Identification of large-scale sparse linear dynamic systems: a regularization based approach

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

Identification of sparse high dimensional linear systems pose sever challenges to off-the-shelf techniques for system identification. This is particularly so when relatively small data sets, as compared to the number of inputs and outputs, have to be used.

While input/output 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 [24,38]. 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 [29,23].

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. [10].

In the dynamic case, i.e. when observed data are trajectories of (possibly stationary) stochastic processes, 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 [1]. In general both the dynamical systems and the interconnection structure is unknown and have to be inferred from data.

When the number of of measured variables is very large

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

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

[1] 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.

and possibly larger than the number of data available (i.e. the number of "samples" available for statistical inference), even though there is no "physical" underlying network, then 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 [40,30,26,27] and references therein. In the paper [40] coupled nonlinear oscillators (Kuramoto type) are considered where the coupling strengths are to be estimated; in [30] nonlinear dynamics are allowed and the attention is restricted to the linear term [2] 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. Instead, [26,27] consider linear models and the methodology is based on smoothing *a la* Wiener, where interconnections are found by putting a threshold on the estimated transfer functions.

In this paper we shall focus on modeling the relation between one node in this network (the "output" variable) and all the other variables measured (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 [20,12].

In the classical identification scenario, Prediction Error Methods (PEM) represent the most used approaches to optimal prediction of discrete-time systems [24]. 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 [1,36]. 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 [39] and Least Angle Regression (LAR) [13] 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 [50,51]. Extensions of this procedure for group selection include Group Lasso and Group LAR (GLAR) [49] where the sum of the Euclidean norms of each group (in place of the absolute value of the single components) is used. Theoretical analyses of these approaches and connections with the multiple kernel learning problem can be found in [5,28]. However, most of the work has been done in the "static" scenario while very little, with some exception [45,21], 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 [34] that relies on nonparametric estimation of impulse responses (see also [32] for extensions to predictor estimation). Rather than postulating finite-dimensional structures for the system transfer function, e.g. ARX, ARMAX or Laguerre [24], 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 [35], in [34] 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.

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 [49] and can be seen as a variation of the so-

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

called "elastic net" [52]; 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 [41]; see also the discussion on scale-mixture distributions in [17]. 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 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, indeed, SSGLAR provides a robust and computationally attractive way on 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: **TO BE DONE**

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. [16,15,8]) 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. [42]) in that under mild excitation conditions it guarantees identifiability.

In particular we define the sets of past measurements at time t

$$Y^t = [y_{t-1} \quad y_{t-2} \ldots], \qquad U^t = [u_{t-1} \quad u_{t-2} \ldots]$$

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

$$\begin{aligned} e_t &:= y_t - \hat{y}_{t|t-1} := y_t - \mathbb{E}[y_t|Y^t, U^t] \\ \mathbb{E}[y_t|Y^t, U^t] &:= \textstyle\sum_{j=1}^{m-1} \left[\sum_{k=1}^{\infty} h_k^{[j]} u_{t-k}^{[j]}\right] + \sum_{k=1}^{\infty} h_k^{[m]} y_{t-k}. \end{aligned} \tag{3}$$

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

[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.

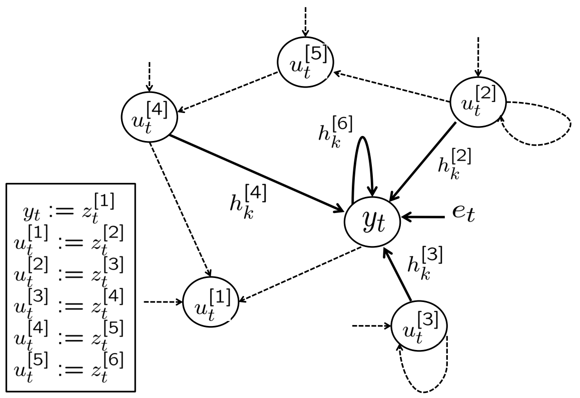

Fig. 1. A dynamical network representing the interaction between $m = 6$ variables. The solid edges represent the links related to the dynamical model for node $y_t := z_t^{[1]}$ given all the others. With reference to equation (3), absence of links from $u_t^{[i]} = z_t^{[i+1]}$, $i = 1,5$ to $y_t := z_t^{[1]}$ 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 $j \neq 1$.

estimation of the predictor impulse responses h_k in (3) from a finite set of input-output data $\{u_t, y_t\}_{t=1,..,N}$. We specifically address situations in which m is very large as compared to the number of available data N 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 [24,38] 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 [46,19] while we refer to the survey [18] for a more recent overview.

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 Kernel-based regularization

A widely used approach to reconstruct a function from indirect measurements $\{y_t\}$ consists of minimizing a regularization functional in a reproducing kernel Hilbert space (RKHS) $\mathscr{H}$ associated with a symmetric and positive-definite kernel K [3]. Given N data points, least-squares regularization in $\mathscr{H}$ estimates the unknown function as

$$\hat{h} = \arg\min_{h} \sum_{t=1}^{N} (y_t - \Gamma_t[h])^2 + \eta \|h\|_{\mathscr{H}}^2 \qquad (4)$$

where $\{\Gamma_t\}$ are linear and bounded functionals on $\mathscr{H}$ related to the measurement model while the positive scalar η trades off empirical error and solution smoothness [44]. Under the stated assumptions and according to the representer theorem [22], the minimizer of (4) is the sum of N basis functions defined by the kernel filtered by the operators $\{\Gamma_t\}$, with coefficients obtainable solving a linear system of equations. Such solution enjoys also an interpretation in Bayesian terms. It corresponds to the minimum variance estimate of h when h is a zero-mean Gaussian process with autocovariance K and $\{y_t - \Gamma_t[h]\}$ is white Gaussian noise independent of h [37]. Often, prior knowledge is limited to the fact that the signal, and possibly some of its derivatives, are continuous with bounded energy. In this case, f is often modeled as the p-fold integral of white noise. If the white noise has unit intensity, the autocorrelation of h is W_p where

$$W_p(s,t) = \int_0^1 G_p(s,u) G_p(t,u) du, \qquad (5)$$

$$G_p(r,u) = \frac{(r-u)_+^{p-1}}{(p-1)!}, \quad (u)_+ = \begin{cases} u & \text{if } u \geq 0 \\ 0 & \text{if } u < 0 \end{cases} \qquad (6)$$

This is the autocovariance associated with the Bayesian interpretation of p-th order smoothing splines [43]. In particular, when $p = 2$, one obtains the cubic spline kernel.

3.2 Kernels for system identification

In the system identification scenario, the main drawback of the kernel (5) is that it does not account for impulse response stability. In fact, the variance of h increases over time. This can be easily appreciated by looking at Fig. 2 (left) which displays 100 realizations drawn from a zero-mean Gaussian process with autocovariance proportional to W_2. One of the key contributions of [34] is the definition of a kernel specifically suited to linear system identification leading to an estimator with favorable bias and variance properties. In particular, it is easy to see that if the autocovariance of h is proportional to W_p, the variance of $h(t)$ is zero at $t = 0$ and tends to ∞ as t increases. However, if f represents a stable

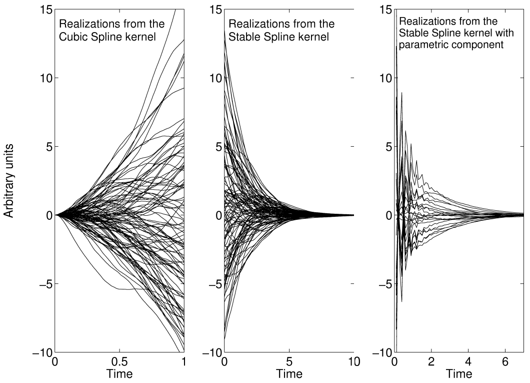

Fig. 2. Realizations of a stochastic process h with autocovariance proportional to the standard Cubic Spline kernel (left), the new Stable Spline kernel (middle) and its sampled version enriched by a parametric component defined by the poles $-0.5 \pm 0.6\sqrt{-1}$ (right).

impulse response, we would rather let it have a finite variance at $t = 0$ which goes exponentially to zero as t tends to ∞. This property can be ensured by considering autocovariances proportional to the class of kernels given 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 [34]. In practice, β will be unknown so that it is convenient to treat it as a hyperparameter to be estimated from data.
In view of (7), if $p = 2$ the autocovariance becomes the Stable Spline kernel introduced in [34]:

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

Proposition 1 *[34] 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.*

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

In practice we shall be working in discrete time and therefore we shall consider sampled versions of $h(t)$ (say h_k) so that $\{h_k\}_{k\in\mathbb{Z}^+}$ can be seen as realizations from the "sampled' Kernel $K_s(i,j)$ $i,j \in \mathbb{Z}^+$, so that $h \in \ell_1(\mathbb{Z}^+)$ almost surely.

3.3 Prior for predictor impulse responses

We shall model $\{h^{[k]}\}$ as independent Gaussian processes whose kernels share the same hyperparameters apart from the scale factors. In particular, each $h^{[k]}$ is proportional to the convolution of a zero-mean Gaussian process, with autocovariance given by the sampled version of K_2, with a parametric impulse response r, used to capture dynamics hardly represented by a smooth process, e.g. high-frequency oscillations. For instance, the zeta-transform $R(z)$ of r can be parametrized as follows

$$R(z) = \frac{z^2}{P_\theta(z)}, \qquad P_\theta(z) = z^2 + \theta_1 z + \theta_2, \qquad \theta \in \Theta \subset \mathbb{R}^2 \tag{9}$$

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. To better appreciate the role of the finite-dimensional component of the model, Fig. 2 (right panel) shows some realizations (with samples linearly interpolated) drawn from a discrete-time zero-mean normal process with autocovariance given by K_2 enriched by $\theta = [1 \quad 0.61]$ in (9). Notice that, in this way, an oscillatory behavior is introduced in the realizations by enriching the Stable Spline kernel with the poles $-0.5 \pm 0.6\sqrt{-1}$.
The kernel of h^k defined by K_2 and (9) is denoted by $K : \mathbb{N} \times \mathbb{N} \mapsto \mathbb{R}$ and depends on β, θ. Thus, letting $\mathbb{E}[\cdot]$ denote the expectation operator, the prior model on the impulse responses is given by

$$\mathbb{E}[h_j^k h_i^k] = \lambda_k^2 K(j,i;\theta,\beta), \quad k = 1,\ldots,m, \quad i,j \in \mathbb{N} \tag{10}$$

3.4 Sparsity inducing priors

Let us consider the problem of estimating the parameter $\theta \in \mathbb{R}^m$ in the linear model

$$Y = X\theta + W \tag{11}$$

where $Y \in R^N$ is the output vector data, $X \in \mathbb{R}^{N\times m}$ is the "regression vector" and $W \in \mathbb{R}^N$ is a noise term which we shall assume to be a zero mean vector with $\mathbb{E}[WW^\top] = \sigma^2 I$.

When the number m of regressors is very large (e.g. as compared to the number N 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. [39,6,41,19,13,14,7] and references therein.

A pioneering work in this direction has been the so called Lasso (Least Absolute Shrinkage and Selection Operator) [39] in which regressor selection has been performed by solving a problem of the form

$$\hat{\theta} := \arg\min_{\theta} \|Y - X\theta\|_2^2 \quad s.t.\ \|\theta\|_1 \le t \tag{12}$$

or, equivalently, the ℓ_1-penalized problem

$$\hat{\theta} := \arg\min_{\theta} \|Y - X\theta\|_2^2 + \gamma_1 \|\theta\|_1. \tag{13}$$

which in turn can also be seen as the Maximum a Posteriori (MAP) estimator in a Bayesian framework by assuming that W has a Gaussian distribution and θ a double exponential-type prior

$$p(\theta) \propto e^{-\lambda\|\theta\|_1}. \tag{14}$$

Despite its nice properties it has been argued (see [25]) that Lasso had not had a significant impact in statistical practice due to its relative computational inefficiency. The Least Angle Regression (LAR) algorithm [13] has provided a new approach to regressor selection and, with minor modifications (the "Lasso modification", [13]), also an efficient implementation of the Lasso.

Recently the Lasso has been proposed for estimation of regression models with autoregressive noise [45] and for Vector Autoregressive with eXogenous inputs (VARX) models [21]. This is a rather straightforward application once the regressor matrix X in (13) is formed with past inputs and outputs and θ 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 θ as independent Gaussian random variables $p(\theta_i|\sigma_i) = \mathcal{N}(\theta; 0, \sigma_i^2)$ where

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

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

$$p(\theta) = \prod_i \int p(\theta_i|\sigma_i)p(\sigma_i)d\sigma_i \tag{15}$$

which is a so-called "scale-mixture" distribution [2,47,31,17]. It is well known [2,47,31] that, if θ_i^2 has an exponential distribution itself, then $p(\theta)$ in (15) has the "double exponential" form (14). This is also related to the so called "Relevance Vector Machine" introduced in [41] which, however, uses a Gamma-type of prior on σ_i^{-2}.

In this paper we shall be concerned with a version of the problem (11) where each of the components θ_i of θ lives in fact in an infinite dimensional RKHS with kernel $K(s,t)$ as in (7) or, equivalently, θ_i is a Gaussian Process [35] with covariance $K(s,t)$. The regressor matrix X (which to be precise will be a linear operator whose representation has infinitely many columns) will contain the past histories of u and y. Details are found in the next section.

4 Variable selection as group sparsity

In this section we shall see how variable selection can be posed as a than of obtaining sparse solution of a linear problem similar to (11) 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 [49]; 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. [24,38]) each group of parameters would describe one impulse response. If we restrict our interest to ARX/FIR models this naturally yield to an algorithm for variable selection which we shall call "ARX-GLAR" (since we shall used the "group LAR" algorithm rather than the "group Lasso"). 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 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" live in an RKHS; this can be handled, as discussed in Section 3 via regularization/priors. This yields to formulations similar to multiple kernel learning [4].

In order to be more precise we need now to set up some notation: let us define

$$Y^+ := \begin{bmatrix} y_t \\ \vdots \\ y_{t+N-1} \end{bmatrix} \quad E^+ := \begin{bmatrix} e_t \\ \vdots \\ e_{t+N-1} \end{bmatrix} \quad h := \begin{bmatrix} h^{[1]} \\ \vdots \\ h^{[m]} \end{bmatrix}; \tag{16}$$

the predictor in (3) can be rewritten as:

$$Y^+ = \underbrace{\begin{bmatrix} A_1 & \dots & A_m \end{bmatrix}}_{:=A} h + E^+ \tag{17}$$

where

$$\begin{aligned} A_i^{[jk]} &:= u_{t-j-k}^{[i]}, \; i = 1,..,m-1, \quad j,k \in \mathbb{Z}^+ \\ A_m^{[jk]} &:= y_{t-j-k}, \; j,k \in \mathbb{Z}^+ \end{aligned}$$

Our identification problem can be stated as that of estimating h in (17),(16), subject to $h^{[i]} \in \ell_1(\mathbb{Z}^+)$, $i = 1,..,m$. Recall that we are interested in estimators which automatically selects, among $u^{[1]},..,u^{[m-1]},y$, the variables which are useful for predicting y and which are not. This is equivalent to saying that certain impulse responses $\hat{h}^{[i]}$ are expected to be exactly zero. Solving this problem entails estimation in "grouped" variables [49,48]; however a peculiarity here is that each "group" lives in an infinite dimensional space (or equivalently has infinitely many components) as in Multiple Kernel Learning [4].

The two approaches we consider in this paper are:

(i) SS-GLAR: a "group version" [49] of (12) extended to a non-parametric setup where the "groups" $h^{[i]}$ (see (17)) live in an infinite dimensional space; in order to include the penalty in the infinite dimensional space we have to solve a mixed $\ell_1 - \ell_2$ regularization problem which can be seen as a "group" version of the so-called "elastic-net" [52]. It is well known that the ℓ_2 penalty in the elastic net helps in selecting groups of correlated variables [52]. Note that this is different from the standard formulation of Multiple Kernel Learning [4]. 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 scale factors λ_i's have an exponential distribution, which will favor sparsity on the space of scale factors. As mentioned in [9] (see also [11]) this is also related to multiple kernel learning. This second approach will actually allow to introduce more flexibility in the Kernels enriching them with a parametric component as done in Section 3.3; as argued in [32] this may be advantageous in situations where the impulse responses contain "fast" dynamics which are penalized by the regularization term, see also [33]. Details will be given in Section 6.

5 Stable Splines Group LAR (SSGLAR) algorithm

6 Stable Splines with Exponential Hyperprior (SSEH) Algorithm

Acknowledgments

This research has been partially supported by the PRIN Projects New Methods and Algorithms for Identification and Adaptive Control of Technological Systems and Artificial Pancreas: Physiological Models, Control Algorithms and Clinical Test, by the Progetto di Ateneo CPDA090135/09 funded by the University of Padova and by the European 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] D.F. Andrews and C.L. Mallows. Scale mixtures of normal distributions. *Journal of the Royal Statistical Society*, 36:99–102, 1974.

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

[4] F. R. Bach. Consistency of the Group Lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9:1179–1225, 2008.

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

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

[9] 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.

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

[11] 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.

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

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

[14] 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.

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

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

[17] 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.

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

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

[20] 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.

[21] Nan-Jung Hsu, Hung-Lin Hung, and Ya-Mei Chang. Subset selection for vector autoregressive processes using lasso. *Computational Statistics and Data Analysis*, 52:36453657, 2008.

[22] G. Kimeldorf and G. Wahba. Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, 33(1):82–95, 1971.

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

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

[25] David Madigan and Greg Ridgeway. [Least Angle Regression]: Discussion. *Annals of Statistics*, 32:465–469, 2004.

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

[27] 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.

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

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

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

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

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

[33] 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.

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

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

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

[37] A. J. Smola and B. Schölkopf. Bayesian kernel methods. In S. Mendelson and A. J. Smola, editors, *Machine Learning, Proceedings of the Summer School, Australian National University*, pages 65–117, Berlin, Germany, 2003. Springer-Verlag.

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

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

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

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

[42] 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.

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

[44] G. Wahba. Support vector machines, reproducing kernel Hilbert spaces and randomized GACV. Technical Report 984, Department of Statistics, University of Wisconsin, 1998.

[45] 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.

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

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

[48] Ming Yuan, Ali Ekici, Zhaosong Lu, and Renato Monteiro. Dimension reduction and coefficient estimation in multivariate linear regression. *Journal of the Royal Statistical Society Series B*, 69(3):329–346, 2007.

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

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

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

[52] 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.