A fully Bayesian approach to kernel-based regularization for impulse response estimation ⋆

Rodrigo A. González * **and Cristian R. Rojas** *

* *Department of Automatic Control, KTH Royal Institute of Technology, 10044 Stockholm, Sweden (e-mails: grodrigo@kth.se, crro@kth.se).*

Abstract: Kernel-based regularization has recently been shown to be a successful method for impulse response estimation. This technique usually requires choosing a vector of hyperparameters in order to form an appropriate regularization matrix. In this paper, we develop an alternative way to obtain kernel-based regularization estimates by Bayesian model mixing. This new approach is tested against state-of-the-art methods for hyperparameter tuning in regularized FIR estimation, with favorable results in many cases.

Keywords: Linear system identification; kernel-based regularization; Bayesian estimation; model mixing.

1. INTRODUCTION

The least squares method for parametric estimation has played a central role in linear system identification for many decades. This simple estimator was thought to be completely understood by the system identification community, but surprisingly, new approaches and improvements upon this method have continued to arise in the recent years.

One of the most interesting ideas recently explored is kernel-based regularization. Kernel-based regularization has roots in Machine Learning (Wahba et al. (1999); Cucker and Smale (2002)), and was introduced for system identification by Pillonetto and De Nicolao (2010). Afterwards, many contributions and applications have followed (see, e.g., Chen et al. (2012)). The main goal behind kernel-based regularization is to find an appropriate regularization matrix such that the regularized least squares problem gives result to a vector parameter estimate that has a smaller mean square error (MSE) than parametric maximum likelihood/prediction error methods. This method has been proven successful in extensive simulations and real data problems in Pillonetto et al. (2014), with goodness of fit often better than the standard prediction error method.

A key insight in kernel-based regularization is that it is convenient to parametrize the regularization matrix by a low-dimensional vector of hyperparameters, which is usually estimated by input-output data. The state-of-the-art methods for estimating the hyperparameter vector are the empirical Bayes method, used in Chen et al. (2012) and the Stein's unbiased risk estimation (SURE) method, as in Pillonetto and Chiuso (2015). Empirical Bayes estimates the hyperparameters by maximizing the marginal likelihood (ML) of the data given the hyperparameter vector, while the SURE approach selects the hyperparameter vector that minimizes an unbiased estimate of the MSE. In both cases, it is not taken in account that improved estimates can possibly be obtained by mixing regularized estimators, instead of choosing a specific hyperparameter vector for the identification procedure.

Model mixing has been studied extensively in many different fields of science. In system identification, a Bayesian framework was employed in Hjalmarsson and Gustafsson (1995) to form a joint frequency response function and estimation error based on estimates from a set of possible model structures. In econometrics, Hansen (2007) develops an asymptotically optimal model weighting procedure for least squares regressions, where optimality was defined in the sense of achieving the lower possible square error among a class of estimators. In statistics, Yang (2000) showed that it is possible to combine a countable collection of estimators in order to construct a minimax-rate optimal adaptive estimator. These contributions show that an averaged model is a preferable solution to the identification problem when there exists uncertainty in model selection.

In this paper we provide an alternative approach to hyperparameter selection in kernel-based regularization methods. We present an estimator that approximates the conditional mean (over the hyperparameters) of the parameter vector given the input-output data, and show that this method is deeply related to model mixing. Numerical examples over benchmarks and data sets show that the proposed estimator compares favorably against the empirical Bayes approach. Thanks to an anonymous reviewer, we have been informed that similar work was reported in Prando et al. (2016), where a different approximation scheme has been put forward, and a flat prior was proposed.

⋆ This work was supported by the Swedish Research Council under contract number 2016-06079 (NewLEADS).

The paper is organized as follows. In Section 2 we formulate the problem. In Section 3 we review regularization techniques and kernel methods, while in Section 4 we study hyperparameter estimation and introduce the proposed full Bayesian method for regularization. Simulations of the novel method are found in Section 5, and Section 6 concludes this paper.

2. PROBLEM FORMULATION

Consider a linear time-invariant causal single input single output stable system

$$y_t = G_0(q)u_t + v_t.$$

Here, q is the shift forward operator, $qh_t = h_{t+1}$, $G_0(q)$ is the transfer function relating input $\{u_t\}$ and output $\{y_t\}$ sequences, and $\{v_t\}$ is a zero-mean additive noise sequence independent of the input. The standard system identification problem is to obtain a correct model for $G_0(q)$, or equivalently, a good estimation of the impulse response of the LTI system. For this purpose, we usually select a particular parametrization of $G_0(q)$ so that

$$y_t = G(\theta, q)u_t + v_t,$$

where the parametrized model of the transfer function is

$$G(\theta, q) = \sum_{k=1}^{\infty} g_k(\theta)q^{-k}. \tag{1}$$

The N data pairs collected from the system are denoted as $\{(u_t, y_t)\}_{t=1}^N$, where $\{u_t\}_{t=1}^N$ is assumed known and $\{y_t\}_{t=1}^N$ are measured output data. Given this setup, the goal is to obtain the finite dimensional parameter vector θ so that $G(\theta, q)$ is close to $G_0(q)$ in some specified sense.

An important problem in system identification is to determine the model parametrization or model order. One way to proceed is to estimate $G_0(q)$ from an FIR parametrized model

$$G(\theta, q) = \sum_{k=1}^{n} g_k(\theta)q^{-k}, \quad g_k(\theta) = \theta_k, k = 1, \ldots, n. \tag{2}$$

In this paper, we assume that n can be chosen large enough to make differences between (2) and the true system negligible, provided θ is chosen correctly.

It has been observed in Pillonetto and De Nicolao (2010) and in Chen et al. (2012), that using kernel regularization techniques over high order FIR models can give rise to robust and suitable impulse response estimations. These estimators intend to minimize the mean square error (MSE)

$$\mathrm{MSE}(\hat{\theta}) = E\left[\|\hat{\theta} - \theta_0\|_2^2\right], \tag{3}$$

where θ_0 is the real parameter vector of the system $G_0(q)$.

The underlying strategy of the recently introduced kernel-based estimators is to parametrize the regularization matrix. The parameters from which the regularization matrix depends on are commonly called *hyperparameters*, and they are estimated from the data. The problem discussed in this paper is how to improve over these regularization techniques of parameter estimation under a pure Bayesian framework.

In the following section we introduce the key concepts related to least squares, regularization and kernel methods.

3. LEAST SQUARES, REGULARIZATIONS AND KERNEL METHODS

There are many ways to estimate the parameter vector θ of the FIR model (2). It is well known (see, e. g., Ljung (1999)) that the problem of estimating this model can be written as the linear regression

$$Y_N = \Phi_N\theta + V_N,$$

where [1]

$$Y_N = [y_{n+1} \ \ldots \ y_N]^T,$$
$$V_N = [v_{n+1} \ \ldots \ v_N]^T,$$
$$\Phi_N = \begin{bmatrix} u_n & u_{n-1} & \ldots & u_1 \\ u_{n+1} & u_n & \ldots & u_2 \\ \vdots & \vdots & \ddots & \vdots \\ u_{N-1} & u_{N-2} & \ldots & u_{N-n} \end{bmatrix}.$$

If the objective is to minimize the loss function given by the 2-norm of the vector of prediction errors, we obtain the popular least squares (LS) estimate:

$$\begin{aligned}\hat{\theta}_{\mathrm{LS}} &= \arg\min_{\theta} \|Y_N - \Phi_N\theta\|_2^2 \\ &= (\Phi_N^T\Phi_N)^{-1}\Phi_N^T Y_N.\end{aligned}$$

The MSE of this estimator is given by

$$\mathrm{MSE}(\hat{\theta}_{\mathrm{LS}}) = \mathrm{Trace}(\mathrm{cov}(\hat{\theta}_{\mathrm{LS}})) = \sigma^2(\Phi_N^T\Phi_N)^{-1}.$$

Least squares is known to have excellent statistical properties. One main quality is that least squares is the maximum likelihood estimator, provided that the additive noise is Gaussian and white (Ljung, 1999). If the model structure contains the true system, this implies that least squares is asymptotically efficient, and of course unbiased. Although unbiasedness is generally a desirable property, in many occasions the user may tolerate some bias in the model, provided that the variance is reduced in a greater amount in order to minimize the MSE.

Therefore, if the design objective is to minimize the MSE (3) of the impulse response, we should search for estimators that purposely add some bias in order to reduce variance. To pursue this idea, we introduce regularization.

3.1 Regularization

The regularized LS estimate of θ, denoted by $\hat{\theta}_{\mathrm{RLS}}$, is computed by

$$\begin{aligned}\hat{\theta}_{\mathrm{RLS}} &= \arg\min_{\theta} \|Y_N - \Phi_N\theta\|_2^2 + \gamma\theta^T P^{-1}\theta \\ &= (\Phi_N^T\Phi_N + \gamma P^{-1})^{-1}\Phi_N^T Y_N\end{aligned} \tag{4}$$

If we define $R := \Phi_N^T\Phi_N + \gamma P^{-1}$, this estimator has an MSE given by

$$\mathrm{MSE}(\hat{\theta}_{\mathrm{RLS}}) = R^{-1}(\sigma^2\Phi_N^T\Phi_N + \gamma^2 P^{-1}\theta_0\theta_0^T P^{-1})R^{-1}.$$

Here γ is a regularization parameter that weighs the balance between fit to experimental data and shrinking for MSE improvement, and P is the regularization matrix, which gives shrinking penalization factors to each element θ_k and the relationship between them. Apart from positive-definiteness, P and γ can be chosen freely.

[1] Since the inputs before initial time are unknown, the first n output measurements are not reliable. Therefore, we do not consider this data for the regression.

Roughly speaking, as γ increases, variance reduces but the bias increases, shrinking the vector parameter towards the zero vector.

It can be proven that there exists a regularization parameter γ that leads to an estimator with smaller MSE than least squares. It is known by Chen et al. (2012) that, if the model describes the true system perfectly with $\theta = \theta_0$, the regularization matrix and parameter (P and γ) that minimize the MSE satisfy the equation $\gamma^{-1}P = \sigma^{-2}\theta_0\theta_0^T$, where σ^2 is the noise variance (which is generally not known). For this optimal case, the decrease in MSE compared to least squares is

$$\min_{P\geq 0,\gamma\geq 0} \mathrm{MSE}(\hat{\theta}_{\mathrm{LS}}) - \mathrm{MSE}(\hat{\theta}_{\mathrm{RLS}}) = \sigma^4(\Phi_N^T\Phi_N\theta_0\theta_0^T\Phi_N^T\Phi_N + \sigma^2\Phi_N^T\Phi_N)^{-1},$$

which suggests that the MSE will be improved more substantially when the noise variance is high, when not much data is used, or when the true vector parameter has small norm.

The practical downside of the insight given by the result in Chen et al. (2012) is that the optimal regularization constants depend on precisely what we want to estimate.

3.2 Kernel methods

The previous result leads to the natural question of how to choose P and γ on an identification experiment. A common strategy for obtaining an appropriate regularization matrix is to make it depend on hyperparameters β. The hyperparameters (γ, β) are estimated based on the data, such that the regularized least squares estimator (4) achieves close to optimal performance in terms of MSE.

The hyperparameters β impose a structure on P that is supposed reasonable for describing the optimal regularization matrix. The matrices obtained by these structures are commonly called kernels. Gaussian processes and deterministic perspectives have been used to motivate the appropriate selection of the kernel (see, e.g., Pillonetto and De Nicolao (2011) or Chen et al. (2011)). Further developments have been shown in Dinuzzo (2015).

A complete description of the different kernels used for regularization can be found in Pillonetto et al. (2014). The most commonly used kernels and their admissible set Γ of hyperparameters are stated next. If we assume that the system's impulse response is exponentially stable, we can describe the regularization matrix appropriately by the Diagonal/Correlated (DC) kernel:

$$\text{DC}\quad P_{ij}(\beta) = \lambda\alpha^{(i+j)/2}\rho^{|i-j|};$$
$$\beta = [\lambda, \alpha, \rho];\quad \Gamma = \{\lambda > 0, 0 \leq \alpha < 1, |\rho| \leq 1\}.$$

A special case of the DC kernel is when $\rho = \sqrt{\alpha}$, leading to the Tuned/Correlated (TC) kernel:

$$\text{TC}\quad P_{ij}(\beta) = \lambda\alpha^{\max(i,j)};$$
$$\beta = [\lambda, \alpha];\quad \Gamma = \{\lambda > 0, 0 \leq \alpha < 1\}.$$

Finally, an important kernel deduced in Pillonetto and De Nicolao (2010) is the Stable Spline (SS) kernel:

$$\text{SS}\quad P_{ij}(\beta) = \lambda\left(\frac{\alpha^{i+j+\max(i,j)}}{2} - \frac{\alpha^{3\max(i,j)}}{6}\right); \tag{5}$$
$$\beta = [\lambda, \alpha];\quad \Gamma = \{\lambda > 0, 0 \leq \alpha < 1\}.$$

4. A FULLY BAYESIAN PERSPECTIVE ON REGULARIZATION

In this section we explore the main ideas of this paper. The goal is to formulate an estimator of the structure (2) that minimizes the MSE of the model parameters θ by applying Bayesian techniques to obtain a parameter estimate that is an appropriate mix of an arbitrary number of regularized least square estimates.

We review the current techniques for estimating β, and then introduce the proposed method of impulse response estimation by model averaging.

4.1 Empirical Bayes method

If we fix $\gamma = \sigma^2$, in Chen et al. (2012) it has been shown that the regularization matrix P has a Bayesian interpretation. Let the vector θ be a random variable, with Gaussian distribution $\mathcal{N}(0, P(\beta))$. If the additive noise is also Gaussian and white, independent of θ, with zero mean and variance σ^2, the random vectors Y_N and θ will be jointly Gaussian random variables:

$$\begin{bmatrix}\theta \\ Y_N\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}P(\beta) & P(\beta)\Phi_N^T \\ \Phi_N P(\beta) & \Phi_N P(\beta)\Phi_N^T + \sigma^2 I_N\end{bmatrix}\right).$$

This means that, given β, the distribution of Y_N is

$$p(Y_N|\beta) = \frac{1}{\sqrt{\det(2\pi Z(\beta))}} e^{-\frac{1}{2}Y_N^T Z(\beta)^{-1}Y_N},$$

where $Z(\beta) = \Phi_N P(\beta)\Phi_N^T + \sigma^2 I_N$.

It is known by Pillonetto et al. (2014) that the regularized estimate $\hat{\theta}$ is the mean of the posterior distribution of θ given Y_N, provided we choose $\gamma = \sigma^2$ and the prior mean and covariance of θ as 0 and $E[\theta_0\theta_0^T]$ respectively. Then, a standard approach for estimating β is to maximize the marginalized log-likelihood

$$\begin{aligned}\hat{\beta}_{\mathrm{ML}} &= \arg\max_{\beta\in\Gamma} \log(p(Y_N|\beta)) \\ &= \arg\min_{\beta\in\Gamma} \frac{1}{2}(Y_N^T Z(\beta)^{-1} Y_N + \log\det(Z(\beta))).\end{aligned} \tag{6}$$

In the following subsections, we denote by $h(\beta)$ the cost function in (6).

4.2 SURE approach

As mentioned before, one identification objective in vector parameter estimation is to obtain an estimator with the lowest MSE possible. Since the MSE is a quantity that depends on the true parameters, it is not possible to minimize directly. The SURE method intends to obtain an unbiased estimate of an MSE of choice, and select the hyperparameter vector which minimizes this risk quantity.

In Pillonetto and Chiuso (2015), SURE estimates of the impulse response and of the output prediction have been put forward for the SS kernel respectively as follows:

$$\begin{aligned}\hat{\mathrm{MSE}}_\theta(\hat{G}) &= \|(\Phi_N^T\Phi_N)^{-1}\Phi_N^T Y_N - \hat{\theta}\|^2 \\ &\quad + \sigma^2\mathrm{Tr}(2R^{-1} - (\Phi_N^T\Phi_N)^{-1}), \\ \hat{\mathrm{MSE}}_y(\hat{G}) &= \|Y_N - \Phi_N\hat{\theta}\|^2 + \sigma^2\mathrm{Tr}(2\Phi_N P\Phi_N^T Z^{-1}),\end{aligned}$$

where R and Z are defined in Sections 3.1 and 4.1 respectively, with the dependence on β excluded only for simplicity.

In the work of Pillonetto and Chiuso (2015), the performance of the empirical Bayes tuning procedure is shown superior to that of the SURE techniques and thus, we will focus only on empirical Bayes comparisons.

4.3 Fully Bayesian method

Instead of choosing an appropriate hyperparameter vector β, our approach is the following. We seek an estimate $\hat{\theta}_{\mathrm{B}}$ that is generated by the conditional mean of the RLS estimate $\hat{\theta}_{\mathrm{RLS}}(\beta)$ given the measurements Y_N:

$$\hat{\theta}_{\mathrm{B}} = \mathrm{E}_{\beta}[\hat{\theta}_{\mathrm{RLS}}(\beta)|Y_N] = \int_{\Gamma} \hat{\theta}_{\mathrm{RLS}}(\beta) p(\beta|Y_N) d\beta. \tag{7}$$

Using Bayes' theorem,

$$p(\beta|Y_N) = \frac{p(Y_N|\beta)p(\beta)}{p(Y_N)},$$

and then expanding $p(Y_N)$, we can write (7) as

$$\hat{\theta}_{\mathrm{B}} = \frac{\int_{\Gamma} \hat{\theta}_{\mathrm{RLS}}(\beta) p(\beta) e^{-h(\beta)} d\beta}{\int_{\Gamma} p(\beta) e^{-h(\beta)} d\beta}. \tag{8}$$

The proposed method has a direct relation with model mixing. If we set

$$w(\beta) = \frac{p(\beta)e^{-h(\beta)}}{\int_{\Gamma} p(\xi)e^{-h(\xi)}d\xi},$$

what the full Bayesian method proposes is to mix an arbitrary number of models with a normalized weighting function given by $w(\beta)$. This weighting function measures the relative fitness of the hyperparameter vector to the data set, and can be adjusted in practice by changing the prior distribution $p(\beta)$, which is unknown and should be provided by the user (even though it will be empirically shown in Section 5 that the effect of this hyper-prior is asymptotically negligible). In other words, (8) corresponds to weighting the estimators $\hat{\theta}_{RLS}(\beta)$, for different β's, by a weighting factor $w(\beta)$. So instead of selecting a specific β, (8) suggests to *combine* or *mix* these estimators.

The problem now is how to compute (8) in a suitable way. Inspiring ideas regarding Bayesian perspectives on model averaging can be found in Madigan et al. (1999), where implementation of Bayesian model averaging is discussed in detail. Also, Bottegal et al. (2014) and Pillonetto (2016) have explored Markov Chain Monte Carlo methods for kernel-based system identification. For the computation of the integrals we will use Monte Carlo integration with importance sampling (Mackay (1998)). Denote by $q(\beta)$ a proposal distribution. We can estimate the quotient of integrals in (8) by

$$\hat{\theta}_{\mathrm{B}} = \frac{\mathrm{E}_q\left[\dfrac{\hat{\theta}_{\mathrm{RLS}}(\beta)e^{-h(\beta)}p(\beta)}{q(\beta)}\right]}{\mathrm{E}_q\left[\dfrac{e^{-h(\beta)}p(\beta)}{q(\beta)}\right]} \approx \frac{\sum_{i=1}^{M} \dfrac{\hat{\theta}_{\mathrm{RLS}}(\beta_i)e^{-h(\beta_i)}p(\beta_i)}{q(\beta_i)}}{\sum_{i=1}^{M} \dfrac{e^{-h(\beta_i)}p(\beta_i)}{q(\beta_i)}}, \tag{9}$$

where β_i are samples from the distribution $q(\beta)$, and M is the number of samples. What is left to know is how to choose the distribution q. We will approximate $p(\beta|Y)$ by a Gaussian distribution centered around the peak of $h(\beta)$, as in a Laplace approximation procedure (De Bruijn, 1970). Thus, a reasonable sampling distribution can be given by the truncated Gaussian

$$q(\beta) = \begin{cases} \eta e^{-\frac{1}{2}(\beta-\hat{\beta}_{\mathrm{ML}})^T\Sigma^{-1}(\beta-\hat{\beta}_{\mathrm{ML}})}, & \beta \in \Gamma \\ 0, & \beta \notin \Gamma, \end{cases} \tag{10}$$

where Σ is the inverse of the Hessian matrix of $h(\beta)$ at $\beta = \hat{\beta}_{\mathrm{ML}}$, and η is a normalizing constant. This distribution intends to capture the most relevant data in the high-peaked integrals in (8) so that less variance is incorporated in the Monte Carlo integration procedure.

Remark 1. Note that for all of the methods explained in this section, the noise variance σ^2 needs to be estimated if it is not known. There are several ways to estimate this quantity. As suggested in Goodwin et al. (1992) and Pillonetto et al. (2014), a standard way to proceed is to calculate the sample variance of a high model order FIR model, and use this value as a low biased estimate of σ^2. Another possibility in the empirical Bayes approach is to include this unknown quantity as an extra hyperparameter in β, and obtaining its value via empirical Bayes. The approach followed in this paper is the former one.

Remark 2. The distribution (10) is one possible option for sampling. If the Hessian matrix of $h(\beta)$ is not directly obtainable or poorly conditioned, other distributions can be used, as long as the distribution has support Γ.

5. SIMULATIONS

In this section, the proposed model mixing estimator will be tested against the standard kernel-based regularization estimator. In these simulations we have used the algorithm proposed in Chen and Ljung (2013) for tuning the hyperparameters of the empirical Bayes method.

5.1 Benchmark systems Monte Carlo simulations

We consider 4 discrete-time transfer functions taken from Pillonetto and De Nicolao (2010). These are expressed as

$$G_1(q) = \frac{0.0355q^2 + 0.02465q}{q^3 - 1.273q^2 + 0.333},$$
$$G_2(q) = \frac{0.36q}{5(q^2 + 0.24q + 0.36)},$$
$$G_3(q) = \frac{0.01q^3 + 0.0074q^2 + 0.000924q - 0.000017642}{q^4 - 2.14q^3 + 1.5549q^2 - 0.4387q + 0.042025},$$
$$G_4(q) = \frac{q^3 + 0.5q^2}{q^4 - 2.2q^3 + 2.42q^2 - 1.87q + 0.7225}.$$

The system input for every example is unit variance Gaussian white noise. The measurement noise is Gaussian and white, with standard deviation set to 10% of the maximum noiseless output value. The goal is to estimate the first 35 impulse response samples, with $N = 100$ measurements of input-output data pairs.

For every transfer function, we perform 500 Monte Carlo simulations with varying input and noise sequences. Later we obtain the regularized least square estimator empirical Bayes hyperparameter estimation using the DC, TC and SS kernels, and compare them with the impulse response

estimations by the proposed fully Bayesian approach, using the same kernels. We have set the number of sampling points of the fully Bayesian method to 1000 for each Monte Carlo iteration, and have used a flat prior on the parameters in β for all estimates. For completeness, the standard least squares estimator was also tested.

In order to test estimation accuracy, we compute the goodness of fit of each estimator with the `compare` command of the System Identification Toolbox in MATLAB, which is defined as

$$W = 100\left(1 - \left[\frac{\sum_{k=1}^{N}|y_k^0 - \hat{y}_k|^2}{\sum_{k=1}^{N}|y_k^0 - \bar{y}^0|^2}\right]^{1/2}\right), \quad \bar{y}^0 = \frac{1}{N}\sum_{k=1}^{N} y_k^0,$$

where $\{y_k^0\}$ is the output sequence of the (noise-less) true system, which in this case was obtained by exciting the system with Gaussian white noise of unit variance as input, and $\{\hat{y}_k\}$ is the output sequence of the model under the same input excitation.

Table 1 shows that all kernel-based methods outperform LS. As hinted before, the full Bayesian method can lead to better estimates than the other kernel-based methods based on empirical Bayes for some of the systems, such as G_2 and G_4. Also, the system's dynamics have an influence on the performance of this estimator compared to the empirical Bayes method, as we will see again in the next subsection.

Table 1. Average fit of the LS, the DC, TC and SS kernel-based regularization estimators and the full Bayesian counterparts.

Estimator\MC study	G_1	G_2	G_3	G_4
LS	82.997	79.708	83.148	80.629
DC	91.893	92.446	92.906	87.346
DC-Bayes	91.851	92.568	92.674	87.356
TC	91.793	92.034	91.963	87.512
TC-Bayes	91.744	92.068	91.947	87.522
SS	94.005	91.004	94.731	87.173
SS-Bayes	94.010	92.260	94.704	87.207

5.2 Data bank Monte Carlo simulations

Now, we focus on a more general set of systems. We consider the data bank of Chen et al. (2012), which is a set of high order random stable systems and input-output data that should be representative of real-life systems.

In this collection of data, the systems were split in 500 "fast systems" S1, which have all their poles inside a circle with radius 0.95, and 500 "slow" systems S2, which have at least one pole outside the circle with radius 0.95. The input is white Gaussian noise with unit variance, and the additive output noise is also white and Gaussian, with variance chosen as a factor (SNR) of the variance of the noiseless output. The 4 collections of data sets are described in Chen et al. (2012), and are named S1D1, S2D1, S1D2 and S2D2.

For these systems the stable spline kernel has been tested, with flat priors on the hyperparameters. 500 Monte Carlo simulations have been put forward per set, where for every Monte Carlo iteration the number of samples for the importance sampling was 1000.

The results are shown in Table 2, and a box plot of these fits is given in Figure 1. From them we can conclude that the proposed algorithm outperforms the empirical Bayes method in 3 out of the 4 data banks, and has very similar performance to the standard method for S2D1.

Table 2. Average fit of the SS kernel-based regularization estimators and its full Bayesian counterpart, for each set of random systems.

Estimator\MC study	S1D1	S2D1	S1D2	S2D2
SS	91.855	78.643	74.780	65.210
SS-Bayes	91.993	78.629	75.441	65.821

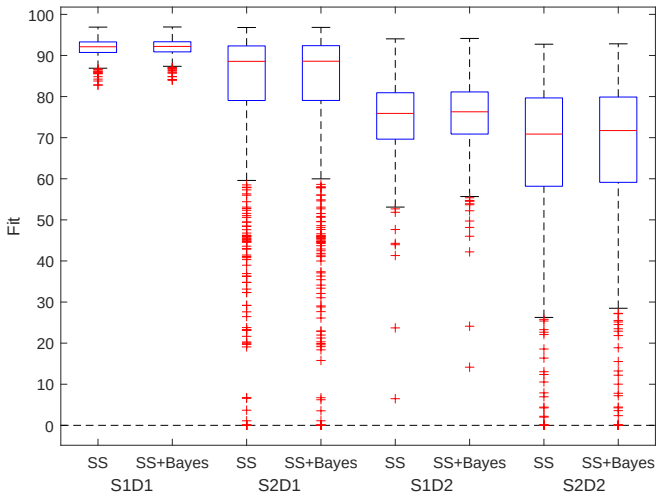

Fig. 1. Box plots of fits of the SS-kernel-basel estimates and their full Bayesian counterpart. SS+Bayes refers to the full Bayes approach.

5.3 Influence of the prior distribution

In Subsection 5.1, we have only worked with flat priors. This practical decision has given favorable results, but it is important to analyze the dependence of our method on the choice of the prior on β.

We consider the SS kernel as in (5), and the 4 data sets under the same conditions as the previous simulations but with prior density $p(\beta)$ given by

$$p(\beta) = p(\lambda, \alpha) = \frac{1}{2\sqrt{2\pi}} e^{-\frac{(\log(\lambda))^2}{8}} (2 - 4|\alpha - 0.5|), \quad (11)$$

that is, a zero-mean Gaussian pdf with variance 2 for $\log(\lambda)$, and a triangular pdf for α, centered in 0.5.

The results of 500 Monte Carlo simulations are shown in Table 3, and a direct comparison between fits for every Monte Carlo simulation can be found in Figure 2. These plots compare the goodness of fit of empirical Bayes versus the proposed method, for each Monte Carlo experiment. The red dots correspond to Monte Carlo simulations where the full Bayesian method outperforms the empirical Bayes approach. The blue dots represent the opposite, and the dashed green line is the separatrix. We refer to Rojas et al. (2015) for more details on benchmarks of random systems.

It is clear that the Bayesian approach outperforms the standard empirical Bayes method in most data sets, and

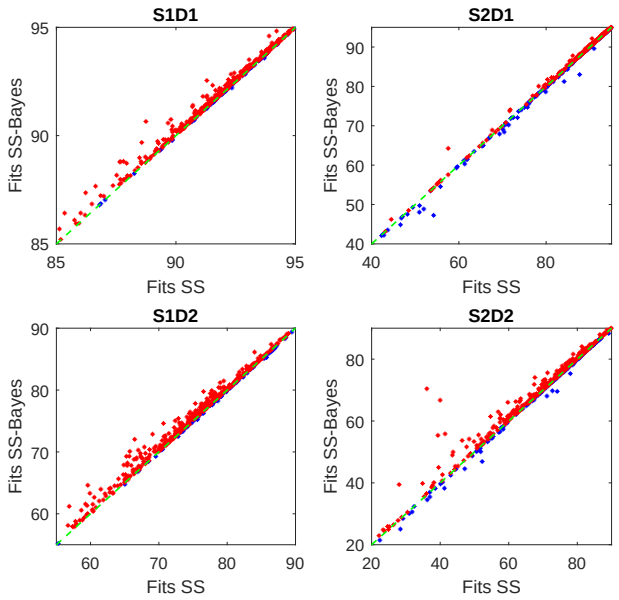

Fig. 2. Direct comparison between SS and Bayesian-SS fits for every Monte Carlo simulation and data set.

in most individual Monte Carlo simulations. A simple comparison between the results in Tables 2 and 3 suggests that the prior distribution of the hyperparameters does not lead to big differences in the performance of this estimator.

Table 3. Average fit of the SS kernel-based regularization estimators and its full Bayesian counterpart with prior (11), for each data set.

Estimator\MC study	S1D1	S2D1	S1D2	S2D2
SS	91.795	79.028	74.597	64.512
SS-Bayes	91.922	79.014	75.162	65.220

6. CONCLUSIONS

We have proposed a novel approach to estimate high-order FIR models by approaching the hyperparameter selection problem in a fully Bayesian manner. This technique avoids the selection of the vector of hyperparameters, and instead focuses on mixing regularized least squares estimates in an optimal way. Extensive simulations under standard data banks have shown that the proposed method increases the goodness of fit compared to standard kernel-based regularization techniques in many systems.

REFERENCES

Bottegal, G., Aravkin, A.Y., Hjalmarsson, H., and Pillonetto, G. (2014). Outlier robust system identification: a bayesian kernel-based approach. *IFAC Proceedings Volumes*, 47(3), 1073–1078.

Chen, T. and Ljung, L. (2013). Implementation of algorithms for tuning parameters in regularized least squares problems in system identification. *Automatica*, 49(7), 2213–2220.

Chen, T., Ohlsson, H., Goodwin, G.C., and Ljung, L. (2011). Kernel selection in linear system identification part II: A classical perspective. In *Decision and Control and European Control Conference (CDC-ECC)*, 4326–4331.

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

Cucker, F. and Smale, S. (2002). On the mathematical foundations of learning. *Bulletin of the American mathematical society*, 39(1), 1–49.

De Bruijn, N.G. (1970). *Asymptotic methods in analysis*, volume 4. Courier Corporation.

Dinuzzo, F. (2015). Kernels for linear time invariant system identification. *SIAM Journal on Control and Optimization*, 53(5), 3299–3317.

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

Hansen, B.E. (2007). Least squares model averaging. *Econometrica*, 75(4), 1175–1189.

Hjalmarsson, H. and Gustafsson, F. (1995). Composite modeling of transfer functions. *IEEE transactions on automatic control*, 40(5), 820–832.

Ljung, L. (1999). *System Identification: Theory for the User, 2nd edition*. Prentice-Hall.

Mackay, D. (1998). Introduction to Monte Carlo Methods. In *Learning in Graphical Models*, 175–204. Springer.

Madigan, J., Hoeting, D., Raftery, C., and Volinsky, C. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, 44(4), 382–417.

Pillonetto, G. (2016). A new kernel-based approach to hybrid system identification. *Automatica*, 70, 21–31.

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. and De Nicolao, G. (2011). Kernel selection in linear system identification part I: A Gaussian process perspective. In *Decision and Control and European Control Conference (CDC-ECC)*, 4318–4325.

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 *European Control Conference (ECC)*, 1365–1370.

Rojas, C.R., Valenzuela, P.E., and Rojas, R.A. (2015). A critical view on benchmarks based on randomly generated systems. *IFAC-PapersOnLine*, 48(28), 1471–1476.

Wahba, G. et al. (1999). Support vector machines, reproducing kernel hilbert spaces and the randomized GACV. *Advances in Kernel Methods-Support Vector Learning*, 6, 69–87.

Yang, Y. (2000). Combining different procedures for adaptive regression. *Journal of multivariate analysis*, 74(1), 135–161.