# Nonparametric sparse estimators for identification of large scale linear systems

Alessandro Chiuso and Gianluigi Pillonetto

***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 a new nonparametric technique which borrows ideas from a recently introduced Kernel estimator called "stable-spline" as well as from sparsity inducing priors which use $\ell_1$ penalty. We compare the new method with a group LAR-type of algorithm applied to estimation of sparse Vector Autoregressive models and to standard PEM methods.**

## I. INTRODUCTION

Several application domains, ranging from chemical engineering to economic systems, from computer vision to environmental modeling and monitoring are characterized by large amounts of measured variables. When it comes to estimating dynamic relationship among these variables one is faced with the problem of estimating dynamical systems with large numbers of inputs and outputs. We call these systems "large scale systems". Even restricting to the linear world, these systems pose severe challenging to state of the art tools for linear system identification. On the one side handling large scale systems may render parametric (nonlinear) optimization based methods like Prediction Error Minimization (PEM) methods [1], [2] computationally challenging due to

(i) non-convexity
(ii) parametrization issues
(iii) large numbers of parameters to be estimated as compared to the number of data available.

Remaining within the optimization based methods, the parametrization issue may be circumvented by the so-called Data Driven Local Coordinates (DDLC) [3]. Alternatively the so-called subspace methods [4], [5], [6] have been developed; these are numerically stable procedures which are not based on iterative non-linear optimization and rather deliver the estimates in two steps, relying in robust tools in numerical linear algebra like QR and SVD decomposition. The reader is referred to [7], [8] for a survey on recent results. These methods overcome limitations (i) and (ii) mentioned above. However item (iii) is still an issue. In this paper we are concerned with problems in which the number of inputs and outputs may be large as compared to the number of data and the system is structured in that, e.g., only subset of past inputs and outputs is needed to predict a certain output channel. In such situations both parametric estimation methods as well as subspace methods may run into troubles. The statistics and machine learning literatures have addressed these issues quite extensively; however, to the best of the authors knowledge most work has been done is the "static" scenario while very little (with some exception [9],[10], [11]) can be found regarding estimation of dynamic systems.

Parametric estimation techniques which result in sparse models have nowadays become very popular and include, among others, the well known Lasso [12], Least Angle Regression (LAR) [13] and their "grouped" versions Group Lasso and Group LAR (GLAR) [14].

In this paper we introduce a new nonparametric technique which borrows ideas from a recently introduced Kernel estimator called "stable-spline" as well as from sparsity inducing priors which use $\ell_1$ penalty. We compare the new method with a group LAR-type of algorithm applied to estimation of sparse Vector Autoregressive models and to standard PEM methods.

The structure of the paper is as follows: Section II states the problem and set up notation. In Section III we briefly recall LASSO, LAR and GLAR and discuss how the GLAR algorithm can be utilized to perform input selection in Vector Autoregressive with exogenous inputs (VARX) models. The non parametric estimator is introduced in Section IV and simulation results are presented in Section V. Section VI contains conclusions and directions for future work.

## II. STATEMENT OF THE PROBLEM AND NOTATION

Let $\{y_t\}_{t\in\mathbb{Z}}$, $y_t \in \mathbb{R}^p$ and $\{u_t\}_{t\in\mathbb{Z}}$, $u_t \in \mathbb{R}^m$ be a pair of jointly stationary stochastic processes which are, respectively, the output and input of an unknown time-invariant dynamical system. With some abuse of notation the symbol $y_t$ will both denote a random variable (from the random process $\{y_t\}_{t\in\mathbb{Z}}$) and its sample value. In particular we define the sets of past measurements at time $t$

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

This research has been partially supported by the PRIN Project "Sviluppo di nuovi metodi e algoritmi per l'identificazione, la stima bayesiana e il controllo adattativo e distribuito" and by the Progetto di Ateneo CPDA090135/09 funded by the University of Padova.

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

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

The symbols $\mathbb{E}[\cdot]$ and $\mathbb{E}[\cdot|\cdot]$ denote, respectively, expectation and conditional expectation while $\hat{E}[\cdot|\cdot]$ 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[1] $A^i$ will be used in place of $A^{i1}$ or $A^{1i}$. The symbol $I$ denotes the identity matrix of suitable dimensions and $A^\top$ is the transpose of the matrix $A$. The symbol $\|x\|_p$ denotes the $p-$norm of the vector $x$.

Our purpose is to identify a linear dynamical system[2] of the form

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

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

$$\begin{array}{rcl} e_t & := & y_t - \hat{E}[y_t|Y^t, U^t] \\ & = & y_t - \hat{y}_{t|t-1} \\ & = & y_t - \sum_{k=1}^{\infty} h_k u_{t-k} - \sum_{k=1}^{\infty} q_k y_{t-k}. \end{array} \tag{2}$$

The matrix sequences $h_k \in \mathbb{R}^{p\times m}$ and $q_k \in \mathbb{R}^{p\times p}$, $k \in \mathbb{Z}^+$ are the predictor impulse response coefficients.

Following the Prediction Error Minimization framework, in this paper we shall convert identification of the dynamical system in (1) in estimation of the predictor impulse responses $h_k$ and $q_k$ in (2) from a finite set of input-output data $\{u_t, y_t\}_{t=1,..,N}$. We shall consider large scale systems in which the numbers $p$ and/or $m$ are very large as compared to the number of available data $N$; in addition we shall also assume that only few inputs and/or outputs are needed to predict the $i-th$ component of $y_t$. Mathematically this can be formulated as follows: consider prediction of the $i-th$ component of $y_t$. Assume the $j-th$ component of $y$ and $\ell-th$ component of $u$ are not needed to predict $y_t$. This means that $h_k^{i\ell} = q_k^{ij} = 0$, $\forall k \in \mathbb{Z}^+$. For simplicity of exposition we shall restrict the attention to MISO systems (i.e. $p = 1$); however the methodologies described in this paper are by no means limited to this situation.

The problem we consider from now on is, therefore, that of estimating $q_k$ and $h_k^i$ in

$$\hat{y}_{t|t-1} = \sum_{i=1}^{m} \left[ \sum_{k=1}^{\infty} h_k^i u_{t-k}^i \right] + \sum_{k=1}^{\infty} q_k y_{t-k} \tag{3}$$

In practice one does not know whether a measured input and/or output is significant for prediction of $y_t$. Standard PEM methods [1], [2] do not attempt to perform input selection and estimate a "full" model which use 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 [15], [16] while we refer to the survey [17] for a more recent overview.

In this paper we shall adapt recent work on estimation of sparse models [13] to identification of linear predictors.

## III. SPARSITY INDUCING PRIORS FOR ESTIMATION OF AUTOREGRESSIVE MODELS

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

$$Y = X\theta + W \tag{4}$$

where $Y \in R^N$ is the output vector data, $X \in \mathbb{R}^{N\times n}$ 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 $n$ 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. [12], [18], [16], [13], [19], [20] and references therein.

A pioneering work in this direction has been the so called Lasso (Least Absolute Shrinkage and Selection Operator) [12] in which regressor selection has been performed by solving a problem of the form

$$\hat{\theta} := \arg\min_{\theta} \|Y - X\theta\|_2^2 \quad s.t. \; \|\theta\|_1 \leq t \tag{5}$$

Equivalently, (5) can be formulated as an $\ell_1$-penalized regularization problem of the form

$$\hat{\theta} := \arg\min_{\theta} \|Y - X\theta\|_2^2 + \gamma_1 \|\theta\|_1 \tag{6}$$

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 $\theta$ a double exponential-type prior

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

Despite its nice properties it has been argued (see [21]) 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 (which shall be called "Lasso modification", see [13] for details), also an efficient implementation of the Lasso.

### A. Input selection in VARX models

Recently the Lasso, possibly implemented via the LAR algorithm, has been proposed for estimation of regression models with autoregressive noise [9] and for Vector Autoregressive with eXogenous inputs (VARX) models [10]. This is a rather straightforward application once the regressor matrix $X$ in (5) is formed with past inputs and outputs. In fact consider (3) and assume the predictors have finite memory $M$ (i.e. we restrict to $VARX(M)$ models). Then

$$\begin{array}{rcl} y_t & = & \sum_{i=1}^{m} \left[ \sum_{k=1}^{M} h_k^i u_{t-k}^i \right] + \sum_{k=1}^{M} q_k y_{t-k} + e_t \\ & = & X_t \theta + e_t \end{array} \tag{7}$$

[1] There is no risk of confusion with matrix powers since we shall *never* use matrix powers in this paper.

[2] In order to streamline notation we shall assume one delay from $u_t$ to $y_t$.

where

$$
\begin{array}{rcl}
X_t & := & [u^1_{t-1},..,u^1_{t-M},...,u^m_{t-1},..,u^m_{t-M},y_{t-1},...,y_{t-M}] \\
\theta & := & [h^1_1,..,h^1_M,...,h^m_1,..,h^m_M,q_1,...,q_M]^\top .
\end{array}
\tag{8}
$$

Taking into account that data $y_t, u_t$ are available for $t = 1,..,N$, (7) can be written compactly as

$$
\begin{bmatrix} y_N \\ y_{N-1} \\ \vdots \\ y_{M+1} \end{bmatrix} = \begin{bmatrix} X_N \\ X_{N-1} \\ \vdots \\ X_{M+1} \end{bmatrix} \theta + \begin{bmatrix} e_N \\ e_{N-1} \\ \vdots \\ e_{M+1} \end{bmatrix}
$$

which is of the form (4).

However, these approaches do not enforce any structure in the estimated zero pattern of $\hat{\theta}$. For instance, with reference to (7) and (8), in order to construct a sparse predictor which uses only a subset of the available inputs $u^i_t$, $i = 1,..,m$ and output $y_t$, one has to guarantee that groups of components in $\hat{\theta}$ are zeros. With reference to (8), if e.g the first $M$ entries of $\hat{\theta}$ are zero, i.e. $\hat{h}^1_1 = ... = \hat{h}^1_M = 0$, then the first input ($u^1_{t-k}$ $\forall k > 0$) does not enter in the prediction of $y_t$.

This entails estimation in "grouped" variables and has been addressed in [14], [22]. Basically one needs to extend the Lasso/LAR ideas to guarantee that selection is performed among groups of variables and no longer on single variables. Such extensions are called, respectively, *Group Lasso* and *Group LAR* in [14].

While the LAR algorithm can be modified to obtain the solution to the Lasso problem (5), this is no longer true for their group versions as discussed in [14]. Unfortunately the LAR algorithm (without the "Lasso modification") does not have an immediate interpretation as a regularized optimization problem of the form (6). However due to its computational simplicity we prefer to work with the Group LAR algorithm which we shall describe in the next section. Later in the paper we shall extend this algorithm to perform sparse estimation in Reproducing Kernel Hilbert Spaces, which will be the core of our contribution.

### B. Group Least Angle Regression (GLAR)

Consider the problem

$$
Y = \sum_{i=1}^{K} X_{(i)} \theta_{(i)} + W \tag{9}
$$

where $Y \in R^N$ is the output vector data, $X_{(i)} \in \mathbb{R}^{N \times n_i}$, $i = 1,..,K$, are the groups of regression vectors, $\theta_{(i)} \in \mathbb{R}^{n_i \times 1}$ is the $i-th$ group of variables and $W \in \mathbb{R}^N$ is a noise term which we shall assume to be a zero mean vector with $\mathbb{E}[WW^\top] = \sigma^2 I$.

The Group LAR (GLAR) algorithm proceeds as follows [14]:

**GLAR Algorithm**[3]

```
(a) set $\hat{Y}_{(0)} = 0$, $\mathscr{X}_{sel} = \emptyset$
(b) for $i = 0 : K-1$ do:
    (1) Among the groups that have not been already selected (i.e. $X_{(i)} \notin \mathscr{X}_{sel}$) find the group $X_{(j)}$ that has the smallest canonical angle with $Y - \hat{Y}_{(i)}$ and add it to the set $\mathscr{X}_{sel}$ of selected groups;
    (2) let $E_{(i)}$ be the projection of $Y - \hat{Y}_{(i)}$ onto the space spanned all the groups in the set $\mathscr{X}_{sel}$;
    (3) move along the direction $E_{(i)}$, i.e. set $\hat{Y}_{(i+1)} = \hat{Y}_{(i)} + \alpha E_{(i)}$, $\alpha > 0$ until $Y - \hat{Y}_{(i+1)}$ has as large correlation with some other group $X_{(\bar{j})} \notin \mathscr{X}_{sel}$ as with $X_{(j)}$. Let $\bar{\alpha}$ be the value of $\alpha$ which satisfies this condition
    (4) set $\hat{Y}_{(i+1)} = \hat{Y}_{(i)} + \bar{\alpha} E_{(i)}$
    (5) go back to (b).
```

The predictors $\hat{Y}_{(i)}$, $i = 1,..,K$, have the form

$$
\hat{Y}_{(i)} = \sum_{j=1}^{K} X_{(j)} \hat{\theta}_{(j)}
$$

where only $i$ of the groups $\hat{\theta}_j$, $j = 1,..,K$, have some nonzero entry. Choosing among the predictors $\hat{Y}_{(i)}$, $i = 1,..,K$ can be done, for instance, using the $C_p$ statistic as suggested in [14]. We shall use instead a validation based approach described in Section V; this approach seems to be more robust that $C_p$.

It is worth recalling that the whole GLAR algorithm producing all estimators $\hat{Y}^{(i)}$ has a complexity which is of the same order as that of solving a standard least squares problem for the model (9).

This algorithm can be directly employed to estimate a VARX model as in (7) once the order $M$ has been specified. However both the number of nonzero components as well as $M$ have to be estimated from data. This could be done using order estimation techniques (AIC,BIC) for $M$ (see e.g. [1], [2]) together with $C_p$-type statistics for estimation of the number of nonzero components (see [14]). For reasons which will become clear later on, we shall adopt a validation based approach to select both $M$ and the number of nonzero groups, see Section V for details. We shall call the resulting algorithm *VARX-GLAR*.

### C. GLAR with $\ell_2$ penalty

Consider now a regularized problem similar to (6) where also an $\ell_2$ penalty on the coefficients is considered[4]

$$
\hat{\theta} := \arg\min_{\theta} \|Y - X\theta\|_2^2 + \gamma_1 \|\theta\|_1 + \gamma_2^2 \|\theta\|_2^2 \tag{10}
$$

*Remark 1:* A mixed $\ell_1$-$\ell_2$ optimization problem of the form (10) has been called in the literature "elastic net" (see [23]). Discussion regarding drawbacks of this formulation for simultaneous selection (Lasso, $\ell_1$ penalty) and shrinkage (ridge-regression, $\ell_2$ penalty), as well as for possible remedies can be found in [23]. We shall not pursue this avenue here.

Problem (10) can be equivalently formulated in the form

$$
\hat{\theta} := \arg\min_{\theta} \|\bar{Y} - \bar{X}\theta\|_2^2 + \gamma_1 \|\theta\|_1 \tag{11}
$$

[3] This is actually a simplified version. See [14] for more details.

[4] The reason for introducing the $\ell_2$ penalty will become clear in Section IV.

where

$$\bar{Y} := \begin{bmatrix} Y \\ 0_{n\times 1} \end{bmatrix} \qquad \bar{X} := \begin{bmatrix} X \\ \gamma_2 I_{n\times n} \end{bmatrix}. \tag{12}$$

As discussed in [13], problem (11) can be efficiently solved via the LAR algorithm. Since, as mentioned in the previous section, we are interested in the "grouped" version, we shall apply the GLAR algorithm to the regression problem

$$\bar{Y} = \sum_{i=1}^{K} \bar{X}_{(i)}\theta_{(i)} + \bar{W} \tag{13}$$

where $\bar{Y}$ is as in (12), $\bar{X}_{(i)} \in \mathbb{R}^{(N+n)\times n_i}$ is the $i-th$ group of columns of $\bar{X}$ in (12) and

$$\bar{W} := \begin{bmatrix} W \\ W_\theta \end{bmatrix} \qquad \bar{W} \sim \mathcal{N}\left(0, \sigma^2 I\right) \tag{14}$$

## IV. STABLE SPLINE KERNEL

The VARX-GLAR algorithm introduced in Section III has two important limitations: (i) the VARX order M has to be either known or estimated from data and (ii) only finite-memory predictors are described within this class. As a consequence of this second item, in order to describe predictors with very long (possibly infinite) memory, a large number of parameters is needed. This may become troublesome in particular when only small data set are available. To overcome these limitations, we consider an alternative approach, initiated in [24], [25], in which we assume $\{q_k\}$ and $\{h_k^i\}$ are realizations from a Gaussian processes, mutually independent and independent of $\{e_t\}$. Their auto-covariances (kernels) are denoted by

$$\begin{aligned} cov(q_t, q_\tau) &= \gamma_2^2 K_\beta(t,\tau), \quad t,\tau \in \mathbb{N} \qquad (15)\\ cov(h_t^i, h_\tau^j) &= \gamma_2^2 K_\beta(t,\tau), \quad t,\tau \in \mathbb{N} \qquad (16)\end{aligned}$$

where $\gamma_2$ in an unknown positive scale factor and $\beta$ is an asymptotic exponential decay-rate parameter to be determined from data.

Following [24], we introduce the so called *Stable Spline* Kernel $K_\beta(t,\tau)$ as a time-warped version of the autocovariance of the integrated Wiener process; this guarantees that realizations from (15) are almost surely BIBO stable sequences.

To define the kernels, recall that, assuming zero initial condition at time zero, the autocovariance of the integrated Wiener process is (see e.g. [26])

$$W(s,\tau) = \frac{s\tau \min\{s,\tau\}}{2} - \frac{(\min\{s,\tau\})^3}{6}, \quad (s,t) \in \mathbb{R}^+ \times \mathbb{R}^+$$

Following [24], stability can be included by means of an exponential time-transformation which leads to definition of the so-called stable spline kernel

$$K_\beta(t,\tau) = W(e^{-\beta t}, e^{-\beta\tau}) \qquad (t,\tau) \in \mathbb{R}^+ \times \mathbb{R}^+ \tag{17}$$

Note that there is a duality between zero-mean Gaussian processes with autocovariance $K$ and the RKHS $\mathscr{H}_K$, e.g. see Section 1.4 in [27]; hence we shall interchangeably use the terms autocovariance an Kernel in the sequel.

An in-depth treatment on RKHS can be found in the seminal work [28]. Here, we recall that given a symmetric and positive-definite kernel $K$ defined on a metric space $X$, the associated RKHS $\mathscr{H}_K$ is the Hilbert space of functions on $X$ which are the completion of the manifolds given by all the finite linear combinations

$$\sum_{i=1}^{l} m_i K(\cdot, t_i) \tag{18}$$

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

$$< \sum_i m_i K(\cdot,t_i), \sum_j n_j K(\cdot,s_j) >_{\mathscr{H}_K} = \sum_{i,j} m_i n_j K(t_i,s_j) \tag{19}$$

Since $< f(\cdot), K(\cdot,x) >_{\mathscr{H}_K} = f(x)$, $K$ is also named reproducing kernel of $\mathscr{H}_K$.

In our context, it will be useful to introduce a finite dimensional subspace of $\mathscr{H}_K$ as follows: let $\mathbb{L}^2(\mathbb{R}^+)$ indicate the space of square integrable functions on $\mathbb{R}^+$ with respect to the Lebesgue measure. Recall that, under suitable technical conditions, Mercer theorem on noncompact domains [29] states that $K$ admits an eigenfunctions-eigenvalue decomposition with eigenfunctions $\{\rho_j\}$ and corresponding eigenvalues $\{\lambda_j\}$. The functions $\{\rho_j\}$ are orthogonal in $\mathbb{L}_\beta(\mathbb{R}^+)$, the space of square integrable functions on $\mathbb{R}^+$ with norm weighted by the density $\beta e^{-\beta t}$. If $\lambda_1 \geq \lambda_2 \geq \ldots > 0$ [30], $\mathscr{H}_K$ is defined by

$$\mathscr{H}_K = \left\{ f \in \mathbb{L}^2 \quad | \quad f(s) = \sum_{j=1}^{\infty} a^j \rho_j(s), \quad \sum_{j=1}^{\infty} \frac{(a^j)^2}{\lambda_j} < \infty \right\}$$

Then, for $J \in \mathbb{N}$, we define

$$\mathscr{H}_K^J = \text{span}\{\rho_1, \ldots, \rho_J\} \tag{20}$$

so that, if $h \in \mathscr{H}_K^J$, for a suitable $a \in \mathbb{R}^J$ it holds that

$$h(x) = \sum_{j=1}^{J} a^j \rho_j(x), \quad \|h\|_{\mathscr{H}}^2 = \sum_{j=1}^{J} \frac{(a^j)^2}{\lambda_j}$$

The following proposition, taken from [24], is especially useful for numerical purposes. In fact, it provides the eigenfunctions of the stable spline kernel (17) in closed form.

*Proposition 2:* The eigenvalues $\{\lambda_i\}$ of the stable spline kernel (17) are defined by

$$\lambda_i = (1/\alpha_i)^4 \quad i = 1,2,\ldots \tag{21}$$

where $\alpha_i$ denotes the solution of

$$1/\cosh(\alpha) + \cos(\alpha) = 0 \tag{22}$$

which is closest to $(i-1/2)\pi$.
In addition, the associated eigenfunctions $\{\rho_i\}$ are

$$\rho_i(\tau,\beta;\alpha_i) = \phi_i(e^{-\beta\tau};\alpha_i) \quad \tau \in X \tag{23}$$

where

$$\begin{aligned}\phi_i(t;\alpha_i) &= C_1(\alpha_i)\cos(\alpha_i t) + C_2(\alpha_i)\sin(\alpha_i t) \\ &+ C_3(\alpha_i)e^{-\alpha_i(1-t)} + C_4(\alpha_i)e^{-\alpha_i t} \qquad t \in [0,1]\end{aligned}$$

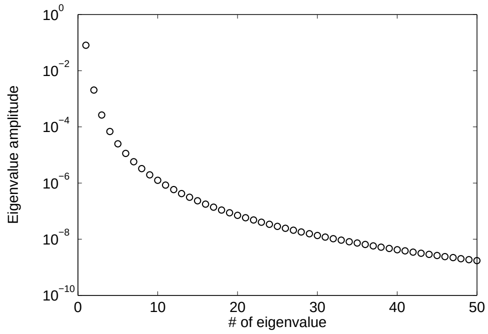

Fig. 1. Eigenvalues $\{\lambda_i\}_{i=1}^{50}$ of the stable spline kernel.

and $\{C_k\}$ are scalars satisfying

$$\begin{array}{rcl} C_4(\alpha) & = & (\int_S [C_1(1)\cos(\alpha t) + C_2(1)\sin(\alpha t) \\ & + & C_3(1)e^{-\alpha(1-t)} + e^{-\alpha t}]^2 dt)^{-1/2} \\ C_3(C_4) & = & C_4(\alpha)\left[\dfrac{2}{1+e^{-2\alpha}} - 1\right]/\sin(\alpha) \\ C_2(C_4) & = & C_4(\alpha) - C_3(C_4)e^{-\alpha} \\ C_1(C_4) & = & -C_4(\alpha) - C_3(C_4)e^{-\alpha} \end{array}$$

In Fig. 1 the first 50 eigenvalues of the stable spline kernel are displayed.

### A. GLAR in Reproducing Kernel Hilbert Spaces

Let us assume now that the impulse responses $h^i$, $q$ are (sampled versions of) functions in $\mathscr{H}_K$. The problem of estimating the impulse responses $h^i$, $q$ from measured data can be formulated as the following Tikhonov-type regularization problem:

$$\{\hat{h}^i, \hat{q}\} = \underset{h^i, q \in \mathscr{H}_K}{\arg\min} \sum_{t=t_0}^{N} (y_t - \hat{y}_{t|t-1})^2 + \gamma_2^2 \left( \|q\|^2_{\mathscr{H}_K} + \sum_{i=1}^{m} \|h^i\|^2_{\mathscr{H}_K} \right) \tag{24}$$

subject to

$$\hat{y}_{t|t-1} = \sum_{i=1}^{m} \left[ \sum_{k=1}^{\infty} h_k^i u_{t-k}^i \right] + \sum_{k=1}^{\infty} q_k y_{t-k}$$

The parameter $\gamma_2$ is the so called regularization parameter which has to trade fit $y_t - \hat{y}_{t|t-1}$ vs. regularity of $q$ and $h^i$.

In order to reframe this problem in a finite dimensional setup, we assume from now on that $h^i, q \in \mathscr{H}_K^J$, i.e. that for suitable $a_i \in \mathbb{R}^J$ and $b \in \mathbb{R}^J$, it holds that

$$\begin{array}{ll} h_k^i = \sum_{j=1}^{J} a_i^j \rho_j(k), & \|h\|^2_{\mathscr{H}} = \sum_{j=1}^{J} \frac{(a_i^j)^2}{\lambda_j} \\ q_k = \sum_{j=1}^{J} b^j \rho_j(k), & \|q\|^2_{\mathscr{H}} = \sum_{j=1}^{J} \frac{(b^j)^2}{\lambda_j}; \end{array}$$

the number $J$ does not have to trade bias vs. variance but is just related to computational issues. Define the filtered past input and output data as follows[5]:

$$\begin{array}{rcl} \phi_t^{ij} & = & \sum_{k=1}^{\infty} u_{t-k}^i \rho_j(k) \\ \psi_t^j & = & \sum_{k=1}^{\infty} y_{t-k} \rho_j(k) \end{array} \tag{25}$$

Using (25) equation (3) can be rewritten in the form

$$\hat{y}_{t|t-1} = \sum_{i=1}^{m} \left[ \sum_{j=1}^{J} a_i^j \phi_t^{ij} \right] + \sum_{j=1}^{J} b^j \psi_t^j. \tag{26}$$

Hence, under the restriction $h^i, q \in \mathscr{H}_K^J$, the solution to problem (24) can be rewritten as

$$\hat{h}_k^i = \sum_{j=1}^{J} \hat{a}_i^j \rho_j(k), \quad \hat{q}_k = \sum_{j=1}^{J} \hat{b}^j \rho_j(k) \tag{27}$$

where $\{\hat{a}_i, \hat{b}\}$ solve the problem[6]

$$\underset{a_i, b \in \mathbb{R}^J}{\arg\min} \sum_{t=t_0}^{N} (y_t - \hat{y}_{t|t-1})^2 + \gamma_2^2 \sum_{j=1}^{J} \left( \frac{(b^j)^2}{\lambda_j} + \sum_{i=1}^{m} \frac{(a_i^j)^2}{\lambda_j} \right) \tag{28}$$

subject to (26).

Problem (28) is an $\ell_2$-penalized linear regression problem which can be rewritten in the form

$$\bar{Y} = \sum_{i=1}^{m+1} \bar{X}_{(i)} \theta_{(i)} + W \tag{29}$$

provided we define $\bar{Y} := [y_N^\top, y_{N-1}^\top, \ldots, y_{t_0}^\top, 0_{1\times(J(m+1))}]^\top$,

$$\begin{array}{rcl} \theta_{(i)} & := & a_i \quad i = 1, .., m \\ \theta_{(m+1)} & = & b \end{array} \tag{30}$$

$$\begin{array}{rcl} X & := & [X_{(1)}\; X_{(2)}\; \ldots\; X_{(m+1)}] \\ X_{(i)} & := & \begin{bmatrix} \phi_N^{i1} & \ldots & \phi_N^{iJ} \\ \vdots & \ldots & \vdots \\ \phi_{t_0}^{i1} & \ldots & \phi_{t_0}^{iJ} \end{bmatrix} \quad i = 1, \ldots, m \\ X_{(m+1)} & := & \begin{bmatrix} \psi_N^{1} & \ldots & \psi_N^{J} \\ \vdots & \ldots & \vdots \\ \psi_{t_0}^{1} & \ldots & \psi_{t_0}^{J} \end{bmatrix} \end{array} \tag{31}$$

$$\begin{array}{rcl} \bar{X} & := & [X^\top\; \bar{\Lambda}]^\top \quad \bar{\Lambda} := I_{m+1} \otimes \Lambda \\ \Lambda & := & \mathrm{diag}\{\gamma_2/\sqrt{\lambda_1}, \ldots, \gamma_2/\sqrt{\lambda_J}\} \end{array} \tag{32}$$

and $\bar{X}_{(i)}$ is the $i-th$ block column of $\bar{X}$ (corresponding to the partition of $X$ in (31)).

Performing input selection can be tackled, as discussed in Section III, via the Group Least Angle Regression algorithm in subsection III-C applied to the regression problem (29). 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)**

[5] These infinite sums will have to be truncated in practice since only a finite amount of data is available. This is not a critical issue since the eigenfunctions $\rho_j$ decay exponentially to zero.

[6] Note that, since data are available only in the interval $[1, N]$, the predictor $\hat{y}_{t_0|t_0-1}$ can only use data in the time interval $[1, t_0 - 1]$. The initial time $t_0$ is chosen so that it is comparable to the "practical" length of the predictor impulse responses.

1) fix the parameter $\beta$ in (17), compute the eigenfunctions $\rho_j$ in (23), the regressors $\phi_t^{ij}$ and $\psi_t^j$ in (25);
2) fix the parameter $\gamma_2$ in (24) and (32); form the regressor $\bar{X}_{(i)}$ in (29) as described in formulas (31),(32);
3) estimate $\theta_{(i)}$ applying the LAR algorithm to problem (29);
4) estimate $\hat{h}^i$ and $q$ as in (27) where $\hat{a}_i$ and $\hat{b}$ are found from $\hat{\theta}$ according to (30).

Note that, in order to do so, the following parameters have to be chosen:

(a) the $\ell_2$ penalty $\gamma_2$ in (32) (regularity of $h_k^i, q_k$ in the space $\mathscr{H}_K$)
(b) the parameter $\beta$ in (17) (decay rate of the eigenfunctions $\rho_j$)
(c) the number of non-zero blocks estimated via the GLAR algorithm.

We shall describe in Section V a validation based approach to estimate these "hyperparameters".

## V. SIMULATION RESULTS

In order to validate the proposed approaches we compare, in two simulation studies, (i) SS-GLAR, (ii) VARX-GLAR and (iii) PEM from the Matlab toolbox.

Before describing the experimental setup and the results, we discuss the choice of the hyperparameters used in SS-GLAR, order estimation for PEM and VARX-GLAR and the choice of the sparsity in SS-GLAR and VARX-GLAR.

### A. *Choice of hyperparameters, sparsity and order estimation*

In this paper we propose a very simple validation based approach as follows. Let $\{y_t, u_t\}_{t=1,..,N}$ be the available data; in this paper $N = 500$. 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 for fixed hyperparameters on the identification data set and we validate the identified model on the validation data set. We grid the hyperparameter space ($\beta \in \mathbb{R}^+$, $\gamma_2 \in \mathbb{R}^+$) so that only a finite (and possibly small) number of alternatives is tested[7]. Also different level of sparsity (i.e. different number of non-zero groups $i = 1,..,m+1$) are tested. We finally select the identified model which performs best (as measured by the root-mean-squared error in one-step-ahead prediction error (33)) on validation data.

We follow a similar approach for choosing the autoregressive order as well as the level of sparsity for the VARX-GLAR algorithm.

The PEM method is applied to state space models of order $n$; the order is either fixed (Example 1) or selected using the validation approach described above selecting $n \in [1:20]$ (Example 2).

### B. *Performance Evaluation*

The predictive performance of all identified models are then tested on new data, called *test set* composed of 500 data points generated using the same data generating mechanism used for identification (but of course with different realizations of the inputs and the innovation process).

For simplicity of exposition let us define $\delta_{q0}$ to be 1 if the true impulse response $q = 0$ and zero otherwise and let the $I_0$ be the subset of $\{1,..,m\}$ of indexes corresponding to $h^i = 0$, $i \in I_0$.

The following performance indexes are considered:

1) k-step-ahead root mean squared error

$$RMS_k := \sqrt{\frac{1}{500}\sum_{i=1}^{500}(y_t - \hat{y}_{t|t-k})^2} \tag{33}$$

2) k-step-ahead root Coefficient of Determination (COD)

$$COD_k := 1 - \frac{RMS_k^2}{\frac{1}{500}\sum_{i=1}^{500}(y_t - \bar{y})^2} \tag{34}$$

where $\bar{y}$ is the sample mean of $y_t$.

3) Norm of estimated impulse responses corresponding to inputs/outputs not present in the model generating the data.

$$Err_0 := \sum_{i \in I_0} \|\hat{h}^i\|_2 + \delta_{q0}\|\hat{q}\|_2 \tag{35}$$

4) Normalized error in estimating impulse responses corresponding to inputs/outputs present in the model generating the data

$$Err_1 := \sum_{i \notin I_0} \frac{\|h^i - \hat{h}^i\|_2}{\|h\|_2} + (1 - \delta_{q0})\frac{\|q - \hat{q}\|_2}{\|q\|_2} \tag{36}$$

We perform $N_{MC} = 125$ Monte Carlo runs and, for each run $= 1,..,N_{MC}$, we compute the indexes above which are denoted by $RMS_k^j$, $COD_k^j$ $Err_0^j$, $Err_1^j$. Average values over the Monte Carlo study of these quantities will be denoted with a bar on top, e.g. $\overline{RMS}_k := \frac{1}{N_{MC}}\sum_{j=1}^{N_{MC}} RMS_k^j$.

### C. *Example 1*

We consider a MISO linear system with 15 inputs of the form

$$y_t = \sum_{i=1}^{15} F(z)\delta_i u_t^i + G(z)e_t \tag{37}$$

where

$$F(z) = \frac{z^2 - 0.81}{z(z^2 - 1.6z + 0.89)} \quad G(z) = \frac{z^2 - 0.8z + 0.97}{z^2 - 1.6z + 0.89}$$

The inputs are independent white Gaussian noises, zero mean and unit variance. The innovation process $\{e_t\}$ is also a zero mean, unit variance Gaussian process. The variables $\delta_i$, $i = 1,..,15$ are either zero or one and determine which inputs actually affect $y(t)$. We perform $N_{MC} = 125$ Monte Carlo experiments and, for each run, $\delta_i$ are chosen as realizations of independent Bernoulli random variables with parameter $p = 0.3$, i.e. $P[\delta_i = 1] = p$ and $P[\delta_i = 0]$.

We compare the following estimators:

[7] Experimental evidence shows that the results are not very sensitive to choice of hyperparameters, so that a rather rough grid suffices.

1) SS-GLAR described in Section IV; the parameters $\beta, \gamma_2$ and the level of sparsity (i.e. the number of nonzero impulse responses) is determined using the validation based approach described above.
2) VARX-GLAR described in Section III. The order of the VARX models is constrained to be between 1 and 30; both this order and the number of nonzero components have been estimated using the same validation based approach as above.
3) Matlab PEM assuming the order to be *known* (the exact order is 3).

Note that this example is extremely challenging for both SS-GLAR and VARX-GLAR as they compete against PEM which assumes the correct parametric structure with *known* order.

### D. Example 2

We consider a MISO ARMAX linear system with 15 inputs of the form

$$A(z)y_t = \sum_{i=1}^{15} B_i(z)u_t^i + C(z)e_t \tag{38}$$

The inputs are independent white Gaussian noises, zero mean and unit variance. The innovation process $\{e_t\}$ has unit variance. We perform $N_{MC} = 125$ Monte Carlo experiments and, for each run, a random ARMAX model is generated as follows:

- first the number $m_1$ of inputs affecting the system is generated as an uniform random variable in the set $\{0,1,2,..,8\}$.
- Then, w.l.o.g., we fix $B_{m_1+1}(z) = ... = B_m(z) = 0$ while the polynomials $A(z)$, $B_i(z)$, $i = 1,..,m_1$ are generated using the MATLAB function `drmodel.m`; the order is chosen at random in the interval $[1,30]$.
- The polynomial $C(z)$ is, with probability 0.5, fixed equal to $A(z)$ (i.e. we consider and Output Error model) and with probability 0.5 it is randomly generated using again the function `drmodel.m`.

The system and the predictor poles are restricted to lie inside the circle of radius 0.95 while the $\ell_2$ norm of each impulse response is bounded by 5 (`drmodel.m` is repeatedly called at any run until such requirements are fulfilled).

We compare the following estimators:

1) SS-GLAR described in Section IV; the parameters $\beta, \gamma_2$ and the level of sparsity (i.e. the number of nonzero impulse responses) is determined using the validation based approach described above.
2) VARX-GLAR described in Section III. The order of the VARX models is constrained to be between 1 and 30; both this order and the number of nonzero components have been estimated using the same validation based approach as above.
3) Matlab PEM, with order estimated using the validation based approach discussed above.

Average $RMS_1$

| | SS-GLAR | VARX-GLAR | PEM |
|---|---|---|---|
| Example 1 | 1.2378 | 1.2680 | 1.5644 |
| Example 2 | 1.2647 | 1.4095 | 2.4482 |

Average $Err_0$

| | SS-GLAR | VARX-GLAR | PEM |
|---|---|---|---|
| Example 1 | 0.1861 | 0.2582 | 0.2153 |
| Example 2 | 0.0718 | 0.1899 | 0.3658 |

Average $Err_1$

| | SS-GLAR | VARX-GLAR | PEM |
|---|---|---|---|
| Example 1 | 0.2806 | 0.2386 | 0.1066 |
| Example 2 | 0.3940 | 0.4824 | 0.7240 |

TABLE I

AVERAGE $RMS_1$ (TOP), $Err_0$ (CENTER) AND $Err_1$ (BOTTOM).

### E. Discussion of the results

The two examples highlight the potential of both VARX-GLAR and, in particular, SS-GLAR in estimating high-dimensional sparse dynamical systems. Example 1 shows that exploiting sparsity, as done by SS-GLAR and VARX-GLAR, allows to obtain very good performance even when compared to PEM which uses the correct order. Instead, when a more realistic setup in which the correct order is not known by PEM (and hence it has to be estimated from data), both SS-GLAR and VARX-GLAR outperform PEM; see Table V-E and Figures 2 and 3. Note that in this second simulation study the systems are randomized.

## VI. CONCLUSIONS AND FUTURE WORKS

In this paper we have proposed a new non-parametric estimator for high dimensional linear systems; input selection is performed efficiently via the Group LAR algorithm. Experimental evidence shows that the new algorithm outperforms classical parametric estimation methods when applied to systems with a large number of candidate inputs.

Future work will include more efficient methods for hyperparameters selection. Also richer structures which account for high frequency components in the predictors (as done in [25]) as well as flexible Kernels allowing different levels of regularity for different input channels will be studied.

## REFERENCES

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

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

[3] T. McKelvey, A. Helmersson, and T. Ribarits, "Data driven local coordinates for multivariable linear systems and their application to system identification," *Automatica*, vol. 40, pp. 1629–1635, 2004.

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

[5] W. Larimore, "System identification, reduced-order filtering and modeling via canonical variate analysis," in *Proc. American Control Conference*, 1983, pp. 445–451.

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

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

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

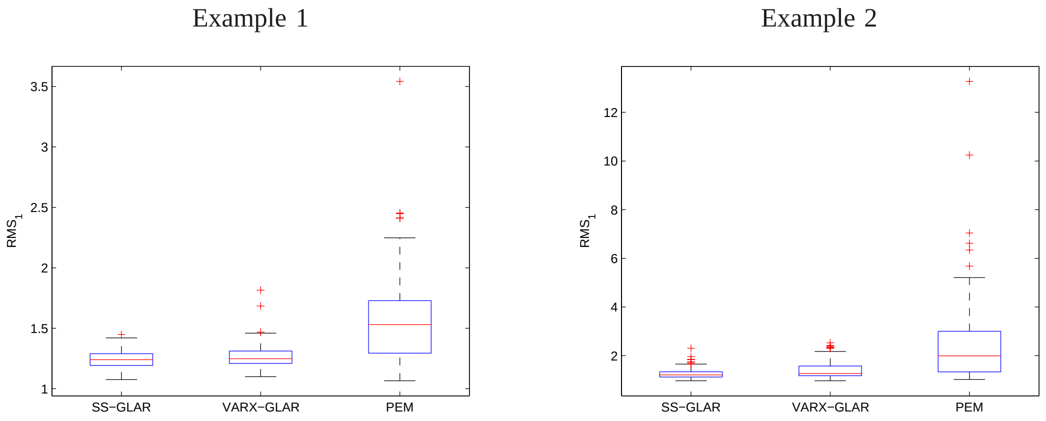

Fig. 2. Boxplots of one step-ahead prediction error. Left: Example 1. Right: Example 2.

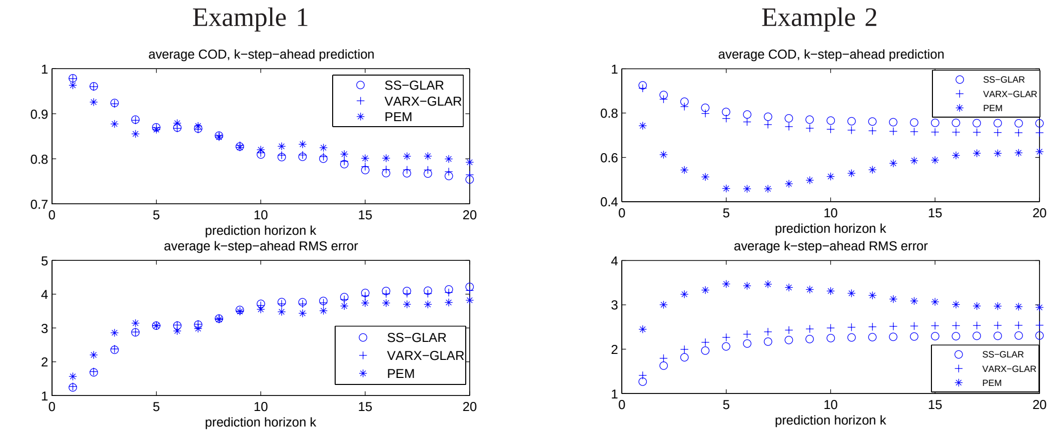

Fig. 3. Average $RMS_k$ (k-one step-ahead prediction error) and average Coefficient of Determination (COD). Left: Example 1. Right: Example 2.

[9] 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*, vol. 69, no. 1, pp. 63–78, 2007.

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

[11] J. Haupt, W. Bajwa, G. Raz, and R. Nowak, "Toeplitz compressed sensing matrices with applications to sparse channel estimation," in *Proc. 42nd Annual Conference on Information Sciences and Systems*, 2008.

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

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

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

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

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

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

[18] S. Bakin, "Adaptive regression and model selection in data mining problems," Ph.D. dissertation, The Australian National University, 1999.

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

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

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

[22] 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*, vol. 69, no. 3, pp. 329–346, 2007.

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

[24] G. De Nicolao and G. Pillonetto, "A new kernel-based approach for system identification," in *Proceedings of the 2008 American Control conference, Seattle, USA*, 2008.

[25] G. Pillonetto, A. Chiuso, and G. De Nicolao, "Predictor estimation via Gaussian regression," in *Proc. of IEEE Conf. on Dec. and Control*, 2008.

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

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

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

[29] H. Sun, "Mercer theorem for RKHS on noncompact sets," *Journal of Complexity*, vol. 21, pp. 337–349, 2005.

[30] F. Cucker and S. Smale, "On the mathematical foundations of learning," *Bulletin of the American mathematical society*, vol. 39, pp. 1–49, 2001.