Available online at www.sciencedirect.com

ScienceDirect

IFAC PapersOnline CONFERENCE PAPER ARCHIVE

IFAC PapersOnLine 51-15 (2018) 31–36

Kernel-based SPS

Gianluigi Pillonetto *, **Algo Carè** **, **Marco C. Campi** **

* *Dept. of Information Engineering, University of Padova, Italy*
** *Dept. of Information Engineering, University of Brescia, Italy*

Abstract: One of the central issues in system identification consists not only in obtaining a good model of the process under study but also an informative confidence interval around it. This problem is often referred to as *robust identification* in the literature. Following the classical paradigm, one first obtains a model through prediction error minimization. Asymptotic theory is then invoked to extract quality tags from the normal approximation of the estimates' distribution. This paper proposes an alternative route for robust linear system identification. Our procedure relies on the use of kernel-based regularization for both impulse response estimation and confidence intervals computation. The main novelty is that the kernel is not used to define a Gaussian density for the impulse response but just a prior satisfying some symmetry properties forming the basis of the recently developed *sign-perturbed sums* (SPS) framework. For system identification, SPS is then combined with the *stable spline* (SS) kernel to account for impulse response regularity and exponential stability. Numerical experiments show that SS+SPS can provide more accurate confidence intervals than those commonly achieved in the Gaussian regression framework (which, in turn, were already shown to outperform those based on the classical paradigm).

© 2018, IFAC (International Federation of Automatic Control) Hosting by Elsevier Ltd. All rights reserved.

Keywords: linear system identification; Gaussian processes; sign-perturbed sums; kernel-based regression; stable spline

1. INTRODUCTION

Classical approaches to system identification are based on maximum likelihood and prediction error minimization (PEM) (Ljung, 1997; Söderström and Stoica, 1989). The general workflow includes first the selection of a set of candidate models that increase in complexity, e.g. FIR, ARX or ARMAX of different dimensions in the linear scenario. Each of them is then fitted to data by PEM and the "best" model is selected using a variety of complexity measures such as the Akaike or the Bayesian information criterion. One important limitation of these criteria is that they rely on asymptotic arguments, being derived assuming the availability of infinite data lengths. As illustrated e.g. in (Pillonetto et al., 2014), this can lead to models with poor prediction capability on future data. In this regard, recent research has shown that the use of regularization may lead to significant improvements (Pillonetto and De Nicolao, 2010; Chen et al., 2012). The impulse response of the unknown system is modeled in a Bayesian setting as a zero-mean Gaussian process whose covariance is also called *kernel* in the machine learning literature (Schölkopf and Smola, 2001). The suggested model is the so called *stable-spline kernel* which includes information on system exponential stability. In this framework, the choice of model (discrete) orders is replaced by the (continuous) tuning of few kernel parameters through non-asymptotic approaches, e.g. empirical Bayes or Stein's unbiased risk (Efron and Morris, 1973; Maritz and Lwin, 1989; Hastie et al., 2001). Empirical and theoretical arguments which support this approach to linear system identification are also described in (Bell and Pillonetto, 2004; Aravkin et al., 2014; Pillonetto and Chiuso, 2015).

However, beyond a good model of the process under study, one of the central issues in system identification is also the determination of an informative confidence interval. This problem is often referred to as *robust identification* in the literature (Goodwin et al., 1992). The classical approach still relies on asymptotic theory: quality tags are obtained from normal approximation of the estimates' distribution. Under the Bayesian framework mentioned above, uncertainty regions can instead be directly derived from the a posteriori distribution. In fact, once the Gaussian prior on the impulse response is accepted, the posterior becomes available in closed form. The numerical studies illustrated in (Prando et al., 2016) have shown that the uncertainty regions so obtained are in general more accurate than the ones returned by the asymptotic approximation. One of the key reasons is the prior's ability to constrain all the estimates in the stability region whereas the "asymptotic" region cannot guarantee this. Note also that comparison between the confidence intervals derived under a frequentist framework and the Bayes intervals is a widely discussed topic, e.g. see (Efron, 2005) and also (Wahba, 1983) for a discussion focused on the smoothing splines case.

This paper proposes an alternative route for robust linear system identification. As in (Pillonetto and De Nicolao, 2010; Prando et al., 2016), our procedure relies on the use of kernel-based regularization for both impulse response estimation and confidence intervals computation. But the main novelty here is that the kernel is not used to define a Gaussian density for the impulse response. The prior instead incorporates a much milder symmetry property which forms the basis of the recently developed *sign-perturbed sums* (SPS) framework (Campi and Weyer, 2005; Csáji et al., 2015; Carè et al., 2018). SPS is then combined with the *stable spline* (SS) kernel to include information on impulse response stability. The method is tested using numerical experiments where noisy data come from output error models defined by randomly generated rational transfer functions. Results show that SS+SPS can provide more accurate confidence intervals than those achieved in the Gaussian regression framework.

The paper is organized as follows. Section 2 formulates the problem. Section 3 reports four numerical procedures for robust

2405-8963 © 2018, IFAC (International Federation of Automatic Control) Hosting by Elsevier Ltd. All rights reserved.
Peer review under responsibility of International Federation of Automatic Control.
10.1016/j.ifacol.2018.09.086

linear system identification, then tested via numerical experiments in Section 4. Conclusions then end the paper.

2. PROBLEM STATEMENT

Our aim is to identify a discrete-time linear and stable system from noisy output measurements. An output-error structure is postulated. In particular, a FIR of (possibly high) dimension m is adopted and the measurements model is

$$y = \Phi\theta^0 + v \tag{1}$$

where the vector $y \in \mathbb{R}^n$ contains the noisy outputs, the components of $\theta^0 \in \mathbb{R}^m$ are the impulse response coefficients, Φ is a known exogenous regression matrix independent of v built with the system inputs (with $\Phi^T\Phi$ of full rank) and v is the noise vector. Following the framework developed in (Csáji et al., 2015), the following assumption is stated.

Assumption 1. The components of the noise v in (1) are independent random variables with a symmetric probability distribution around zero.

The problem is now to obtain an estimate of θ^0 from y and also an informative confidence interval around it.

3. LS+SPS, RELS+GAUSS AND RELS+SPS

In this section, four procedures to derive the estimate of θ^0 and a confidence interval around it are described. The first one is called LS+SPS. It relies on least squares and coincides with that discussed in (Csáji et al., 2015). The other three, namely ReLS+Gauss and two versions of ReLS+SPS, make instead use of regularization.

3.1 LS+SPS

Algorithm 1 SPS-initialization given a matrix $\Omega \in \mathbb{R}^{N\times m}$

1: Define a (rational) confidence probability $p \in (0,1)$ and set integers $r > q > 0$ such that $p = 1 - q/r$;
2: Calculate R_N and $R_N^{1/2}$ where
$$R_N = \frac{\Omega^T\Omega}{N}, \quad R_N^{1/2}(R_N^{1/2})^T = R_N;$$
3: Generate $N(r-1)$ i.i.d. random signs $\{\alpha_{i,t}\}$ with
$$\mathbb{P}(\alpha_{i,t} = 1) = \mathbb{P}(\alpha_{i,t} = -1) = 1/2,$$
for $i = 1,\dots r-1$ and $t = 1,\dots,N$;
4: Generate a random perturbation π of the set $\{0,1,\dots,r-1\}$, where each of the $r!$ possible perturbations has the same probability to be selected.

The simplest approach to estimate θ^0 is the least squares (LS) estimator

$$\hat{\theta}^{LS} = (\Phi^T\Phi)^{-1}\Phi^T y. \tag{2}$$

As for the uncertainty around $\hat{\theta}^{LS}$, under Assumption 1 an exact confidence interval (CI) is characterized by the SPS procedure developed in (Csáji et al., 2015) and summarized in Algorithms 1 and 2. In the procedures, θ is a given vector and Algorithm 2 checks whether the given θ belongs to the CI. In what follows, we use $\mathscr{I}$ to denote a set of candidate θ whose choice will be discussed later on in the numerical experiments section 4.2. If $\mathscr{I}$ is sufficiently rich, a good CI approximation is then achieved by Algorithm 3.

Algorithm 2 SPS-indicator(θ) given a matrix $\Omega \in \mathbb{R}^{N\times m}$ and a vector $z \in \mathbb{R}^N$

1: For the given θ, compute the prediction errors
$$\varepsilon_t(\theta) = z_t - \Omega(t,:)\theta, \quad t = 1,\dots,N$$
where $\Omega(t,:)$ is the t-th row of Ω;
2: Evaluate for $i = 1,2,\dots,r-1$
$$S_0(\theta) = R_N^{-1/2}\frac{1}{N}\sum_{t=1}^{N}\Omega(t,:)^T\varepsilon_t(\theta)$$
and
$$S_i(\theta) = R_N^{-1/2}\frac{1}{N}\sum_{t=1}^{N}\alpha_{i,t}\Omega(t,:)^T\varepsilon_t(\theta);$$
3: Order the scalars $\{\|S_i(\theta)\|\}$ in increasing order. If $\|S_a(\theta)\| = \|S_b(\theta)\|$, $\|S_a(\theta)\|$ preceeds $\|S_b(\theta)\|$ iff $\pi(a) < \pi(b)$;
4: Compute the rank $\mathscr{R}(\theta)$ of $\|S_0(\theta)\|$ in the ordering, e.g. $\mathscr{R}(\theta) = 1$ if $\|S_0(\theta)\|$ is the smallest one;
5: Return "accept" if $\mathscr{R}(\theta) \le r - q$.

Algorithm 3 LS+SPS

1: Compute the LS estimate (2);
2: Define a set $\mathscr{I}$ of candidate impulse responses;
3: Initialize the SPS procedure using Algorithm 1 with $\Omega = \Phi$, setting e.g. $q = 5$ and $r = 100$ to obtain a 95% CI;
4: For each $\theta \in \mathscr{I}$ use Algorithm 2 with $\Omega = \Phi$ and $z = y$ to accept or refuse the candidate impulse response. Call the accepted subset of $\mathscr{I}$ the CI sampled version and denote it with $\mathscr{C}$;
5: Return LS estimate (2) and CI sampled version $\mathscr{C}$.

3.2 ReLS+Gauss

The main problem of the estimator (2) is that it can suffer of high variance due to ill-conditioning. In these cases the introduction of regularization is important and one option is to resort to Bayesian estimation. In particular, in the Gaussian regression setting, both θ^0 and the noise v are modeled as (independent) normal vectors, i.e.

$$\theta^0 \sim \mathscr{N}(\mu, \lambda^2\Sigma), \quad v \sim \mathscr{N}(0, \sigma^2 I_n), \tag{3}$$

with λ^2 and σ^2 positive scale factors. The mean μ and covariance $\lambda^2\Sigma$ thus embed our prior information on θ^0. As for the noise, note that v satisfies Assumption 1 but is now constrained to be stationary and Gaussian.
Under the normal assumptions reported above, the posterior distribution of θ^0 given y is

$$\theta^0|y \sim \mathscr{N}\left(\hat{\theta}^B, (\frac{\Phi^T\Phi}{\sigma^2} + \frac{\Sigma^{-1}}{\lambda^2})^{-1}\right), \tag{4}$$

where $\hat{\theta}^B$ is the minimum variance estimate characterized by

$$\hat{\theta}^B = \arg\min_{\theta} \|y - \Phi\theta\|^2 + \eta^2(\theta-\mu)^T\Sigma^{-1}(\theta-\mu) \tag{5a}$$

$$= \mu + (\Phi^T\Phi + \eta^2\Sigma^{-1})^{-1}\Phi^T(y - \Phi\mu). \tag{5b}$$

In (5), the scalar $\eta^2 = \sigma^2/\lambda^2$ is the so called regularization parameter which balances the adherence to experimental data and to the prior information on θ^0.
For future developments, it is important to stress that the regularized least squares (ReLS) estimator (5) can be derived also in a Fisherian context. We can come back to see θ^0 as deterministic and add the following *virtual measurements* to (1)

$$\widetilde{y} = \widetilde{\Phi}\theta^0 + \widetilde{v} \tag{6a}$$
$$\widetilde{y} := \eta\Sigma^{-1/2}\mu, \quad \widetilde{\Phi} := \eta\Sigma^{-1/2}, \quad \widetilde{v} \sim \mathcal{N}(0, \sigma^2 I_m). \tag{6b}$$

The new measurements model becomes

$$z = \Omega\theta^0 + e \tag{7a}$$
$$z = \begin{pmatrix} y \\ \widetilde{y} \end{pmatrix}, \quad \Omega = \begin{pmatrix} \Phi \\ \widetilde{\Phi} \end{pmatrix}, \quad e \sim \mathcal{N}(0, \sigma^2 I_{n+m}). \tag{7b}$$

The maximum likelihood (least squares) estimate of θ^0 is then given by

$$\hat{\theta}^{ReLS} = (\Omega^T\Omega)^{-1}\Omega^T z \tag{8a}$$
$$= \mu + (\Phi^T\Phi + \eta^2\Sigma^{-1})^{-1}\Phi^T(y - \Phi\mu) \tag{8b}$$

and coincides with (5).

Under the linear Gaussian model (7) we can use as confidence region

$$\Theta = \left\{\theta \ : \ (\theta - \hat{\theta}^{ReLS})^T\Omega^T\Omega(\theta - \hat{\theta}^{ReLS}) \leq \kappa\sigma^2\right\}, \tag{9}$$

where κ regulates the confidence level, see subsection 3.B in (Csáji et al., 2015). Then, given a sufficiently rich set $\mathscr{I}$ of candidates, a good CI approximation is now returned by Algorithm 4.

Algorithm 4 ReLS+Gauss

1: Compute the ReLS estimate (8);
2: Define a set $\mathscr{I}$ of candidate impulse responses;
3: Determine the $\theta \in \mathscr{I}$ which falls in the confidence region (9), with the desired level defined by κ. Call the accepted subset of $\mathscr{I}$ the CI sampled version $\mathscr{C}$;
4: Return ReLS estimate (8) and CI sampled version $\mathscr{C}$.

Example 2. (Stable Spline). To embed information on system exponential stability, one can adopt the stable spline prior (Pillonetto et al., 2014) setting $\mu = 0$ and using as covariance the matrix $\lambda^2\Sigma_\alpha$, where the (i, j) entry of Σ_α is

$$[\Sigma_\alpha]_{i,j} = \alpha^{\max(i,j)}, \ \ 0 \leq \alpha < 1. \tag{10}$$

The scalar α is an additional hyperparameter which regulates the impulse response decay rate. Using the inverse of the Cholesky factor of Σ_α (Chen et al., 2016), after simple calculations the model (6) becomes

$$0 = \eta\frac{\theta_1^0 - \theta_2^0}{\sqrt{\alpha(1-\alpha)}} + \widetilde{v}_1 \tag{11a}$$
$$\vdots \tag{11b}$$
$$0 = \eta\frac{\theta_{m-1}^0 - \theta_m^0}{\sqrt{\alpha^{m-1}(1-\alpha)}} + \widetilde{v}_{m-1} \tag{11c}$$
$$0 = \eta\frac{\theta_m^0}{\sqrt{\alpha^m}} + \widetilde{v}_m. \tag{11d}$$

Note that the *virtual measurements* provide the information that the variances of θ_k^0 and of the Gaussian increments $\theta_k^0 - \theta_{k+1}^0$ decay exponentially to zero as k increases.

3.3 ReLS+SPS1

As graphically depicted in the top panel of Fig. 1, in the Gaussian regression context we can see θ^0 as the output of a linear operator (that defines the kernel) fed with a stationary white Gaussian noise. For instance, in the smoothing splines case, the linear system is a cascade of integrators (Wahba, 1990), whereas the covariance (10) derives from a particular anti-causal integrator (Pillonetto and De Nicolao, 2010). The change of perspective to build a new confidence interval around the ReLS estimate (8) is shown in the bottom panel of the same figure. The difference is that the white noise input used to introduce the expected properties of θ^0 is no more assumed Gaussian but follows the assumptions underlying the SPS framework. Converting this Bayesian view in the Fisherian context based on the *virtual measurements*, model (7) becomes

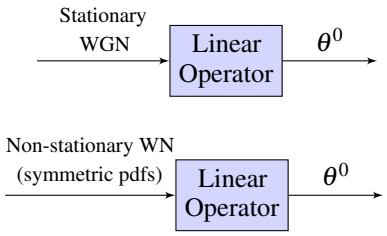

Fig. 1. *Top: Gaussian regression* Prior information is formulated modeling θ^0 as the output of a linear system fed with stationary white Gaussian noise (WGN). *Bottom: SPS regression* The information on θ^0 is built using a non-stationary white noise (WN) with probability density functions (pdfs) just symmetric around zero. Mean μ is assumed null to simplify the figure.

$$z = \Omega\theta^0 + e \tag{12a}$$
$$z = \begin{pmatrix} y \\ \widetilde{y} \end{pmatrix}, \quad \Omega = \begin{pmatrix} \Phi \\ \widetilde{\Phi} \end{pmatrix}, \quad e \text{ satisfies Assumption 3} \tag{12b}$$

with $\widetilde{y}$ and $\widetilde{\Phi}$ as in (6b), and

Assumption 3. The components of the noise e in (12) are independent random variables with a symmetric probability distribution around zero.

The consequence of this new framework can be also appreciated reconsidering the model (11) induced by the stable spline kernel. The information on increments $\theta_k^0 - \theta_{k+1}^0$ of the impulse response is no more connected with a Gaussian measure of known variance but with a (much more general) pdf symmetric around zero. The new procedure is called ReLS+SPS1 and is implemented by Algorithm 5.

Algorithm 5 ReLS+SPS1

1: Compute the ReLS estimate (8);
2: Define a set $\mathscr{I}$ of candidate impulse responses;
3: Initialize the SPS procedure using Algorithm 1 with Ω given by (12), setting e.g. $q = 5$ and $r = 100$ to obtain a 95% CI;
4: For each $\theta \in \mathscr{I}$ use Algorithm 2 with Ω and z in (12) to accept or refuse the candidate impulse response. Call the accepted subset of $\mathscr{I}$ the CI sampled version $\mathscr{C}$;
5: Return ReLS estimate (8) and CI sampled version $\mathscr{C}$.

3.4 ReLS+SPS2

In this section, we present an alternative procedure with additional theoretical properties. Start by noting that in Algorithm 2 the random signs $\alpha_{i,t}$ apply to all terms in the summation in the $S_i(\theta)$ functions, so that, in the case of ReLS+SPS1, random signs apply to terms that come from data as well as

terms associated with the virtual measurements. More explicitly, a function $S_i(\theta)$ in Algorithm 2 as used by ReLS+SPS1 (Algorithm 5) can be written as

$$S_i(\theta) = \frac{R_N^{-1/2}}{N}\left[\sum_{t=1}^{n}\alpha_{i,t}\Phi(t,:)^T\varepsilon_t(\theta) + \sum_{t=1}^{m}\alpha_{i,t+n}\widetilde{\Phi}(t,:)^T\varepsilon_{t+n}(\theta)\right], \quad (13)$$

where we recall that $\varepsilon_t(\theta) = y_t - \Phi(t,:)\theta$, $t = 1,\dots,n$, are associated with the data while $\varepsilon_{t+n}(\theta) = \tilde{y}_t - \widetilde{\Phi}(t,:)\theta$, $t = 1,\dots,m$, with the virtual measurements. In this section we introduce the alternative approach of applying the random signs only to the data based terms and consider the functions

$$S_i(\theta) = \frac{R_N^{-1/2}}{N}\left[\sum_{t=1}^{n}\alpha_{i,t}\Phi(t,:)^T\varepsilon_t(\theta) + \sum_{t=1}^{m}\widetilde{\Phi}(t,:)^T\varepsilon_{t+n}(\theta)\right]. \quad (14)$$

Algorithm 2 where (14) is used in place of (13) is in the sequel referred to as Algorithm 2'. Using Algorithm 2' in Algorithm 5 gives the following procedure.

Algorithm 6 ReLS+SPS2

Same as Algorithm 5 where, at step 4, Algorithm 2' is used in place of Algorithm 2.

To this approach, the following theorem applies which extends the theory valid for LS+SPS, see Csáji et al. (2015).

Theorem 4. Consider model (1) where θ^0 has a deterministic value. Under Assumption 1, for any value of θ^0, Algorithm 2' returns "accept" when applied to θ^0 with exact probability $p = 1 - q/r$.

Remark 5. Model (1) requires that the data generating system is an FIR (finite impulse response) system; however, the theorem can be approximately applied to IIR (infinite impulse response) systems after the system is approximated by a model in the form (1) where Φ contains a long enough tail of past input values. For undermodelling detection and a study of the influence of undermodelling on SPS techniques, see Carè et al. (2017).

Due to space limitations, we here only provide a sketch of the proof of Theorem 4 while the reader is also referred to Volpe (2015) for more details. Corresponding to θ^0, the functions $S_i(\theta)$ in (14) take the form

$$S_i(\theta^0) = \frac{R_N^{-1/2}}{N}\left[\sum_{t=1}^{n}\alpha_{i,t}\Phi(t,:)^T v_t + \eta^2\Sigma^{-1}(\mu - \theta^0)\right], \quad (15)$$

while $S_0(\theta)$ is given by

$$S_0(\theta^0) = \frac{R_N^{-1/2}}{N}\left[\sum_{t=1}^{n}\Phi(t,:)^T v_t + \eta^2\Sigma^{-1}(\mu - \theta^0)\right]. \quad (16)$$

Comparing these two expressions, one notices that the probability distributions of (15) and (16) are identical because $\alpha_{i,t}v_t$ and v_t have the same distribution since v_t has symmetric probability distribution around zero, so that none of the two variables carries a probability higher than the other to be bigger. [1] Extending this reasoning to all functions $S_0(\theta)$, $S_i(\theta)$, $i = 1,2,\dots,r-1$, one can conclude that $\mathscr{R}(\theta^0) \le r - q$, so that θ^0 is accepted by Algorithm 2', with probability $p = 1 - q/r$.

Notice that Theorem 4 holds true independently of the way θ^0 is generated. This result, instead, does not apply to ReLS+SPS1, for which an analysis (which is beyond the scope of this paper and therefore not included in this contribution) can be developed under the more stringent Assumption 3. We only notice that the validity of Assumption 3 requires that $\tilde{y} - \widetilde{\Phi}\theta^0$ be an independent random vector with symmetric distribution.

4. NUMERICAL EXPERIMENTS

4.1 Set-up of a Monte Carlo experiment

We will consider two Monte Carlo studies of 1000 runs each. At any run a different transfer function of order 10 is randomly generated as follows. Poles and zeros are chosen iterating the following procedure: with equal probability a real or a couple of complex conjugate poles is added to the numerator and denominator until their order reaches 10. In the case of a real pole, it is randomly drawn from a uniform distribution on $[-0.95, 0.95]$, while the absolute value and phase of one of the complex conjugate poles are independent random variables uniform on $[0, 0.95]$ and $[0, \pi]$, respectively. Measurement noise is white and Gaussian with standard deviation set to $1/5$ of that of the noiseless output. In the first Monte Carlo study the input is white Gaussian noise of unit variance. In the second, it is white Gaussian noise filtered by a second-order system (randomly generated at any run with the same procedure described above). Identification data comprise $N = 800$ outputs and the dimension of θ is set to $m = 100$.

4.2 Implementation details

ReLS is implemented using the stable spline kernel Σ_α in (10). The structure of the resulting estimator (5) is thus known except for the regularization parameter η^2 and the decay rate α. They are estimated at any Monte Carlo run via marginal likelihood optimization (Pillonetto et al., 2014)[Section 4.4], which also returns the noise variance estimate $\hat{\sigma}^2$ used to define the confidence region Θ in (9).

To implement LS+SPS, the set $\mathscr{I}$ of candidate θ contains 20000 vectors. In particular, 10000 vectors are obtained from the posterior (4) but adopting an improper uniform prior on the impulse response, i.e. from

$$\mathscr{N}\left((\Phi^T\Phi)^{-1}\Phi^T y, \hat{\sigma}^2(\Phi^T\Phi)^{-1}\right).$$

The final 10000 samples are generated via a Metropolis random walk (Gilks et al., 1996) with Gaussian increments of covariance $\hat{\Sigma} := (\frac{\Phi^T\Phi}{\hat{\sigma}^2})^{-1}/10$. Note that the generation of the candidates for LS+SPS does not use the kernel since this procedure does not exploit any kind of regularization.

To implement ReLS+SPS1, ReLS+SPS2 and ReLS+Gauss, the set $\mathscr{I}$ of candidate θ still contains 20000 vectors. However, it is not convenient to use the same set $\mathscr{I}$ adopted by LS+SPS since, thanks to the introduction of the kernel, in practice most of these candidates will be refused. This holds especially for ReLS+SPS1 and ReLS+Gauss where the virtual measurements/prior play a much important role in determining the confidence region. For these reasons, only 10000 candidate vectors are generated by the strategy used for LS+SPS, i.e. 5000 from $\mathscr{N}\left((\Phi^T\Phi)^{-1}\Phi^T y, \hat{\sigma}^2(\Phi^T\Phi)^{-1}\right)$ and 5000 from the Metropolis

[1] The statement that the distributions of (15) and (16) are identical is rigorous as long as η, Σ and μ are deterministic parameters that do not depend on the dataset. It is important to remark, however, that in everyday practice these parameters are often estimated from data, in which case η, Σ and μ carry a dependence on v so that the distributions of (15) and (16) are no longer rigorously identical. On the other hand, it is expected that the stochastic fluctuation associated with η, Σ and μ is moderate and thus Theorem 4 still holds approximately. The numerical results in Section 4 confirm this intuition.

random walk. The other 10000 are obtained exploiting the information that the region's shape depends also on the kernel. Hence, other 5000 candidates are independent samples from (4), i.e. from a Gaussian distribution with mean $\hat{\theta}^B$ and covariance $\hat{\Sigma} := (\frac{\Phi^T\Phi}{\hat{\sigma}^2} + \frac{\Sigma^{-1}}{\lambda^2})^{-1}$. Other 5000 are generated through a random walk Metropolis (Gilks et al., 1996) with Gaussian increments of covariance $\hat{\Sigma}/10 + \kappa I$. The matrix κI allows to generate candidates with components significantly different from zero in the tail of the impulse response. The scale factor κ is initially set to the maximum value of the LS estimate divided by 100 and then tuned at any run via a pilot analysis to obtain an acceptance rate around 30%.

4.3 Performance indexes computation

Let θ^0 and $\hat{\theta}$ indicate at a generic run the true impulse response and its LS or ReLS estimate, respectively. First, it is checked if the true θ^0 is contained in the 95% confidence intervals (CI) defined by LS+SPS, ReLS+Gauss and the two versions of ReLS+SPS. Then, the classical impulse response (IR) fit is computed as

$$100\left(1 - \frac{\|\theta^0 - \hat{\theta}\|}{\|\theta^0\|}\right) \qquad \textbf{IR fit}. \tag{17}$$

Next, sampled versions $\mathscr{C}$ of the 95% CIs associated to LS+SPS, ReLS+Gauss, ReLS+SPS1 and ReLS+SPS2 are obtained by Algorithms 3, 4, 5 and 6. From $\mathscr{C}$ two performance indexes are extracted. The first one is called CI area. Letting h_i be the i-th component of a generic $h \in \mathscr{C}$, it is defined by

$$\sum_{i=1}^{m=100}\left(\max_{h\in\mathscr{C}} h_i - \min_{h\in\mathscr{C}} h_i\right) \qquad \textbf{CI area}. \tag{18}$$

This index is thus a rough measure of the CI dispersion.
The second index is called CI fit and is computed only if the CI does not contain the true θ. It is defined by

$$\max_{h\in\mathscr{C}} 100\left(1 - \frac{\|\theta^0 - h\|}{\|\theta^0\|}\right) \qquad \textbf{CI fit} \tag{19}$$

and thus describes the nearness of the sampled CI to the true impulse response.

4.4 Results

Table 1 allows one to assess the frequency with which the 95% CIs defined by the four procedures contain the true θ^0 for the two Monte Carlo studies. For LS+SPS the value is very close to 95%. This is expected from the SPS theory: letting the number of runs grow to infinity, convergence to 95% would hold. When using ReLS+Gauss, in both the case studies the level is around 70%. A significant improvement is obtained by ReLS+SPS1: the percentage is always larger than 85% and close to 90% for white noise input. Notice, however, that such values would not converge to the confidence level 95% for increasingly many Monte Carlo runs since, as hinted at in Section 3.4, the validity of an exact confidence result for ReLS+SPS1 hinges upon Assumption 3, which is violated by the data generation mechanism used in this example. Finally, from Table 1 we see that the percentage for ReLS+SPS2 is very close to 95%, as expected from Theorem 4 and Remark 5.
Fig. 2 displays the performance indexes (17-19) for input equal to white noise (top panels) and filtered white noise (bottom). As expected, ReLS outperforms LS in terms of IR fit (17), especially in the second Monte Carlo study where most of the problems are severely ill-conditioned. Boxplots of the CI area indexes (18) then show that the uncertainty regions returned by the kernel-based estimators are much more compact than those computed by LS+SPS. This is further illustrated in Fig. 3 which displays the sets $\mathscr{C}$ (which approximate the 95% CIs) obtained in one Monte Carlo run. The performance of ReLS+SPS2 is worse than that of ReLS+SPS1, but still considerably better than LS+SPS. Finally, the right panels of Fig. 2 report the CI fit indexes (19) (the boxplots are separate since they contain a different number of values). The kernel-based approaches provide similar results, much better than LS+SPS.

Table 1. 95% CI accuracy for the two Monte Carlo studies of 1000 runs

	LS+SPS	**ReLS+Gauss**	**ReLS+SPS1**	**ReLS+SPS2**
WN input	95.2 %	70.1%	89.3%	95.5%
filtered WN	95.1%	72.7%	84.7%	94.8%

5. CONCLUSIONS

In this paper, we have introduced kernel-based SPS methods, i.e. methods for robust linear system estimation that rely on kernel-based regularization.
In Gaussian regression, a prior is postulated according to which the model parameter θ^0 is the output of a linear system fed with a stationary sequence of white Gaussian random variables. This prior has been relaxed in this paper by replacing the stationary Gaussian sequence with a non-stationary sequence of independent and symmetric, but otherwise arbitrarily distributed, random variables.
A rigorous analysis of the theoretical properties of ReLS+SPS1 has not been included due to space limitations. In future work it will be shown that ReLS+SPS1 can be cast in a joint frequentist-Bayesian framework (Bayarri and Berger, 2004). In particular, instead of resorting to virtual measurements in a Fisherian context, one can assume that θ^0 satisfies the prior described in the bottom panel of Fig. 1 and then show that the algorithm delivers guaranteed confidence regions on average over θ^0. On the other hand, ReLS+SPS2 delivers guaranteed regions for every value of θ^0, so that this result is immune to misspecifications of the prior. Simulations results show that ReLS+SPS1 builds smaller confidence regions than ReLS+SPS2. This can be interpreted considering that ReLS+SPS1 exploits, at least partly, the information contained in the prior. In future work, we plan to theoretically study the shape and size of the regions provided by these two algorithms. Under a numerical point of view, the reconstruction of SPS regions in sampled form is also an important point which will deserve further study. For this purpose, we plan to design a more sophisticated and efficient MCMC scheme. This will likely allow to define more rigorous regions dispersion indexes to be used also to monitor the convergence of the generated chains.
In conclusion, there is evidence that kernel-based SPS methods combine some of the advantages of regularized methods with the robustness of guaranteed SPS algorithms, and further analysis is required to understand the potentials and the limits of this new approach.

REFERENCES

Aravkin, A., Burke, J., Chiuso, A., and Pillonetto, G. (2014). Convex vs non-convex estimators for regression and sparse estimation: the MSE properties of ARD and GLasso. *The Journal of Machine Learning Research*, 15(1), 217–252.

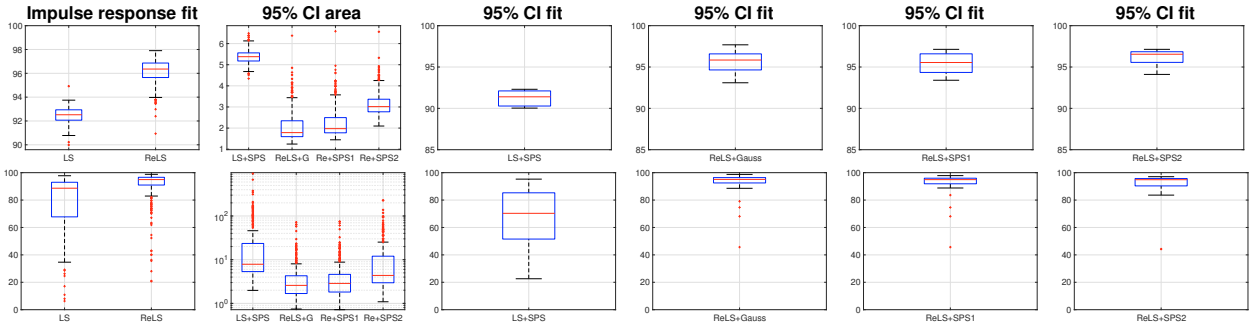

Fig. 2. *Monte Carlo study* Impulse response fits and performance indices of the 95% confidence interval obtained by LS+SPS, ReLS+Gauss, ReLS+SPS1 and ReLS+SPS2 with input equal to white noise (top panels) and to filtered white noise (bottom panels).

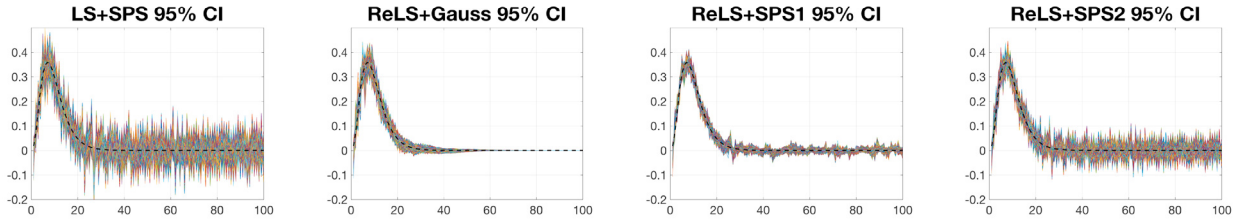

Fig. 3. *Results from a Monte Carlo run* True impulse response (thick dashed line) and 95% confidence intervals returned by LS+SPS, ReLS+Gauss, ReLS+SPS1 and ReLS+SPS2 using filtered white noise as input.

Bayarri, M.J. and Berger, J.O. (2004). The interplay of bayesian and frequentist analysis. *Statistical Science*, 19(1), 58–80.

Bell, B. and Pillonetto, G. (2004). Estimating parameters and stochastic functions of one variable using nonlinear measurement models. *Inverse Problems*, 20(3), 627.

Campi, M.C. and Weyer, E. (2005). Guaranteed non-asymptotic confidence regions in system identification. *Automatica*, 41(10), 1751 – 1764.

Carè, A., Campi, M.C., Csáji, B.Cs., and Weyer, E. (2017). Undermodelling detection with sign-perturbed sums. In *20th IFAC World Congress*, 2744 – 2749.

Carè, A., Csáji, B.Cs., Campi, M.C., and Weyer, E. (2018). Finite-sample system identification: An overview and a new correlation method. *IEEE Control Systems Letters*, 2(1), 61–66. doi:10.1109/LCSYS.2017.2720969.

Chen, T., Ardeshiri, T., Carli, F., Chiuso, A., Ljung, L., and Pillonetto, G. (2016). Max-entropy properties of discrete-time first-order stable spline kernel. *Automatica*, 66(1), 34–38.

Chen, T., Ohlsson, H., and Ljung, L. (2012). On the estimation of transfer functions, regularizations and Gaussian processes - revisited. *Automatica*, 48(8), 1525–1535.

Csáji, B.Cs., Campi, M.C., and Weyer, E. (2015). Sign-perturbed sums: A new system identification approach for constructing exact non-asymptotic confidence regions in linear regression models. *IEEE Transactions on Signal Processing*, 63(1), 169–181.

Efron, B. (2005). Bayesians, frequentists, and scientists. *Journal of the American Statistical Association*, 100, 1–5.

Efron, B. and Morris, C. (1973). Stein's estimation rule and its competitors–an empirical Bayes approach. *Journal of the American Statistical Association*, 68(341), 117–130.

Gilks, W., Richardson, S., and Spiegehalter, D. (1996). *Markov Chain Monte Carlo in Practice*. London: Chapman and Hall.

Goodwin, G., Gevers, M., and Ninness, B. (1992). Quantifying the error in estimated transfer functions with application to model order selection. *IEEE TAC*, 37(7), 913–928.

Hastie, T.J., Tibshirani, R.J., and Friedman, J. (2001). *The Elements of Statistical Learning. Data Mining, Inference and Prediction*. Springer, Canada.

Ljung, L. (1997). *System Identification, Theory for the User*. Prentice Hall.

Maritz, J.S. and Lwin, T. (1989). *Empirical Bayes Method*. Chapman and Hall.

Pillonetto, G. and Chiuso, A. (2015). Tuning complexity in regularized kernel-based regression and linear system identification: the robustness of the marginal likelihood estimator. *Automatica*, 58, 106–117.

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

Pillonetto, G., Dinuzzo, F., Chen, T., De Nicolao, G., and Ljung, L. (2014). Kernel methods in system identification, machine learning and function estimation: a survey. *Automatica*, 50(3), 657–682.

Prando, G., Romeres, D., Pillonetto, G., and Chiuso, A. (2016). Classical vs. bayesian methods for linear system identification: Point estimators and confidence sets. In *2016 European Control Conference (ECC)*, 1365–1370.

Schölkopf, B. and Smola, A.J. (2001). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. (Adaptive Computation and Machine Learning). MIT Press.

Söderström, T. and Stoica, P. (1989). *System Identification*. Prentice-Hall.

Volpe, V. (2015). *Identification of dynamical systems with finitely many data points*. Un. of Brescia, M. Sc. Thesis.

Wahba, G. (1983). Bayesian confidence intervals for the cross-validated smoothing spline. *Journal of the Royal Statistical Society. Series B (Methodological)*, 45(1), 133–150.

Wahba, G. (1990). *Spline models for observational data*. SIAM, Philadelphia.