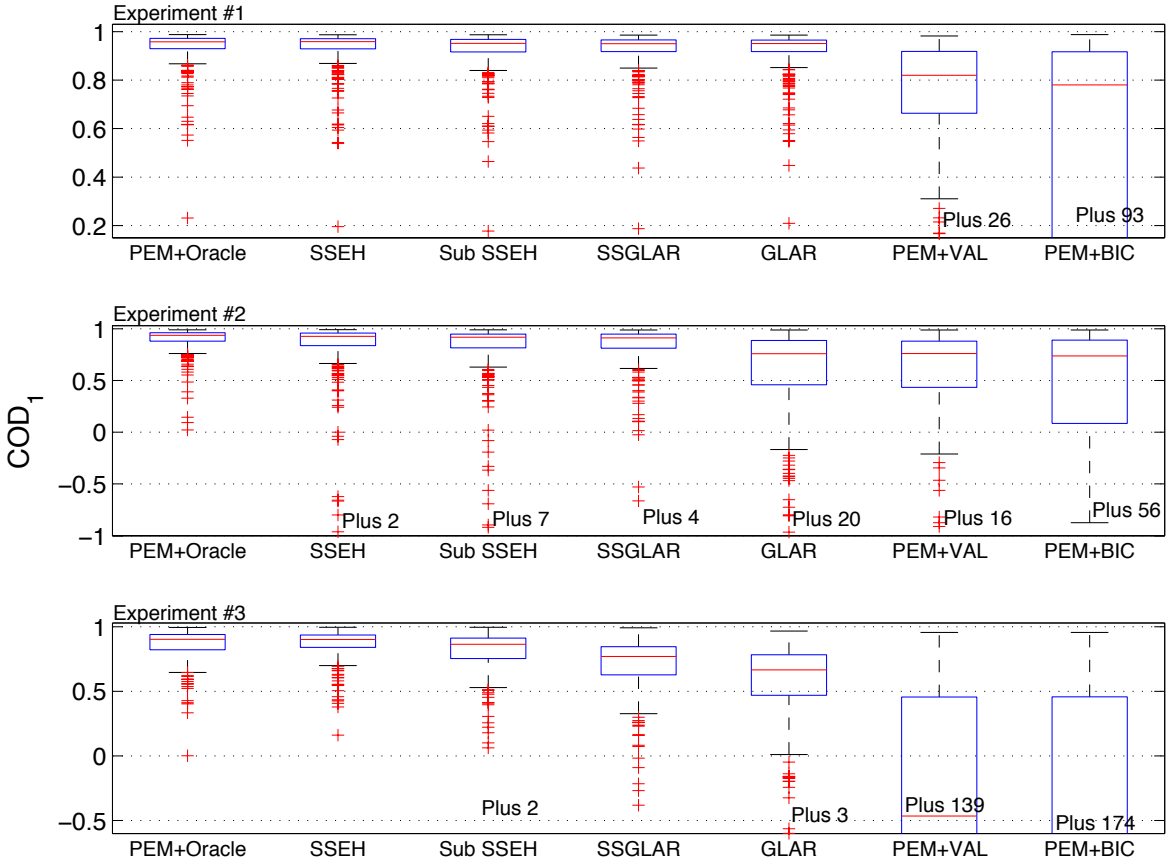

Fig. 4. Boxplot of Coefficient of Determination for one-step-ahead prediction (COD_1). PEM + Oracle is PEM with an oracle who knows which impulse responses are zero and has access to validation data in order to select the best performing system order. The notation "Plus *xx*" means that there are *xx* "outliers" which are left out of the plot.

System Identification Toolbox [33], equipped with an oracle that, at every run, knows which predictor impulse response are zero and, having access to the test set, selects those model orders that provide the best prediction performance.

(3) SSGLAR: this is the approach which combines GLAR and the Stable Spline prior for impulse responses, as detailed in Section 5. The first 40 available input/output pairs enter the $\{A_i\}$ in (20), i.e. $t_0 = 40$ in (19). For computational reasons the predictor length is set to 40, a number that does not establish any trade-off between bias and variance but is just sufficiently large to capture the predictor dynamics.

(4) SSEH: this is the approach described in Section 6, which is based on the full Bayesian model of Fig. 3. As done before, we set both t_0 and the predictor length to 40.

(5) Suboptimal SSEH: the same as above except that we exploit the reduced Bayesian model of Fig. 3 complemented with the procedure described at the end of subsection 6.2, with the refined step ii' used only in the third Monte Carlo experiment.

(6) PEM+VAL: this is the classical PEM approach that uses validation data for model order selection. The order of the polynomials in the ARMAX model are not allowed to be different each other since this would lead to a combinatorial explosion of the number of competitive models.

(7) PEM+BIC: this is the classical PEM approach that uses BIC for model order selection. The order of the polynomials in the ARMAX model are not allowed to be different each other since this would lead to a combinatorial explosion of the number of competitive models.

(8) PEM+BIC+or2: this is the same as PEM + BIC with and additional oracle knowing which impulse responses are zero.

(9) PEM+VAL+or2: this is the same as PEM + BIC + or2 besides the fact that the order is estimated using validation data rather than BIC.

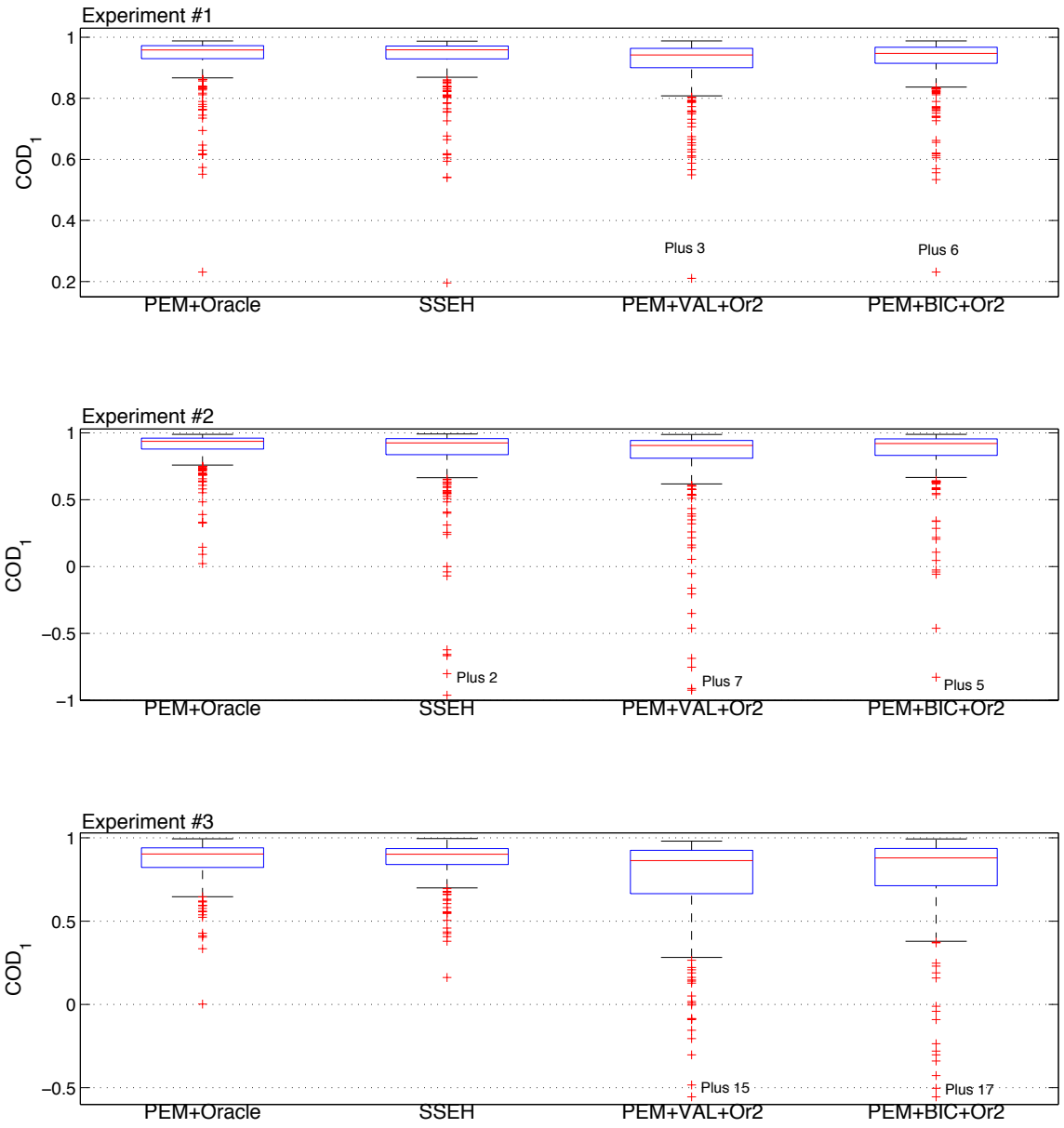

Fig. 5. Boxplot of Coefficient of Determination for one-step-ahead prediction (COD_1). PEM + Oracle is PEM with an oracle who knows which impulse responses are zero and has access to validation data in order to select the best performing system order. PEM + BIC +or2 and PEM +VAL + or2 are equipped with an oracle who knows which impulse responses are zero. The notation "Plus *xx*" means that there are *xx* "outliers" which are left out of the plot.

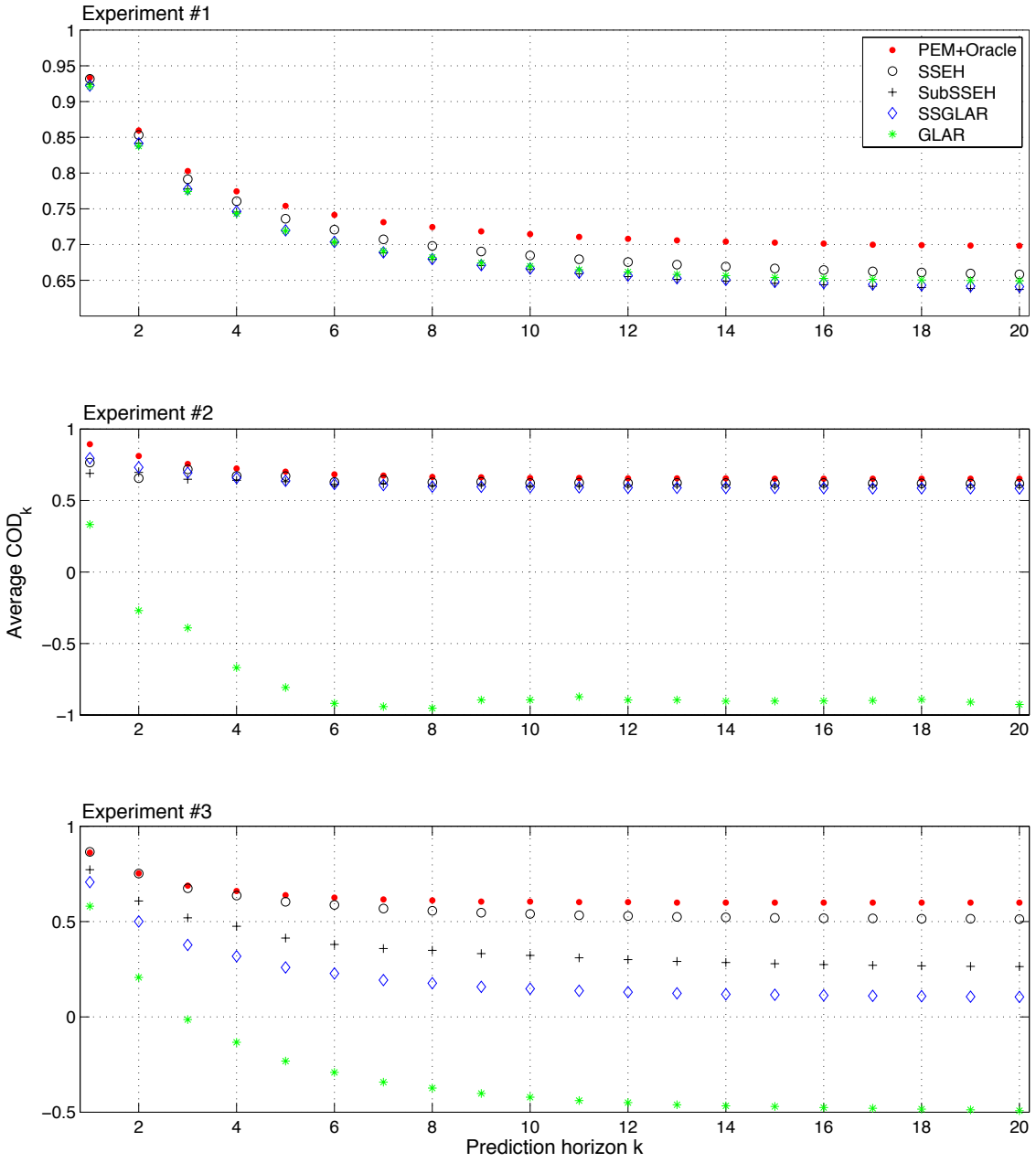

Fig. 6. $\overline{COD}_k$, i.e. average coefficient of determination relative to k-step ahead prediction, obtained during the Monte Carlo study #1 (top), #2 (center) and #3 (bottom), using PEM+Oracle (•), SSEH (○), Suboptimal SSEH (×), SSGLAR (◇) GLAR (∗)

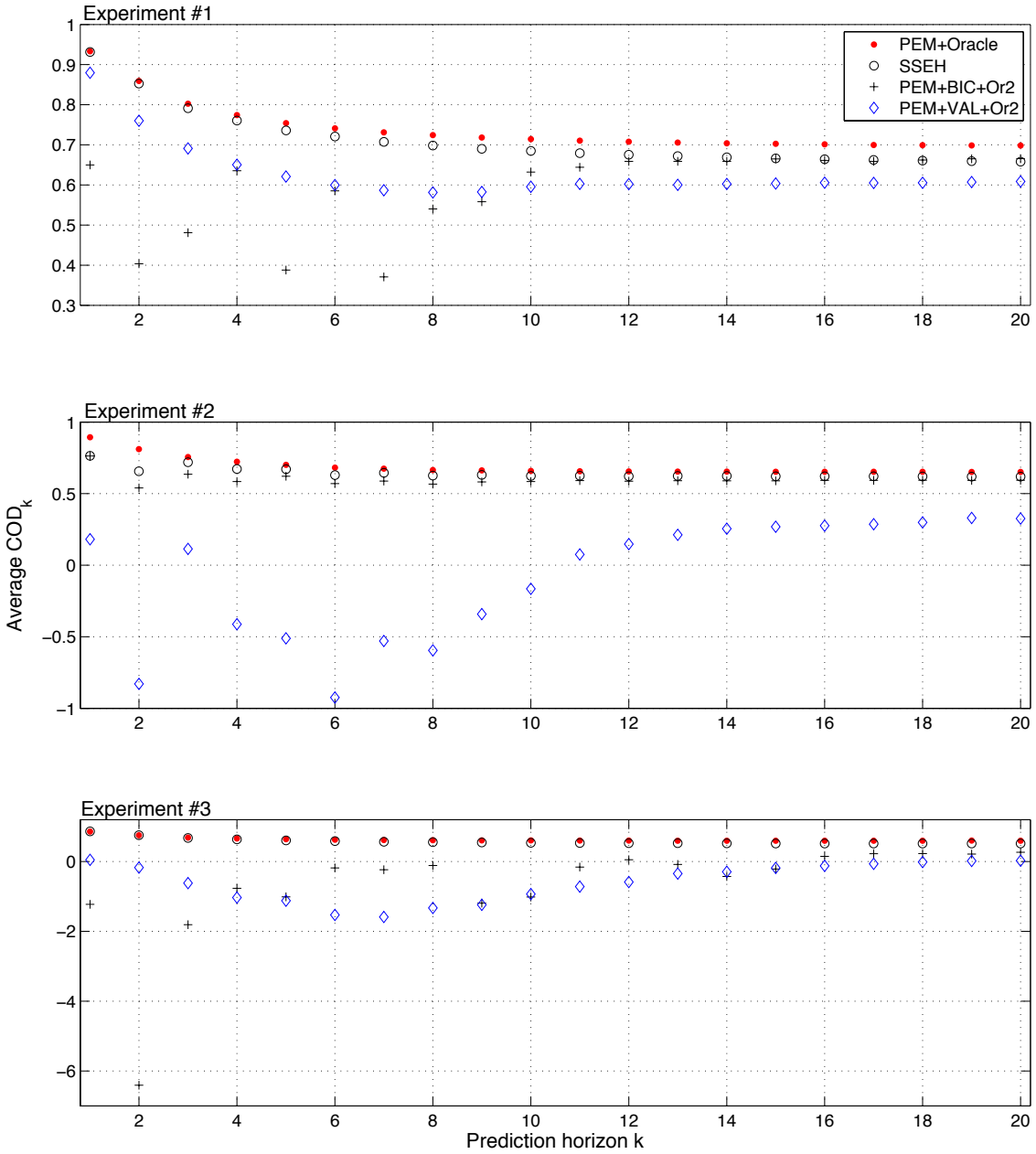

Fig. 7. $\overline{COD}_k$, i.e. average coefficient of determination relative to k-step ahead prediction, obtained during the Monte Carlo study #1 (top), #2 (center) and #3 (bottom), using PEM+Oracle (•), SSEH (○), PEM with order estimated via BIC and knowledge of which impulse responses are zero (×), PEM with order estimated via validation and knowledge of which impulse responses are zero (◇)

Experiment	SSEH	Suboptimal SSEH	SSGLAR	GLAR
#1	98.8%	99.1%	45.93%	63.41%
#2	98.64%	98.39%	49.76%	70.09%
#2	95.05%	92.70%	56.58%	67.16%

Table 1
Percentage of the $h^{[i]}$ equal to zero correctly set to zero by the employed estimator.

The following performance indexes are considered:

(1) Percentage of the impulse responses equal to zero correctly set to zero by the estimator.
(2) k-step-ahead Coefficient of Determination, denoted by COD_k, quantifying how much of the test set variance is explained by the forecast. It is computed at each run as

$$COD_k := 1 - \frac{RMS_k^2}{\frac{1}{1000}\sum_{i=1}^{1000}(y_t^{test} - \bar{y}_t^{test})^2}$$
$$RMS_k := \sqrt{\frac{1}{1000}\sum_{t=1}^{1000}(y_t^{test} - \hat{y}_{t|t-k}^{test})^2} \quad (31)$$

where $\bar{y}^{test}$ is the sample mean of the test set data $\{y_t^{test}\}_{t=1}^{1000}$ and $\hat{y}_{t|t-k}^{test}$ is the k-step ahead prediction computed using the estimated model. The average index obtained during the Monte Carlo study, as a function of k, is then denoted by $\overline{COD}_k$.

Notice that, in both of the cases, the larger the index, the better is the performance of the estimator.
In every experiment the performance of PEM+VAL and PEM+BIC has been largely unsatisfactory, providing strongly negative values for $\overline{COD}_k$. This is illustrated e.g. in Fig. 4 showing the boxplots of the 300 values of COD_1 obtained by the employed estimators on the three Monte Carlo studies. We have also assessed that results do not improve using AIC. In view of this, in what follows other results from PEM+VAL and PEM+BIC will not be shown; for sake of comparison we add in Figs. 7 and 5 comparison with PEM+BIC+or2 and PEM+VAL+or2 which use knowledge of which impulse responses are zero.
Table 1 reports the percentage of the predictor impulse responses equal to zero correctly estimated as zero by the estimators. In terms of predictive performance the Stable Spline estimators (SSEH and SSGLAR) outperform GLAR, with a slight advantage of SSEH; instead, in terms of sparsity of the estimated model, SSEH and Suboptimal SSEH show a definite advantage over GLAR-based techniques, achieving the remarkable performance of 99% correct detection of the "zero" impulse responses.

We conjecture that the superior performance of SSEH can be attributed to the fact that it combines the advantages of the Stable-Spline regularization (giving good performance in prediction, [44]) and those of the exponential hyperprior favoring sparsity. On the other hand note that the SSGLAR algorithm, which combines ℓ_2 and ℓ_1 penalties like the elastic-net, tends to overestimate the number of nonzero impulse responses. While a rigorous explanation of this behavior is the subject of current research, its is worth stressing that the SSEH procedure combines a "forward selection" initialization with an optimization based refinement; this can be seen as an instance of the "screening" procedure analyzed in [56].

Finally, Figs. 6 and 7 display $\overline{COD}_k$ as a function of the prediction horizon obtained during the Monte Carlo study #1 (top), #2 (center) and #3 (bottom). The performance of Stable Spline appears superior than that of GLAR and is comparable with that of PEM+Oracle also when the reduced Bayesian model of Fig. 3 is used.

8 Conclusions

Identification of large scale dynamical systems in the framework of dynamical Bayesian networks has been discussed. It has been argued that estimation of network connectivity and dynamic interaction can be framed as identification of a sparse multi-input, single-output dynamical system. Two new methods have been presented which combine recently developed non-parametric methods for system identification and sparsity-favoring algorithms. The two methods (SSGLAR and SSEH) have been compared via extensive simulation studies with state-of-the art algorithms such as the Group Least Angle Regression (LARS) algorithm applied to ARX models and Prediction Error Methods (PEM). Several simulation setups have been considered including low pass input excitation as well as highly correlated inputs. The advantages of the new methods (especially SSEH) is apparent both in terms of predictive capabilities on new data as well as regarding the ability of detecting the network connectivity (i.e. the percentage of correctly detected zeros). Future work will concentrate on the analysis of the proposed methods. In particular we envision that properties of the so-called multivariate Laplace distribution [18] may be relevant. In addition also dedicated numerical optimization procedures will be developed.

Acknowledgments This research has been partially supported by the PRIN Project "New Methods and Algorithms for Identification and Adaptive Control of Technological Systems", by the Progetto di Ateneo CPDA090135/09 funded by the University of Padova and by the European Communitys Seventh Framework Programme under agreement n. FP7-ICT-223866-FeedNetBack.

References

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

[2] H. An, H. Da, Y. Qiwei, and Z. Cun-Hui. Stepwise searching for feature variables in high-dimensional linear regression. Technical report, Department of Statistics, London School of Economics, 2008.

[3] D.F. Andrews and C.L. Mallows. Scale mixtures of normal distributions. *Journal of the Royal Statistical Society*, 36:99–102, 1974.

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

[5] F.R. Bach. Consistency of the group lasso and multiple kernel learning. *J. Mach. Learn. Res.*, 9:1179–1225, 2008.

[6] S. Bakin. *Adaptive regression and model selection in data mining problems*. PhD thesis, The Australian National University, 1999.

[7] M. Banbura, D. Giannone, and L. Reuchlin. Large Bayesian VARs. Working Paper Series 966, European Central Bank, 2008.

[8] D.R. Brillinger. *Time series: Data analysis and theory*. Holden-Day, 1981.

[9] E. Candes and T. Tao. The Dantzig selector: statistical estimation when p is much larger than n. *Annals of Statistics*, 35:2313–2351, 2007.

[10] A. Chiuso and G. Picci. Consistency analysis of some closed-loop subspace identification methods. *Automatica*, 41(3):377–391, 2005.

[11] A. Chiuso and G. Pillonetto. Learning sparse dynamic linear systems using stable spline kernels and exponential hyperpriors. In *Proceedings of Neural Information Processing Symposium*, Vancouver, 2010.

[12] A. Chiuso and G. Pillonetto. Nonparametric sparse estimators for identification of large scale linear systems. In *Proceedings of IEEE Conf. on Dec. and Control*, Atlanta, 2010.

[13] R. Dahlhaus and M. Eichler. *Highly structured stochastic systems*, chapter Causality and graphical models in time series analysis, pages 115–137. Oxford University Press, 2003.

[14] A.P. Dempster. Covariance selection. *Biometrics*, 28(1):157–175, 1972.

[15] F. Dinuzzo. Kernel machines with two layers and multiple kernel learning. Technical report, Preprint arXiv:1001.2709, 2010. Available at http://www-dimat.unipv.it/ dinuzzo.

[16] D. Donoho. Compressed sensing. *IEEE Trans. on Information Theory*, 52(4):1289–1306, 2006.

[17] B. Efron, T. Hastie, L. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32:407–499, 2004.

[18] T. Eltolft, T. Kim, and T.W. Lee. On the multivariate Laplace distribution. *IEEE Signal Processing Letters*, 13(5):300–303, 2006.

[19] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96:1348–1360, December 2001.

[20] U. Forsell and L. Ljung. Closed loop identification revisited. *Automatica*, 35:1215–1242, 1999.

[21] J. Friedman, T. Hastie, and R. Tibshirani. Applications of the Lasso and grouped Lasso to the estimation of sparse graphical models. Technical report, Stanford University, 2010.

[22] M. Gevers and B.D.O. Anderson. On jointly stationary feedback-free stochastic processes. *IEEE Trans. Aut. Contr.*, 27:431–436, 1982.

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

[24] J.E. Griffin and P.J. Brown. Alternative prior distributions for variable selection with very many more variables than observations. Technical report, University of Warwick, Coventry, UK, 2005.

[25] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[26] T. Hastie, R. Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning*. Springer, July 2003.

[27] T. J. Hastie and R. J. Tibshirani. Generalized additive models. In *Monographs on Statistics and Applied Probability*, volume 43. Chapman and Hall, London, UK, 1990.

[28] N.J. Hsu, H.L. Hung, and Y.M Chang. Subset selection for vector autoregressive processes using lasso. *Computational Statistics and Data Analysis*, 52:36453657, 2008.

[29] J. Huang and T. Zhang. The benefit of group sparsity. Technical report, Rutgers University, 2009.

[30] E.D. Kolaczyk. *Statistical Analysis of Network Data*. Springer Series in Statistics. Springer, 2009.

[31] Hannes Leeb and Benedikt M. Ptscher. Sparse estimators and the oracle property, or the return of hodges' estimator. *Journal of Econometrics*, 142(1):201 – 211, 2008.

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

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

[34] D. Madigan and G. Ridgeway. [Least Angle Regression]: Discussion. *Annals of Statistics*, 32:465–469, 2004.

[35] D. Materassi and G. Innocenti. Topological identification in networks of dynamical systems. *Automatic Control, IEEE Transactions on*, 55(8):1860 –1871, aug. 2010.

[36] D. Materassi and M.V. Salapaka. On the problem of reconstructing an unknown topology. In *Proc. of IEEE American Control Conference (ACC), 2010*, pages 2113 –2118, jun. 2010.

[37] N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34:1436–1462, 2006.

[38] C. A. Micchelli and M. Pontil. Learning the kernel function via regularization. *Journal of Machine Learning Research*, 6:1099–1125, 2005.

[39] J. Mohammadpour and K.M. Grigoriadis. *Efficient Modeling and Control of Large-scale Systems*. Springer, 2010.

[40] D. Napoletani and T.D. Sauer. Reconstructing the topology of sparsely connected dynamical networks. *Phys. Rev. E*, 77(2):026103, Feb 2008.

[41] T. Park and G. Casella. The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, June 2008.

[42] G. Pillonetto, A. Chiuso, and G. De Nicolao. Prediction error identification of linear systems: a nonparametric Gaussian regression approach. *Automatica*, 2010. to appear.

[43] G. Pillonetto, A. Chiuso, and G. De Nicolao. Regularized estimation of sums of exponentials in spaces generated by stable spline kernels. In *Proceedings of the IEEE American Cont. Conf., Baltimora, USA*, 2010.

[44] G. Pillonetto, A. Chiuso, and G. De Nicolao. Prediction error identification of linear systems: a nonparametric Gaussian regression approach. *Automatica (in press)*, 2011.

[45] G. Pillonetto and G. De Nicolao. A new kernel-based approach for linear system identification. *Automatica*, 46(1):81–93, 2010.

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

[47] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464, 1978.

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

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

[50] J. Songsiri and L. Vandeberghe. Topology selection in graphical models of autoregressive processes. *Journal of Machine Learning Research*, 11:2671–2705, 2010.

[51] R. Tibshirani. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society, Series B.*, 58, 1996.

[52] M. Timme. Revealing network connectivity from response dynamics. *Phys. Rev. Lett.*, 98(22):224101, 2007.

[53] M. Tipping. Sparse bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1:211–244, 2001.

[54] P.M.J. van den Hof, D.K. de Vries, and P. Shoen. Delay structure conditions for identifiability of closed loop systems. *Automatica*, 28(5):1047–1050, 1992.

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

[56] H. Wang. Forward regression for ultra-high dimensional variable screening. *Journal of the American Statistical Association*, 104(488):1512–1524, 2009.

[57] H. Wang, G. Li, and C.L. Tsai. Regression coefficient and autoregressive order shrinkage and selection via the lasso. *Journal Of The Royal Statistical Society Series B*, 69(1):63–78, 2007.

[58] S. Weisberg. *Applied Linear Regression*. Wiley, New York.

[59] M. West. On scale mixtures of normal distributions. *Biometrika*, 74(3):646–648, September 1987.

[60] M. Yuan, A. Ekici, Z. Lu, and R. Monteiro. Dimension reduction and coefficient estimation in multivariate linear regression. *Journal of the Royal Statistical Society Series B*, 69(3):329–346, 2007.

[61] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, 68:49–67, 2006.

[62] T. Zhang. Adaptive forward-backward greedy algorithm for sparse learning with linear models. In *Proceedings of NIPS*, 2008.

[63] P. Zhao and B. Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2563, 2006.

[64] H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429, 2006.

[65] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320, 2005.