Outlier robust system identification: a Bayesian kernel-based approach

Giulio Bottegal * Aleksandr Y. Aravkin **
Håkan Hjalmarsson * Gianluigi Pillonetto ***

* ACCESS Linnaeus Centre, School of Electrical Engineering, KTH
Royal Institute of Technology, Stockholm, Sweden
(e-mail: {bottegal; hjalmars}@kth.se)
** IBM T.J. Watson Research Center, Yorktown Heights, NY, USA
(e-mail: saravkin@us.ibm.com)
*** Department of Information Engineering, University of Padova,
Padova, Italy (e-mail: giapi@dei.unipd.it)

Abstract: In this paper, we propose an outlier-robust regularized kernel-based method for linear system identification. The unknown impulse response is modeled as a zero-mean Gaussian process whose covariance (kernel) is given by the recently proposed stable spline kernel, which encodes information on regularity and exponential stability. To build robustness to outliers, we model the measurement noise as realizations of independent Laplacian random variables. The identification problem is cast in a Bayesian framework, and solved by a new Markov Chain Monte Carlo (MCMC) scheme. In particular, exploiting the representation of the Laplacian random variables as scale mixtures of Gaussians, we design a Gibbs sampler which quickly converges to the target distribution. Numerical simulations show a substantial improvement in the accuracy of the estimates over state-of-the-art kernel-based methods.

1. INTRODUCTION

The classic approach to the problem of identifying a linear time-invariant system assumes that its transfer function belongs to a model class described by a small number of parameters that determine important properties, such as zeros and poles positions, time constant, etc. To identify the system, these parameters are estimated by minimizing a cost function related to the variance of the output prediction error. This procedure, called prediction error method (PEM), is motivated by the fact that, when the number of available data tends to infinity, the parameter estimates are consistent and their variance attains the Cramer-Rao bound [Ljung, 1999], [Söderström and Stoica, 1989]. This optimality result is guaranteed only when the "true" model lies in the chosen model class. Clearly, in many situations choosing the appropriate model class may be an issue, and one should rely on model selection criteria such as AIC [Akaike, 1974] or cross validation [Ljung, 1999]. However, these criteria are consistent only asymptotically and may tend to overestimate the model order or provide poor predictive capability [Pillonetto and De Nicolao, 2012].

Motivated by these issues, new identification paradigms have recently gained popularity. Rather than positing a model class described by a small number of parameters and then estimating these, newer methods try to estimate the entire impulse response. In order to overcome the ill-posedness of this problem, these methods estimate *hyper-parameters* in order to regularize the identification process. Hyperparameters can be seen as the counterpart of the parametric model order selection. Kernel-based regularization methods are an important example of this kind of approach, and have had a long history in regression problems [Poggio and Girosi, 1990], [Wahba, 1990]. In the system identification framework, kernel-based methods have been introduced recently [Pillonetto and De Nicolao, 2010], [Pillonetto et al., 2011], [Pillonetto et al., 2014]. The unknown impulse response is modeled as a realization of a Gaussian stochastic process, whose covariance matrix belongs to the class of the so-called *stable spline kernels* [Pillonetto and De Nicolao, 2011]. Introduced in [Pillonetto and De Nicolao, 2010], kernels of this type have been proven to effectively model the behavior of the impulse response of stable systems [Chen et al., 2012], exponential trends [Pillonetto et al., 2010] and correlation functions [Bottegal and Pillonetto, 2013].

In the kernel-based approach, the estimate of the impulse response is computed as the minimum variance Bayes estimate given the observed input/output data. Recall that when the output is corrupted by white Gaussian noise, the impulse response and the output are jointly Gaussian. However, if the white Gaussian noise assumption is violated, then the estimated impulse response may be poor. In particular, this approach fails in the presence of outliers [Aravkin et al., 2011]; see the example below.

A motivating example Suppose we want to estimate the impulse response of a linear system fed by white noise using the kernel-based method proposed in [Pillonetto and
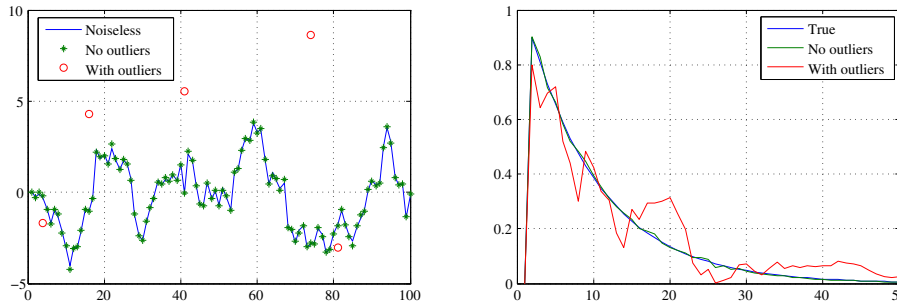
Fig. 1. *Introductory example. Left panel: the noiseless output and the measured outputs in the no-outliers situation (measurements shown using green asterisks) and when outliers are present (shown using red circles). Right panel: the true impulse response and its estimate in the no-outliers situation and when outliers are present.*

De Nicolao, 2010]. We consider two different situations, depicted in Figure 1. In the first one, 100 samples of the output signal are measured with a low-variance Gaussian additive noise; note that the estimated impulse response is very close to the truth. In the second situation we introduce 5 outliers in the measured output, obtaining a much poorer estimate of the same impulse response. This suggests that outliers may have a devastating effect on the standard identification process that relies on the assumption of Gaussianity.

Statement of contribution and organization of the paper
In this paper we introduce an outlier-robust system identification algorithm. We model the measurement noise as realizations of independent Laplacian random variables, which are better suited to modeling outliers because they have heavier tails than the Gaussian distribution. Then, using stable spline kernels, we set a proper prior to the impulse response of the system, which allows us to cast the problem into a Bayesian framework and to solve it using Markov Chain Monte Carlo (MCMC) approach [Andrieu et al., 2010]. Note that MCMC-based approaches are standard in system identification [Ninness and Henriksen, 2010], [Lindsten et al., 2012]. A fundamental point of this work is exploiting the representation of Laplacian random variables as scale mixtures of Gaussians, that is, Gaussian variables whose variance has a prior exponential distribution. This representation allows us to design a Gibbs sampler [Gilks et al., 1996], which does not require any rejection criterion of the generated samples and quickly converges to the target distribution. We evaluate the performance of the proposed algorithm using numerical simulations, and show that in the presence of outliers, there is a substantial improvement of the accuracy of the estimated impulse response compared to the kernel-based method proposed in [Pillonetto and De Nicolao, 2010].

The paper is organized as follows. In Section 2, we formulate our system identification problem. In Section 3 we cast this problem in a Bayesian framework. In Section 4, we describe the proposed algorithm for impulse response estimation, and test it using numerical simulations in Section 5. Some conclusions end the paper.

2. PROBLEM STATEMENT

We consider a SISO linear time-invariant discrete-time dynamic system (see Figure 2)

$$y(t) = G(z)u(t) + v(t), \qquad (1)$$

where $G(z)$ is a strictly causal transfer function representing the dynamics of the system, driven by the input $u(t)$. The measurements of the output $y(t)$ are corrupted by the process $v(t)$, which is zero-mean white noise with variance σ^2. In the typical system identification framework, the distribution of the noise samples is assumed to be Gaussian. Here, instead, we consider a Laplacian probability density for the noise, i.e.

$$p(v(t)) = \frac{1}{\sqrt{2}\sigma} e^{-\frac{\sqrt{2}|v(t)|}{\sigma}}. \qquad (2)$$

We assume that N samples of the input and output measurements are collected, and denote them by $u(1), \ldots, u(N), y(1), \ldots, y(N)$. Our system identification problem is to obtain an estimate of the impulse response $g(t)$ (or, equivalently, the transfer function) for n time instants, namely $\hat{g}(1), \ldots, \hat{g}(n)$. Recall that by choosing n sufficiently large, these samples can be used to approximate $g(t)$ with arbitrary accuracy [Ljung and Wahlberg, 1992].
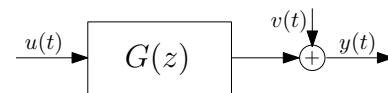

Fig. 2. *Block scheme of the system identification scenario.*

Introducing the vector notation

$$y := \begin{bmatrix} y(1) \\ \vdots \\ y(N) \end{bmatrix}, \ g := \begin{bmatrix} 0 \\ g(1) \\ \vdots \\ g(n) \end{bmatrix}, \ v := \begin{bmatrix} v(1) \\ \vdots \\ v(N) \end{bmatrix}$$

$$U = \begin{bmatrix} u(1) & 0 & & \ldots & 0 \\ u(2) & u(1) & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \ddots & \\ u(N) & u(N-1) & \ldots & u(1) & 0 \end{bmatrix} \in \mathbb{R}^{N \times (n+1)},$$

the input-output relation for the available samples can be written

$$y = Ug + v, \qquad (3)$$

so that our estimation problem can be cast as a linear regression problem.

3. A BAYESIAN FRAMEWORK

In this section we describe probabilistic models used for the quantities of interest in the problem.

3.1 The stable spline kernel

We first focus on setting a proper prior on g. Following a Gaussian regression approach [Rasmussen and Williams, 2006], we model g as a zero-mean Gaussian random vector, i.e.

$$p(g) \sim \mathcal{N}(0, \lambda K_\beta), \qquad (4)$$

where K_β is a covariance matrix whose structure depend on the value of the parameter β and $\lambda \geq 0$ is a scaling factor. In this context, K_β is usually called a *kernel* and determines the properties of the realizations of g. In this paper, we draw K_β from the class of the stable spline kernels [Pillonetto and De Nicolao, 2010], [Pillonetto et al., 2011]. In particular we shall make use of the so-called *first-order stable spline kernel* (or *TC kernel* in [Chen et al., 2012]), defined as

$$\{K_\beta\}_{i,j} := \beta^{\max(i,j)}, \ 0 \leq \beta < 1. \qquad (5)$$

Such a kernel is parameterized by β, which regulates the decaying velocity of the generated impulse responses. Then, once the hyperparameters are fixed, the probability distribution of g is

$$p(g|\lambda, \beta) = \frac{1}{\sqrt{(2\pi)^n \det(\lambda K_\beta)}} e^{-\frac{1}{2} g^T (\lambda K_\beta)^{-1} g}. \qquad (6)$$

Clearly, knowing the values of hyperparameters is of paramount importance to the design of an impulse response estimator. The following result, drawn from [Magni et al., 1998], shows the marginal distribution of the inverse of the hyperparameter λ given g and β.

Lemma 1. The posterior probability distribution of λ^{-1} given g and β is

$$p(\lambda^{-1}|g, \beta) \sim \Gamma\left(\frac{n}{2} + 1, g^T K_\beta^{-1} g\right) \qquad (7)$$

Remark 2. To obtain the result of the above Lemma, we have implicitly set an improper prior on λ with non-negative support.

3.2 Modeling noise as a scale mixture of Gaussians

The assumption on the noise distribution poses a challenge in expressing the conditional probability of g given the input-output data, since it is non-Gaussian. Here, we show how to deal with this problem. The key is to represent the noise samples $v(t)$ as a scale mixture of normals [Andrews and Mallows, 1974]. Specifically, denoting by v_i the i-th entry of the noise vector v, for $i = 1, \ldots, N$, the pdf of v_i can always be expressed as

$$p(v_i|\sigma^2) = \frac{1}{\sqrt{2}\sigma} e^{-\frac{\sqrt{2}|v_i|}{\sigma}} = \int_0^{+\infty} \frac{1}{\sqrt{2\pi\tau_i}} e^{-\frac{v_i^2}{2\tau_i}} \frac{1}{\sigma^2} e^{-\frac{\tau_i}{\sigma^2}} d\tau_i. \qquad (8)$$

The above expression highlights the fact that each noise sample can be thought of as a realization of a Gaussian random variable, whose variance τ_i is in turn the realization of an exponential random variable, i.e.

$$p(\tau_i|\sigma^2) = \frac{1}{\sigma^2} e^{-\frac{\tau_i}{\sigma^2}}, \ \tau_i \geq 0. \qquad (9)$$

Thus,

$$p(v_i|\tau_i, \sigma^2) = \frac{1}{\sqrt{2\pi\tau_i}} e^{-\frac{v_i^2}{2\tau_i}} \qquad (10)$$

The following result establishes a closed-form expression for the conditional probability density $p(\tau_i|v_i)$.

Lemma 3. For any $i = 1, \ldots, N$, the posterior of τ_i given v_i is

$$p(\tau_i|v_i, \sigma^2) \sim GIG\left(\frac{2}{\sigma^2}, v_i^2, \frac{1}{2}\right), \qquad (11)$$

i.e., a generalized inverse Gaussian random variable.

Using the above result, we have that the posterior probability density of τ_i given v_i, $i = 1, \ldots, N$, is available in closed-form. The probability density (11) also depends on σ^2. Instead of establishing a prior for such a parameter, a consistent estimate of its value can be obtained with the following steps:

(1) compute the least-squares estimate of g, i.e.

$$\hat{g}_{LS} = (U^T U)^{-1} U^T y, \qquad (12)$$

in order to obtain an unbiased estimate of g;

(2) compute the empirical estimate of σ^2

$$\hat{\sigma}^2 = \frac{(y - U\hat{g}_{LS})^T (y - U\hat{g}_{LS})}{N - n}. \qquad (13)$$

In the following section, we shall assume that σ^2 is known.

4. SYSTEM IDENTIFICATION UNDER GAUSSIAN AND LAPLACIAN NOISE ASSUMPTIONS

4.1 The Gaussian noise case

In this section, we make use of prior (4) for modeling g, assuming that the noise $v(t)$ is Gaussian. Then, the joint distribution of the vectors y and g, given values of λ, β and σ^2, is jointly Gaussian, namely

$$p\left(\begin{bmatrix} y \\ g \end{bmatrix} \middle| \lambda, \beta\right) \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Sigma_y & \Sigma_{yg} \\ \Sigma_{gy} & \lambda K_\beta \end{bmatrix}\right), \qquad (14)$$

where

$$\Sigma_y = \lambda U K_\beta U^T + \sigma^2 I_N \qquad (15)$$

and $\Sigma_{yg} = \Sigma_{gy}^T = \lambda U K_\beta$. In this case, the minimum mean square error (MSE) estimation of g is given by its Bayesian linear estimate, namely

$$\hat{E}[g|y, \lambda, \beta] = \Sigma_{gy} \Sigma_y^{-1} y. \qquad (16)$$

The above equation depends on unknown values of hyperparameters λ and β. The estimate of such parameters, denoted $\hat{\lambda}$ and $\hat{\beta}$, can be performed by exploiting the Bayesian framework of the problem. More precisely, since y and g are jointly Gaussian, we can obtain $\hat{\lambda}$ and $\hat{\beta}$ by maximizing the marginal likelihood, obtained by integrating out g from the joint probability density of (y, g). Then we have

$$(\hat{\lambda}, \hat{\beta}) = \arg\min_{\lambda, \beta} \log \det(\Sigma_y) + y^T \Sigma_y^{-1} y. \qquad (17)$$

In this paper, we always use this approach to estimate β. Hence, below we shall consider such parameter to be known.

4.2 The Laplacian noise case

We now consider the proposed model, where g has prior (4) and the noise is modeled using the Laplacian distribution. Then, the joint description of y and g given σ^2, λ and β does not admit a Gaussian distribution, since the vector y is itself not Gaussian distributed. However, as shown in Section 3.2, we can cast the problem in the Gaussian

regression framework by introducing variables τ_i, $i = 1, \ldots, N$. In fact, it can be seen that, redefining Σ_y as

$$\Sigma_y = \lambda U K_\beta U^T + D, \, D := \text{diag}\{\tau_1, \ldots, \tau_N\} \quad (18)$$

the joint posterior of y and g given λ, β, σ^2 and all τ_i is again Gaussian:

$$p\left(\begin{bmatrix} y \\ g \end{bmatrix} \middle| \lambda, \{\tau_i\}_{i=1}^N\right) \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Sigma_y & \Sigma_{yg} \\ \Sigma_{gy} & \lambda K_\beta \end{bmatrix}\right), \quad (19)$$

and the best estimator for g is given by

$$\hat{E}[g|y, \lambda, \{\tau_i\}_{i=1}^N] = \Sigma_{gy}\Sigma_y^{-1}y. \quad (20)$$

Unfortunately, the above estimator requires the knowledge of the values of the τ_i's. In principle these parameters could be estimated by adopting a marginal likelihood function analogous to (17). However, the resulting minimization problem is extremely complicated and ill-posed. The number of variables of the same order of the number of measurements and subject to multiple minima. Below, we describe our approach to solve the system identification problem.

The proposed MCMC scheme   The Bayesian solution to the problem here presented accounts also for uncertainty of λ and $\{\tau_i\}_{i=1}^N$. It returns the impulse response estimate as the following integral

$$\hat{g} = \int g\, p(g, \lambda, \{\tau_i\}_{i=1}^N|y)\, dg\, d\lambda \prod_{i=1}^N d\tau_i, \quad (21)$$

which can be computed by Monte Carlo integration. In particular, it is sufficient to draw a large number of samples from the distribution $p(g, \lambda, \{\tau_i\}_{i=1}^N|y)$ and compute the average value

$$\hat{g} = \lim_{M\to\infty} \frac{1}{M}\sum_{k=1}^M g^k, \quad (22)$$

where the g^k are realizations from the posterior. Drawing samples from a distribution is a hard problem in general. However, when all the conditional probability densities of such a distribution are available in closed-form, this can be done efficiently by employing a special case of the Metropolis Hastings sampler, namely the Gibbs sampler (see e.g. [Gilks et al., 1996]). The basic idea is that each conditional random variable is the state of a Markov chain; then, drawing samples from each conditional probability density iteratively, we converge to the stationary state of this Markov chain and generate samples of the conditional distribution of interest. In our case, in view of (21), we set $p(g, \lambda, \{\tau_i\}_{i=1}^N|y)$ as target probability density. Then, the conditional densities are as follows.

(1) $p(\tau_i|g, \lambda, \{\tau_j\}_{j=1,j\neq i}^N, y)$, $i = 1, \ldots, N$. Note that, for any $i = 1, \ldots, N$, τ_i is independent of λ, τ_j and y_j, $j \neq i$ and indeed it depends only on the observed value of the noise sample v_i. Then, recalling that $v_i = y_i - U_i g$, where U_i denotes the i-th row of U, this conditional density has the form (11), namely a generalized inverse Gaussian with parameters $\left(\frac{2}{\sigma^2}, (y_i - U_i g)^2, \frac{1}{2}\right)$.

(2) $p(\lambda^{-1}|g, \{\tau_i\}_{i=1}^N, y)$. Once g is given, λ becomes independent of all the other variables (see Lemma 1). Hence this conditional corresponds to the one stated in Lemma 1, namely a Gamma distribution with parameters $\left(\frac{n}{2} + 1, g^T K_\beta^{-1} g\right)$.

(3) $p(g|\lambda, \{\tau_i\}_{i=1}^N, y)$. This probability density can be easily derived from (19) and has a Gaussian distribution, with mean $\lambda K_\beta U^T \Sigma_y^{-1} y$ and covariance

$$\lambda K_\beta - \lambda^2 K_\beta U^T \Sigma_y^{-1} U K_\beta.$$

Having established the above conditional probabilities, we need to specify the initial values for g and λ, to be used as starting points in the iterative Gibbs sampler. These are obtained by exploiting the estimation procedure proposed in Section 4.1 for the Gaussian noise case.

We now give our system identification algorithm.

Algorithm: Outlier robust system identification
Input: $\{y(t)\}_{t=1}^N$, $\{u(t)\}_{t=1}^N$
Output: $\{\hat{g}\}_{t=1}^n$
(1) Initialization:
    (a) Estimate σ^2 from (13) and β from (17)
    (b) Obtain g^0 from (16) and λ^0 from (17)
(2) For $k = 1$ to M:
    (a) Draw the sample τ_i^k, $i = 1, \ldots, N$ from
$$p(\tau_i|g^{k-1}, \lambda^{k-1}, \{\tau_j^{k-1}\}_{j=1,j\neq i}^N, y)$$
    (b) Draw the sample λ^k from
$$p(\lambda^{-1}|\{\tau_i^k\}_{i=1}^N, g^{k-1}, y)$$
    (c) Draw the sample g^k from
$$p(g|\lambda^k, \{\tau_i^k\}_{i=1}^N, y)$$
(3) Compute $\hat{g} = \frac{1}{M-M_0}\sum_{k=M_0}^M g^k$

In the above algorithm, the parameters M and M_0 are introduced. M the number of samples to be generated; clearly, large values of M should guarantee more accurate estimates of g. M_0 is the number of initial samples drawn from the conditional of g to be discarded. In fact, the conditionals from which those samples are drawn are to be considered as non-stationary, since the Gibbs sampler takes a certain number of iterations to get close to a stationary distribution.

Remark 4. The estimation procedure of β is in a certain sense "non-optimal", since it is based on a different noise model. However, we observe that the sensitivity of the estimator to the value of β is relatively low, in the sense that a large interval of values of β can model a given realization of g efficiently (see Lemma 2 in [Bottegal and Pillonetto, 2013]). Models for β will be introduced in future works.

A block scheme representation of the proposed identification algorithm is shown in Figure 3. From this scheme, it is clear that this algorithm can be seen as a refinement of the algorithm proposed in [Pillonetto et al., 2010] and briefly described in Section 4.1.

5. NUMERICAL EXPERIMENTS

In this section, we report numerical results to illustrate the performance of the proposed algorithm. We evaluate the proposed algorithm by means of 4 Monte Carlo experiments of 100 runs each. At each run, a linear system is randomly generated such that its transfer function $G(z)$ has 30 zeros and 30 poles. These poles are always within
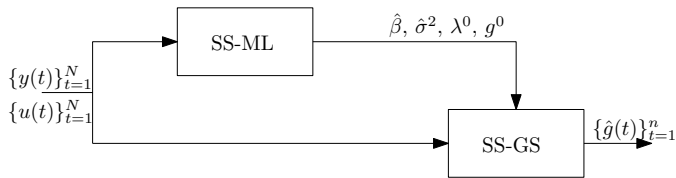
Fig. 3. *Block scheme of the proposed algorithm. The label* SS-ML *represents the marginal likelihood-based system identification method reviewed in Section 4.1. The label* SS-GS *indicates the Gibbs sampler step of the proposed method.*

the circle with center at the origin and radius 0.95 on the complex plane. We consider an input-output delay equal to 1. In order to simulate the presence of outliers in the measurement process, the noise samples $v(t)$ are drawn from a mixture two Gaussian of the form

$$v(t) \sim c_1 \mathcal{N}(0, \sigma^2) + c_2 \mathcal{N}(0, 100\sigma^2), \quad (23)$$

with $c_1 = 0.7$ and $c_2 = 0.3$, so that outliers (observations with 100 times higher variance) are generated with probability 0.3. The value of σ^2 was set to the variance of the noiseless output divided by 100. Two different types of input signals are considered: 1) $u(t)$ is obtained by filtering a white noise sequence through a second-order low pass filter with random bandwidth (labeled as LP); 2) $u(t)$ is white noise (labeled as WN).

At each Monte Carlo run, N samples of the input and output signals are generated; we consider two different situations where the number of available samples is either $N = 200$ or $N = 500$. In all the experiments, we set $n = 50$. Thus, there is a total of 4 different Monte Carlo experiments whose features are summarized in Table 1.

Exp.#	Data set size (N)	Input type
1	200	LP
2	500	LP
3	200	WN
4	500	WN

Table 1: *Features of the 4 Monte Carlo experiments.*

Two different algorithms are tested; their performances are evaluated at any run by computing the fitting score, i.e.

$$FIT_i(\%) = 100 \left(1 - \frac{\|g_i - \hat{g}_i\|_2}{\|g_i\|_2} \right), \quad (24)$$

where g_i and $\hat{g}_i$ represent, respectively, the true and the estimated impulse responses (truncated at the n-th sample) obtained at the i-th run. The estimators tested are specified below.

SS-ML: This is the nonparametric kernel-based identification method proposed in [Pillonetto et al., 2010] and briefly described in Section 4.1. The impulse response is modeled as in (4) and the hyperparameters λ and β are estimated by using a marginal likelihood maximization approach. Note that this estimator does not attempt to model the presence of outliers.

SS-GS: This is the approach proposed in this paper, where a Gibbs sampler is employed for computing (21). The parameter M, denoting the number of samples generated from each conditional probability density, is set to 1500. The first $M_0 = 500$ generated samples are discarded. The validity of the choice of M and M_0 is checked by assessing that quantiles 0.25, 0.5, 0.75 are estimated with good

precision [Raftery and Lewis, 1996]. The initial values of g and λ and the estimated values of β and σ^2 are drawn from the *SS-ML* Algorithm.

Figure 4 shows the box plots of the 100 reconstruction errors obtained by each estimator after the 4 Monte Carlo experiments. The proposed method offers a substantial improvement of the fitting score in the example scenario. This is particularly visible in the case of white noise, where the fitting score is above 90%. When the input is a low-pass signal, one can see that sometimes the performance of the estimators are not so satisfactory. This happens when a high-pass transfer function is fed with a short-band input, a combination that is known to give rise to ill-posed problems [Bertero, 1989].

5.1 An example with no outliers

In order to complete our analysis, we also test our algorithm in the same framework as above, but setting $c_1 = 1$ and $c_2 = 0$, that is, generating errors from a Gaussian noise model with no outliers. We use $N = 500$, and generate inputs by filtering white noise through random second order low-pass filters. The boxplots of Figure 5 show the comparison between SS-ML and SS-GS over 100 Monte Carlo runs. The performance of the proposed algorithm is comparable to the performance of the SS-ML Algorithm, with a slight degradation in the fitting score due to the modeling of the noise, which, in the proposed estimator, is Laplacian instead of Gaussian.

6. CONCLUSIONS

In this paper, we have proposed a novel identification scheme for estimating impulse responses of linear system when the measurements are corrupted by outliers. We have shown that, modeling the measurement noise as Laplacian random variables, we can model our problem using a mixture of Gaussian random variables. The mixture coefficients can be estimated by adopting a MCMC scheme which exploits closed-form expressions of conditional probabilities for the parameters of interest. The performance of the proposed algorithm gives a substantial improvement over the state-of-the-art algorithm, which does not use outlier-robust noise modeling.
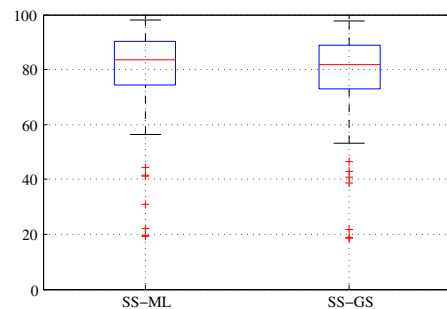

Fig. 5. *Box plot of the fitting score when no outliers are simulated.*

REFERENCES

H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19: 716–723, 1974.
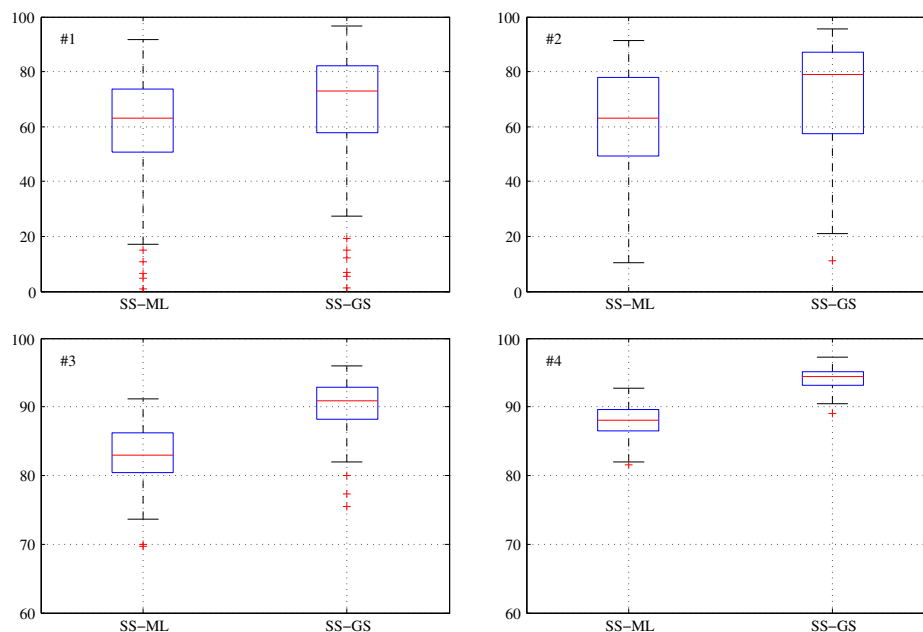
Fig. 4. *Box plots of the fitting scores when measurements are corrupted by outliers. The description of the experiments is summarized in Table 1.*

D. F. Andrews and C. L. Mallows. Scale mixtures of normal distributions. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 99–102, 1974.

C. Andrieu, A. Doucet, and R. Holenstein. Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3):269–342, 2010.

A.Y. Aravkin, B.M. Bell, J.V. Burke, and G. Pillonetto. An ℓ_1-Laplace robust kalman smoother. *Automatic Control, IEEE Transactions on*, 56(12):2898–2911, 2011.

M. Bertero. Linear inverse and ill-posed problems. *Advances in Electronics and Electron Physics*, 75:1–120, 1989.

G. Bottegal and G. Pillonetto. Regularized spectrum estimation using stable spline kernels. *Automatica*, 49 (11):3199–3209, 2013.

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

W.R. Gilks, S. Richardson, and D.J. Spiegelhalter. *Markov chain Monte Carlo in Practice*. London: Chapman and Hall, 1996.

F. Lindsten, T. B. Schön, and M. Jordan I. A semiparametric bayesian approach to wiener system identification. In *Proceedings of the 16th IFAC Symposium on System Identification, Brussels, Belgium*, 2012.

L. Ljung. *System Identification, Theory for the User*. Prentice Hall, 1999.

L. Ljung and B. Wahlberg. Asymptotic properties of the least-squares method for estimating transfer functions and disturbance spectra. *Advances in Applied Probability*, pages 412–440, 1992.

P. Magni, R. Bellazzi, and G. De Nicolao. Bayesian function learning using MCMC methods. *IEEE Transactions on Pattern Analysis Machine Intelligence*, 20 (12):1319–1331, 1998.

B. Ninness and S. Henriksen. Bayesian system identification via Markov chain Monte Carlo techniques. *Automatica*, 46(1):40–51, 2010.

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

G. Pillonetto and G. De Nicolao. Kernel selection in linear system identification – part I: A Gaussian process perspective. In *Proceedings of CDC-ECC*, 2011.

G. Pillonetto and G. De Nicolao. Pitfalls of the parametric approaches exploiting cross-validation or model order selection. In *Proceedings of the 16th IFAC Symposium on System Identification (SysId 2012)*, 2012.

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.

G. Pillonetto, A. Chiuso, and G. De Nicolao. Prediction error identification of linear systems: a nonparametric Gaussian regression approach. *Automatica*, 47(2):291–305, 2011.

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

T. Poggio and F. Girosi. Networks for approximation and learning. In *Proceedings of the IEEE*, volume 78, pages 1481–1497, 1990.

A. E. Raftery and S. M. Lewis. The number of iterations, convergence diagnostics and generic Metropolis algorithms. In *Markov Chain Monte Carlo in practice*. Chapman & Hall, 1996.

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

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

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