Low-Complexity Identification by Sparse Hyperparameter Estimation [⋆]

Mohammad Khosravi, [∗,1] **Mingzhou Yin,** [∗,1] **Andrea Iannelli,** [∗]
Anilkumar Parsi, [∗] **Roy S. Smith** [∗]

[∗] *Automatic Control Laboratory, ETH, Zürich 8092, Switzerland
(e-mail: {khosravm,myin,iannelli,aparsi,rsmith}@control.ee.ethz.ch).*

Abstract: This paper presents a novel kernel-based system identification method, which promotes low complexity of the model in terms of the McMillan degree of the system. The regularization matrix is characterized as a linear combination of pre-selected rank-one matrices with unknown hyperparameter coefficients, and the hyperparameters are derived using a maximum *a posteriori* estimation approach. Each basis matrix is the optimal regularization matrix for a first-order system. With this basis matrix selection, the McMillan degree of the identified model is upper-bounded by the rank of the regularization matrix, which in turn is equal to the cardinality of the hyperparameters. For this reason, a sparsity-promoting prior is chosen for hyperparameter tuning. The resulting optimization problem has a difference of convex program form which can be efficiently solved. The advantages of the proposed method are that the identified model has a low-complexity structure and that an improved bias-variance trade-off is achieved. Numerical results confirm that the proposed method achieves a better bias-variance trade-off as well as a better fit to the model compared to both the empirical Bayes method and the atomic-norm regularization.

Keywords: System identification, regularization, complexity tuning, hyperparameter estimation

1. INTRODUCTION

System identification deals with the problem of fitting a suitable mathematical model to a given set of measurement data corresponding to the input and the output of a dynamic system (Ljung, 1999).

In recent years, starting from the seminal work of Pillonetto and De Nicolao (2010), the idea of integrating prior knowledge in the model estimation other than in the model structure has received extensive attention (Pillonetto et al., 2014). Toward this goal, the system identification problem is framed as a regularized regression problem. The role of the regularization term is to impose a penalty on the feasible solutions which are not consistent with the prior knowledge. This prior knowledge can include various desired features of the model including the stability and the smoothness of the impulse response, the complexity of the model, the time constant and the resonant frequency of the system, etc. (Pillonetto et al., 2014; Chen, 2018; Marconato et al., 2016; Shah et al., 2012). Additionally, the issue of bias-variance trade-off can be addressed in this framework, especially when Tikhonov-like regularizations are utilized (Pillonetto et al., 2014).

The low complexity of the model can be imposed by the regularizer. In (Fazel et al., 2013; Smith, 2014), the rank and the nuclear norm of the Hankel matrix are used for penalizing the order of systems. In (Shah et al., 2012), the notion of the atomic transfer functions is introduced and the atomic norm is utilized for regularization. However, the issue of bias-variance trade-off is not completely addressed in these methods. This issue, as well as Bayesian interpenetration of these methods, is discussed in (Chiuso, 2016; Pillonetto et al., 2016).

Recently, the idea of multiple regularization has been introduced (Chen et al., 2014). It has been observed that multiple regularization can have better performance, especially when dealing with complex systems (Chen et al., 2014; Hong et al., 2018; Chen et al., 2018; Khosravi et al., 2020). In these approaches, the regularization matrix is presented parametrically in terms of basis regularization matrices with simple structure, such as tuned/correlated (TC) kernels (Chen et al., 2014; Hong et al., 2018) and filters (Chen et al., 2018). The parameters in this parameterization are known as *hyperparameters*. A number of approaches have been proposed in the literature to estimate the hyperparameters, such as the empirical Bayes (EB) method (Pillonetto et al., 2014), Stein unbiased risk estimator (SURE) (Hong et al., 2018), cross-validation (CV), and generalized cross-validation (GCV) (Mu et al., 2018a,b). This process is known as hyperparameter tuning. However, these methods are not suitable for encoding the low-complexity feature of the model.

In this paper, a novel multiple regularization method is proposed that promotes low-complexity system structure. The novelty of the method is in both the design of basis regularization matrices and the hyperparameter tuning. The basis regularization matrices are designed to be the

[⋆] This research project is part of the Swiss Competence Center for Energy Research SCCER FEEB&D of the Swiss Innovation Agency Innosuisse, and supported by the Swiss National Science Foundation under grant no.: 200021_178890.

[1] Both authors contributed equally to this work.

optimal regularization matrix for first-order systems. With this design, it is shown that the complexity of the identified model, measured by the McMillan degree of the system, is bounded by the cardinality of the hyperparameters. Then, the hyperparameters are estimated using a maximum *a posteriori* (MAP) approach. By selecting an appropriate prior on the hyperparameters, the MAP hyperparameter tuning gives a sparse estimation of the hyperparameters. This imposes the low-complexity feature on the identified model while maintaining the advantage of Bayesian regularization in terms of a favorable bias-variance trade-off. The resulting optimization problem has the form of the difference of convex programs (DCP) which can be efficiently solved. Simulation results demonstrate that the proposed method achieves a better bias-variance trade-off, as well as a better fit to the model, compared to existing methods.

2. NOTATION

In this paper, the set of natural numbers, the set of integers, the set of real numbers, the set of non-negative real numbers, and the set of complex numbers are shown as $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{R}$, $\mathbb{R}_+$, and $\mathbb{C}$, respectively. The set of complex numbers with magnitude less than one is called the *open unit disk* and is denoted by $\mathbb{D}$. For any $z \in \mathbb{C}$, the complex conjugate, the real and the imaginary parts of z are denoted by z^*, $\mathrm{real}(z)$ and $\mathrm{imag}(z)$, respectively. Let $\mathbb{F}$ be either $\mathbb{R}$ or $\mathbb{C}$. For any $1 \leq n \leq \infty$, the vector space of n-dimensional vectors with entries in $\mathbb{F}$ is denoted by $\mathbb{F}^n$ and $\mathrm{h} \in \mathbb{F}^n$ is expressed as $\mathrm{h} = (h_k)_{k=1}^n$ in terms of its entries, i.e., h_k is the entry of h at k^{th} location. The set of n-by-m matrices with entries in $\mathbb{F}$ is denoted by $\mathbb{F}^{n\times m}$. The set of n-by-n real symmetric positive-definite matrices, n-by-n real symmetric semi-positive-definite matrices, and n-by-n complex Hermitian semi-positive-definite matrices are denoted by $\mathbb{S}_{++}^n$, $\mathbb{S}_+^n$, and $\mathbb{H}_+^n$, respectively. The zero vector, the zero matrix, and the identity matrix are shown by $\mathbf{0}$, $\mathbb{0}$, and $\mathbb{I}$, respectively. The space of real-valued signals defined over $\mathbb{Z}$ is denoted by $\mathbb{R}^{\mathbb{Z}}$. The *forward shift operator*, denoted by q, is an operator on the space of signals, $\mathrm{q} : \mathbb{R}^{\mathbb{Z}} \to \mathbb{R}^{\mathbb{Z}}$, defined by $(\mathrm{qu})_t = u_{t+1}$, for any $t \in \mathbb{Z}$ and any $\mathrm{u} \in \mathbb{R}^{\mathbb{Z}}$. For any vector x, the Euclidean norm of x is denoted by $\|\cdot\|$. The *support* of $\mathrm{x} = (x_i)_{i=1}^n$ is defined as $\mathrm{supp}(\mathrm{x}) = \{i \mid x_i \neq 0\}$. The cardinality of x, denoted by $\|\mathrm{x}\|_0$, is defined as the number of elements of supp(x), i.e. $\|\mathrm{x}\|_0 = |\mathrm{supp}(\mathrm{x})|$. The trace of a matrix A is denoted by $\mathrm{tr}(A)$. The expression $X \sim \mathcal{N}(\mu, \Sigma)$ means that X is a Gaussian random vector with mean μ and covariance Σ. The probability density functions and the conditional probability densities are denoted by $p(\cdot)$ and $p(\cdot|\cdot)$, respectively.

3. REGULARIZED SYSTEM IDENTIFICATION

Consider a strictly causal and stable linear time-invariant (LTI) single-input single-output discrete-time system. The system can be represented with the transfer function $G_0(\mathrm{q})$ defined by $G_0(\mathrm{q}) = \sum_{k>0} g_k \mathrm{q}^k$, where $(g_k)_{k>0}$ is the impulse response of the system. The stability of $G_0(\mathrm{q})$ implies that the impulse response decays exponentially. Thus, it is reasonable to truncate the infinite impulse response at a sufficiently high order, denoted by $n_{\mathrm{g}} \in \mathbb{N}$. Accordingly, the system is approximated with a finite-length impulse response of $\mathrm{g} = [g_1, g_2, \ldots, g_{n_{\mathrm{g}}}]^{\mathsf{T}} \in \mathbb{R}^{n_{\mathrm{g}}}$. This leads to the finite impulse response (FIR) model of system which is defined as $G(\mathrm{q}) = \sum_{k=1}^{n_{\mathrm{g}}} g_k \mathrm{q}^k$.

3.1 Problem Statement

Define the input and the output of the system as $\mathrm{u} = (u_t)_{t\in\mathbb{Z}} \in \mathbb{R}^{\mathbb{Z}}$ and $\mathrm{y} = (y_t)_{t\in\mathbb{Z}} \in \mathbb{R}^{\mathbb{Z}}$, respectively. Assume the system is subject to independent and identically distributed (i.i.d.) Gaussian additive measurement noise $\mathrm{w} = (w_t)_{t\in\mathbb{Z}} \in \mathbb{R}^{\mathbb{Z}}, w_t \sim \mathcal{N}(0, \sigma_w^2), \forall t \in \mathbb{Z}$. Then, we have

$$y_t = \sum_{k=1}^{n_{\mathrm{g}}} g_k u_{t-k} + w_t, \quad \forall t \in \mathbb{Z} \tag{1}$$

Assume the inputs and the outputs of the system are measured at time instants $t = 0, 1, \ldots, n_{\mathcal{D}} - 1$. Define $\mathcal{D}$ as the set of pairs of measured input and output data, i.e., $\mathcal{D} = \{(u_t, y_t) \mid t = 0, 1, \ldots, n_{\mathcal{D}} - 1\}$. In this paper, we are interested in the problem of identifying the FIR model of the system given the data set $\mathcal{D}$.

3.2 Prediction Error Method as an ML Estimate

For $t = 0, \ldots, n_{\mathcal{D}} - 1$, define the vector φ_t as

$$\varphi_t = \begin{bmatrix} u_{t-1} & u_{t-2} & \ldots & u_{t-n_{\mathrm{g}}} \end{bmatrix}^{\mathsf{T}} \in \mathbb{R}^{n_{\mathrm{g}}}. \tag{2}$$

Here, for the sake of simplicity, it is assumed that the input is zero before $t = 0$, i.e., $u_t = 0$, for all $t < 0$. According to (1) and (2), we have $\mathrm{y} = \Phi\mathrm{g} + \mathrm{w}$, where $\mathrm{y} := [y_0, \ldots, y_{n_{\mathcal{D}}-1}]^{\mathsf{T}}$, $\mathrm{w} := [w_0, \ldots, w_{n_{\mathcal{D}}-1}]$ and $\Phi := [\varphi_0, \ldots, \varphi_{n_{\mathcal{D}}-1}]^{\mathsf{T}}$. Since $w_t \sim \mathcal{N}(0, \sigma_w^2)$, we have $\mathrm{y} - \Phi\mathrm{g} \sim \mathcal{N}(0, \sigma_w^2\mathbb{I})$. Therefore, one may estimate the impulse response based on a maximum likelihood (ML) approach as $\mathrm{g}^{\mathrm{ML}} = \mathrm{argmax}_{\mathrm{g}\in\mathbb{R}^{n_{\mathrm{g}}}}\ p(\mathrm{y}|\Phi, \mathrm{g})$, which is equivalent with

$$\mathrm{g}^{\mathrm{ML}} = \mathrm{argmin}_{\mathrm{g}\in\mathbb{R}^{n_{\mathrm{g}}}} \ \|\mathrm{y} - \Phi\mathrm{g}\|^2 = (\Phi^{\mathsf{T}}\Phi)^{-1}\Phi^{\mathsf{T}}\mathrm{y}. \tag{3}$$

According to (3), the ML estimation of the impulse response g^{ML} is the solution to the least squares (LS) problem equivalent to minimizing the prediction error. Therefore, g^{ML} can also be denoted by g^{LS}. Though the ML estimate g^{ML} is unbiased, the variance of the estimation can be high under high noise levels and/or the optimization problem can be ill-posed when $n_{\mathcal{D}}$ is small. These issues can be alleviated by introducing suitable priors and corresponding regularizers into (3) (Pillonetto et al., 2014; Chen, 2018), as shown next.

3.3 Regularization Method as an MAP Estimate

By introducing a suitable prior for the impulse response, a maximum *a posteriori* (MAP) approach can be employed for estimating the impulse response. More precisely, one can choose an appropriate matrix $\mathrm{S} \in \mathbb{S}_+^{n_{\mathrm{g}}}$ and set the prior for the impulse response as $\mathrm{g} \sim \mathcal{N}(\mathbf{0}, \mathrm{S})$. Subsequently, the MAP estimate of g is given by $\mathrm{g}^{\mathrm{MAP}} = \mathrm{argmax}_{\mathrm{g}\in\mathbb{R}^{n_{\mathrm{g}}}}\ p(\mathrm{g}|\Phi, \mathrm{y})$. Because $\mathrm{y} - \Phi\mathrm{g} \sim \mathcal{N}(0, \sigma_w^2\mathbb{I})$ and using the prior, one can easily see

$$\begin{bmatrix} \mathrm{g} \\ \mathrm{y} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathrm{S} & \mathrm{S}\Phi^{\mathsf{T}} \\ \Phi\mathrm{S} & \Phi\mathrm{S}\Phi^{\mathsf{T}} + \sigma_w^2\mathbb{I} \end{bmatrix} \right). \tag{4}$$

From this it follows

$$\mathrm{g}|\Phi, \mathrm{y} \sim \mathcal{N}\big(\mathrm{S}\Phi^{\mathsf{T}}\Psi^{-1}\mathrm{y}, \mathrm{S} - \mathrm{S}\Phi^{\mathsf{T}}\Psi^{-1}\Phi\mathrm{S}\big), \tag{5}$$

where $\Psi := \Phi \mathrm{S} \Phi^{\mathsf{T}} + \sigma_w^2 \mathbb{I}$. Thus, the MAP estimate gives

$$\mathrm{g}^{\mathrm{MAP}} = \mathrm{S}\Phi^{\mathsf{T}}\Psi^{-1}\mathrm{y}. \tag{6}$$

Let the singular value decomposition of S be given as

$$\mathrm{S} = [\mathrm{U}_1 \ \mathrm{U}_2] \begin{bmatrix} \mathrm{S}_1 & \mathbb{0} \\ \mathbb{0} & \mathbb{0} \end{bmatrix} \begin{bmatrix} \mathrm{U}_1^{\mathsf{T}} \\ \mathrm{U}_2^{\mathsf{T}} \end{bmatrix}, \tag{7}$$

where S_1 is a diagonal matrix with positive diagonal entries. We have that $\mathrm{g}^{\mathrm{MAP}}$ is the *unique* solution of the following regularized optimization (Chen et al., 2014)

$$\mathrm{g}^{\mathrm{MAP}} = \operatorname*{argmin}_{\mathrm{g}\in\mathbb{R}^{n_{\mathrm{g}}}, \mathrm{U}_2^{\mathsf{T}}\mathrm{g}=0} \ \|\mathrm{y} - \Phi\mathrm{g}\|^2 + \sigma_w^2 \mathrm{g}^{\mathsf{T}}\mathrm{U}_1\mathrm{S}_1^{-1}\mathrm{U}_1^{\mathsf{T}}\mathrm{g} \tag{8}$$

Consequently, this approach can be referred as the *regularized system identification.* The covariance matrix S, which is also called the *regularization matrix* or the *kernel*, imposes desired features on the estimated impulse response by encoding available prior knowledge like smoothness and stability in the estimation problem. Therefore, choosing an appropriate covariance matrix S has a significant impact on the estimation. In this regard, a suitable parametric family of candidate covariance matrices, $\mathcal{S} := \{\mathrm{S}_\eta \mid \eta \in \mathcal{E}\} \subseteq \mathbb{S}_+^{n_{\mathrm{g}}}$, is considered, where S_η defines the structure of the covariance matrix and η is the vector of hyperparameters and $\mathcal{E} \subseteq \mathbb{R}^{n_\eta}$. Given $\mathcal{S}$, the hyperparameters can be estimated from the measurement data $\mathcal{D}$.

Given a well-tuned covariance matrix, the regularization method leads to an estimate with the desired model structure and a satisfactory bias-variance trade-off. In this paper, the parametric set $\mathcal{S}$ is parameterized as a linear combination of a family of basis covariance matrices S_i, i.e., $\mathrm{S}_\eta = \sum_{i=1}^{n_\eta} \eta_i \mathrm{S}_i$, for all $\eta \in \mathcal{E}$. This structure is known as multiple kernel design in (Chen et al., 2014), and covers a broad class of systems. The next two sections deal with the following three main problems in multiple kernel design.

What is the appropriate choice of S_i*?* In Section 4, a suitable structure of basis covariance matrices S_i is proposed for estimating a model with low complexity.

What is the appropriate approach to tune η_i*?* In contrast to conventional single kernel design, where hyperparameter tuning is often conducted by non-convex optimization or grid search since n_η is small, in multiple kernel design a high-dimensional hyperparameter tuning problem has to be addressed. The high-dimensional problem is then prone to high variance and computational intractability. In Section 5, a sparse hyperparameter tuning approach is presented with regularized MAP estimation. This approach is shown to be computationally efficient.

Why does this method induce a low-complexity estimation? The theoretical foundation of the proposed method is established by Theorem 1.

4. COVARIANCE DESIGN: FROM SYSTEM COMPLEXITY TO HYPERPARAMETER SPARSITY

In real-world applications, many systems have simple structures with low complexity. To reveal the complexity of the system, one can model or closely approximate a system with the following linear fractional expansion of the transfer function

$$\tilde{G}(\mathrm{q}) = \sum_{i=1}^{s} \frac{\tilde{c}_i}{\mathrm{q} - w_i}, \quad \tilde{c}_i \neq 0, \tag{9}$$

where s corresponds to the McMillan degree of the system and assumed to be small. Motivated by this formulation, for any $w \in \mathbb{D}$, define the *atomic transfer function* $G^{(w)}$, as the following stable first-order transfer function (Shah et al., 2012)

$$G^{(w)}(\mathrm{q}) = \frac{\alpha_w}{\mathrm{q} - w}, \tag{10}$$

where α_w is a positive scalar normalizing a given system norm of $G^{(w)}$. In this paper, in line with (Shah et al., 2012), $G^{(w)}$ is normalized with respect to the Hankel nuclear norm with $\alpha_w = 1 - |w|^2$. Correspondingly, define the *atomic impulse response* $\mathrm{g}^{(w)}$ as the finite impulse response of $G^{(w)}$ given by

$$\mathrm{g}^{(w)} := \alpha_w[1, w, \ldots, w^{n_{\mathrm{g}}-1}]^{\mathsf{T}} \in \mathbb{C}^{n_{\mathrm{g}}}, \tag{11}$$

and the *atomic covariance matrix*

$$\mathrm{S}^{(w)} = \mathrm{g}^{(w)} \left(\mathrm{g}^{(w)}\right)^{\mathsf{H}} \in \mathbb{H}_+^{n_{\mathrm{g}}}. \tag{12}$$

As proved in (Chen et al., 2012), when the underlying system is exactly $G^{(w)}(\mathrm{q})$, the optimal kernel selection is given by $\mathrm{S}^{(w)}$.

One can easily see that $\tilde{G}(\mathrm{q})$ given in (9) is a linear combination of atomic transfer functions, i.e., $\{G^{(w)}\}_{w\in\mathbb{D}}$ spans the space of transfer functions of the form given in (9).

Let $n_\eta \in \mathbb{N}$ and $\mathcal{W} = \{w_1, \ldots, w_{n_\eta}\} \subset \mathbb{D}$ be a set of n_η stable distinct poles. For real-valued systems, it is required that for any $w \in \mathcal{W}$, one has $w^* \in \mathcal{W}$. The following structure of S_η is proposed

$$\mathrm{S}_\eta = \sum_{i=1}^{n_\eta} \eta_i \mathrm{S}^{(w_i)} = \sum_{i=1}^{n_\eta} \eta_i \mathrm{g}^{(w_i)} \left(\mathrm{g}^{(w_i)}\right)^{\mathsf{H}}, \tag{13}$$

where $[\eta_1 \ \eta_2 \ \ldots \ \eta_{n_\eta}]^{\mathsf{T}} \in \mathbb{R}_+^{n_\eta}$ is denoted by η. To satisfy the constraint of $\mathrm{S}_\eta \in \mathbb{S}_+^{n_{\mathrm{g}}}$, it is noted that $\mathrm{S}^{(w_i)} = \left(\mathrm{S}^{(w_j)}\right)^*$ for any $w_i = w_j^*$. So, η is constrained to satisfy

$$\eta_i = \eta_j, \qquad \forall i, j \text{ such that } w_i = w_j^*. \tag{14}$$

Based on this definition, we have the following theorem which is the theoretical foundation of the method proposed in this paper.

Theorem 1. Let $\hat{\mathrm{g}}$ be a realization sample of the impulse response prior distribution $\mathcal{N}(\mathbf{0}, \mathrm{S}_\eta)$, where S_η satisfies (13). Then, there exists $\mathrm{c} = [c_1 \ c_2 \ \ldots \ c_{n_\eta}]^{\mathsf{T}} \in \mathbb{C}^{n_\eta}$ such that the following hold.
i) For any i, j satisfying $w_i = w_j^*$, we have $c_i = c_j^*$.
ii) The sample realization impulse response $\hat{\mathrm{g}}$ can be decomposed as $\hat{\mathrm{g}} = \sum_{i=1}^{n_\eta} c_i \mathbf{1}_{\{\eta_i \neq 0\}}(\eta) \ \mathrm{g}^{(w_i)}$.
iii) Let $\mathrm{g}_{\mathrm{c}} \in \mathbb{R}^{n_{\mathrm{g}}}$ be a finite impulse response which corresponds to a transfer function, denoted by G_{c} and defined as $G_{\mathrm{c}}(\mathrm{q}) = \sum_{i=1}^{n_\eta} \left(c_i \mathbf{1}_{\{\eta_i \neq 0\}}(\eta)\right) G^{(w_i)}(\mathrm{q})$. Then, one has $\mathrm{g} = \mathrm{g}_{\mathrm{c}}$.

Corollary 2. The poles of $G_{\mathrm{c}}(\mathrm{q})$ is a subset of $\{w_i \mid i \in \mathrm{supp}(\eta)\}$ and the McMillan degree of $G_{\mathrm{c}}(\mathrm{q})$ is less than or equal to $\|\eta\|_0$. Therefore, any $\mathrm{g} \sim \mathcal{N}(\mathbf{0}, \mathrm{S}_\eta)$ corresponds to a system with a McMillan degree of at most $\|\eta\|_0$.

According to Theorem 1 and Corollary 2, it can be seen that the desired low-complexity structure of the system is induced by the sparsity of the hyperparameters η. However, in order to have a favorable estimation of the impulse response, the set $\mathcal{W}$ should be nearly dense in $\mathbb{D}$. Accordingly, it is advantageous to employ a large set of basis covariance matrices. This suggests performing sparse estimation at the level of hyperparameter tuning, which will be discussed in the next section. Meanwhile, a satisfactory bias-variance trade-off at the level of impulse response estimation is maintained by utilizing the regularized identification method in (8).

Remark 3. This characteristic differs from the atomic-norm regularization in (Shah et al., 2012), where l_1-norm regularization is directly imposed on the decomposition coefficient vector c at the level of impulse response estimation. This direct sparsity regularization adds more regularization to the impulse response estimation. The atomic-norm regularization is known to suffer from high bias as discussed in (Pillonetto et al., 2016).

5. HYPERPARAMETER TUNING: AN MAP APPROACH

This section present a MAP approach that imposes additional sparsity regularization for the estimation of the hyperparameters introduced in Section 4.

5.1 MAP Estimation of Hyperparameters

Define set $\mathcal{I}^+$ as $\mathcal{I}^+ = \{i \in \{1, \dots, n_\eta\} \mid \mathrm{imag}(w_i) \geq 0\}$. According to the structural constraint of S_η in (14), one only needs to estimate η_i, for $i \in \mathcal{I}^+$. The remaining hyperparameters are determined automatically. In this section, with an abuse of notation, η denotes the vector of independent hyperparameters $(\eta_i)_{i\in\mathcal{I}^+}$. Accordingly, n_η is the length of η.

In order to introduce an MAP estimation approach for η, the following prior, that promotes the sparsity of the hyperparameters η, is defined. Since the entries of η are initially (a priori) undiscriminating and independent, the prior of each entry of η is selected as an i.i.d. exponential distribution (Aravkin et al., 2014). More precisely, for any i, we have $p(\eta_i) = \lambda \exp(-\lambda\eta_i)\mathbf{1}_{\{\eta_i\geq 0\}}(\eta_i)$, where $\lambda > 0$ is the *rate parameter* of the distribution. The "hyper-hyperparameter" λ parameterizes the prior on the hyperparameter. In this paper, λ is estimated by cross validation. Accordingly, we have

$$p(\eta) = \lambda^{n_\eta} \exp(-\lambda \sum_{i=1}^{n_\eta} \eta_i)\ \mathbf{1}_{\{\eta\in\mathbb{R}_+^{n_\eta}\}}(\eta). \tag{15}$$

From (4), we know that

$$\mathrm{y}|\eta \sim \mathcal{N}(\mathbf{0}, \Phi \mathrm{S}_\eta \Phi^\mathsf{T} + \sigma_w^2 \mathbb{I}). \tag{16}$$

By Bayes' rule, one can estimate the vector of hyperparameters by MAP estimation as follows.

$$\eta^{\mathrm{MAP}} = \mathrm{argmax}_{\eta\in\mathbb{R}^{n_\eta}}\ p(\mathrm{y}|\eta)p(\eta), \tag{17}$$

where the denominator in the Bayes fraction $p(\mathrm{y})$ is removed due to being independent of η. From (15), (16), (17), and the monotonicity of the logarithm function, one can see that

$$\eta^{\mathrm{MAP}} = \mathrm{argmin}_{\eta\in\mathbb{R}_+^{n_\eta}}\ \frac{1}{2}\mathrm{y}^\mathsf{T}\Sigma_\eta^{-1}\mathrm{y} + \frac{1}{2}\,\mathrm{logdet}\,\Sigma_\eta + \lambda\sum_{i=1}^{n_\eta}\eta_i, \tag{18}$$

where $\Sigma_\eta = \Phi \mathrm{S}_\eta \Phi^\mathsf{T} + \sigma_w^2\mathbb{I}$.

The following proposition provides the theoretical basis for solving the hyper-estimation problem (18) efficiently.

Propositon 4. Optimization problem (18) can be expressed as a DCP problem.

Proof. Let $F, H : \mathbb{R}_+^{n_\eta} \to \mathbb{R}$ be defined as

$$F(\eta) = \frac{1}{2}\mathrm{y}^\mathsf{T}\Sigma_\eta^{-1}\mathrm{y} + \lambda\sum_{i=1}^{n_\eta}\eta_i,\ H(\eta) = -\frac{1}{2}\,\mathrm{logdet}\,\Sigma_\eta. \tag{19}$$

One can easily see that F and H are convex functions and the objective function in (18) is $F(\eta) - H(\eta)$. This concludes the proof. □

According to Proposition 4, one can obtain a stationary point of the optimization problem (18) efficiently. In this regard, the problem can be solved by an appropriate solver designed for disciplined programming of DCP problems (Shen et al., 2016), or a suitable algorithm for solving this type of optimization problem, e.g. Yuille and Rangarajan (2002). Here, the latter approach is applied which is essentially a majorization minimization (MM) algorithm and solves the problem using a sequential convex programming scheme. In detail, a *majorization* function $J : \mathbb{R}_+^{n_\eta} \times \mathbb{R}_+^{n_\eta} \to \mathbb{R}$ can be defined, if we have

$$J(\eta, \eta) = F(\eta) - H(\eta), \qquad \forall \eta \in \mathbb{R}_+^{n_\eta}, \tag{20}$$

$$J(\eta, \gamma) \geq F(\eta) - H(\eta), \qquad \forall \eta, \gamma \in \mathbb{R}_+^{n_\eta}. \tag{21}$$

Following this, the method solves the problem iteratively using the following procedure

$$\eta^{(k+1)} = \mathrm{argmin}_{\eta\in\mathbb{R}_+^{n_\eta}} J(\eta, \eta^{(k)}). \tag{22}$$

Here, the majorization function is defined by replacing $H(\eta)$ with its linearization at γ in the objective function. Then, constraints (20) and (21) are satisfied due to the convexity of $H(\eta)$. More precisely, we have

$$J(\eta, \gamma) := F(\eta) - H(\gamma) - \nabla_\gamma H(\gamma)^\mathsf{T}(\eta - \gamma), \tag{23}$$

where

$$\nabla_{\eta_i} H(\eta) = -\frac{1}{2}\mathrm{tr}\left(\Sigma_\eta^{-1}\Phi \mathrm{S}^{(w_i)}\Phi^\mathsf{T}\right). \tag{24}$$

Note that the term $H(\gamma)$ does not depend on the optimization variable η and can thus be removed.

The hyperparameter tuning procedure, by solving the MAP estimate (18) with the MM algorithm, is summarized in Algorithm 1. Then, the FIR model is identified by (6) with $\mathrm{S} = \mathrm{S}_{\eta^{\mathrm{MAP}}}$.

5.2 Empirical Bayes and Regularized Empirical Bayes

In the empirical Bayes method (Chen et al., 2014), the hyperparameters are estimated by maximizing the marginal likelihood, i.e., $\eta^{\mathrm{EB}} = \mathrm{argmax}_{\eta\in\mathbb{R}_+^{n_\eta}}\ p(\mathrm{y}|\eta)$. Note that there is an implicit prior on the hyperparameters, which is $\eta \in \mathbb{R}_+^{n_\eta}$. The method is essentially a ML approach which is also known as *type-II maximum likelihood* (Good, 1966). The empirical Bayes method performs well when the number of hyperparameters is not large and the data set is not significantly small or noisy. However, as discussed in Section 4, here a considerably large number of hyperparameters are to be estimated . In this situation, the solution of the empirical Bayes method is prone to high variance. Consequently, it is preferable to employ the

Algorithm 1 MAP-based hyperparameter tuning

```
Input: $\mathrm{y},\Phi,\mathcal{W}$
Calculate $\mathrm{S}^{(w)}$, for $w \in \mathcal{W}$ (see (12) and (11)).
$k \leftarrow 0$, initialize $\eta^{(0)}$.
while stopping condition is not met do
    Calculate $\nabla_{\eta^{(k)}} H(\eta^{(k)})$ (see (24)) and
    From (19) and (23), set the function $J$ as
        $J(\eta,\eta^{(k)}) = F(\eta) - \nabla_{\eta^{(k)}} H(\eta^{(k)})(\eta - \eta^{(k)})$.
    Solve convex optimization problem
        $\eta^{(k+1)} = \mathrm{argmin}_{\eta \in \mathbb{R}_+^{n_\eta}} J(\eta,\eta^{(k)})$.
    $k \leftarrow k + 1$.
end
Output: $\eta^{\mathrm{MAP}} = \eta^{(k)}$
```

MAP approach, especially when the prior knowledge on the sparsity of the hyperparameters is available.

From (16), one can see that

$$\eta^{\mathrm{EB}} = \mathrm{argmin}_{\eta \in \mathbb{R}_+^{n_\eta}} \frac{1}{2}\mathrm{y}^{\mathsf{T}}\Sigma_\eta^{-1}\mathrm{y} + \frac{1}{2}\operatorname{logdet}\Sigma_\eta. \tag{25}$$

The difference between (18) and (25) is a single term in the objective function, i.e., $\lambda\sum_{i=1}^{n_\eta}\eta_i$, which comes from the sparse prior. In fact, this term performs a regularization on the estimation of the hyperparameters and subsequently improves the bias-variance trade-off. Therefore, one can alternatively call η^{MAP} as the solution of the *regularized empirical Bayes* estimation and denote it by η^{REB}.

One should note that the empirical Bayes method may also induce sparsity in the estimation of the hyperparameters in a high-dimensional setting. This is due to the implicit prior of non-negative η. However, the induced sparsity is governed by the data rather than assumed prior, and thus can be noise-dependent. In the method proposed here, the sparsity is governed by the prior knowledge (low-complexity in this case) of the system. As a consequence, the estimation encodes more system-dependent features comparing to the empirical Bayes method, which makes it more robust with respect to measurement noise. Here, the sparsity of the estimator is controlled by the rate parameter λ. Meanwhile, one can still obtain the empirical Bayes estimate by setting $\lambda = 0$.

6. NUMERICAL RESULTS

In this section, the proposed sparse hyperparameter tuning method is compared to other existing regularization formulation with the atomic structure. For the baseline performance, the least squares method without regularization and the well-known first-order stable spline kernel, also known as the TC kernel, introduced in (Chen et al., 2012) are also compared. Note that these two methods do not estimate a low-complexity model. Specifically, the following five identification schemes are compared. The least squares method (*LS*) corresponds to the estimate g^{ML} in (3). The system is also identified with a TC kernel (*TCK*) regularization. The hyperparameters are selected by the empirical Bayes method with a non-convex optimization. This is also the defaulted identification method used in the MATLAB command `impulseest`. The atomic-norm method (*Atom*) applies the atomic-norm regularization proposed in (Shah et al., 2012). *Atom* applies a set of atomic transfer function characterized by the poles $w_i = r \cdot e^{j\phi}$, where $\phi = [0 : \pi/15 : \pi]$ in the MATLAB notation. The magnitude r is in a 15-point logspace grid of base 10^6 between 0.8 and 1 to obtain a denser grid near $r = 1$. The empirical Bayes method (*EB*) uses the hyperparameter estimation scheme (25) introduced in (Chen et al., 2014) without explicitly exploiting the sparse kernel structure. The regularized empirical Bayes method (*REB*) refers to the method proposed in this paper (Algorithm 1). Both *EB* and *REB* regularize the problem with the first-order kernel set $\mathcal{S}$ paramerized by (13) with the same set of poles as *Atom*. This gives a total of $n_\eta = 240$ kernels.

To highlight the characteristics of the bias-variance trade-off in these methods, numerical simulations are conducted on a benchmark system under i.i.d Gaussian noise of three different levels ($\sigma^2 = 0.1, 0.01, 0.001$) with 150 different noise realizations each. The transfer function of the chosen fourth-order discrete-time LTI system is

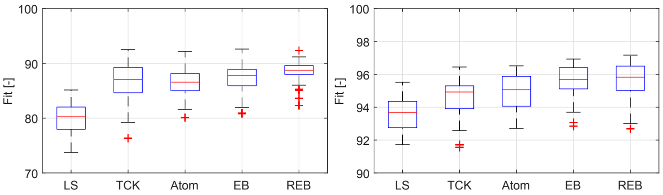
$$G(z) = \frac{z^3 + 0.5z^2}{z^4 - 2.2z^3 + 2.42z^2 - 1.87z + 0.7225} \tag{26}$$

which is one of the benchmark systems tested in (Pillonetto and De Nicolao, 2010). The input to the system is Gaussian with $u(t) \sim \mathcal{N}(0,1)$. The length of the identification data is $n_{\mathcal{D}} = 150$ and the order of the FIR model is $n_{\mathrm{g}} = 50$. For *EB* and *REB*, the noise variance σ^2 is estimated from the variance of the residuals in *LS*. For *Atom* and *REB*, the weighting of the l_1-norm regularization and the rate parameter λ are cross-validated over a five-point grid of `logspace(0,4,5)` and `logspace(-1,1,5)` in the MATLAB notation, respectively. The DCP optimization problem in *REB* is solved by a fixed number of five iterations.

First, the fitting performance of the five methods are compared in Fig. 1 using box plots. The fits of the estimates are defined as

$$W = 100 \times \left(1 - \left[\frac{\sum_{i=1}^{n_g}(g_i - \hat{g}_i)^2}{\sum_{i=1}^{n_g}(g_i - \bar{g})^2}\right]^{1/2}\right), \tag{27}$$

where g_i are the true impulse response coefficients, $\hat{g}_i$ are the estimated coefficients, and $\bar{g}$ is the mean of the true coefficients.



(a) $\sigma^2 = 0.01$

(b) $\sigma^2 = 0.001$

Fig. 1. Comparison of fitting performance under different noise levels. The last three methods estimate a low-complexity model with atomic structure.

As can be seen from Fig. 1, *REB* achieves the best fitting performance at all three noise levels. Compared to *REB*, the performance of *EB* is poor under the high noise level, whereas *Atom* and *TCK* perform worse under the low noise level. In all three cases, *LS* fails to achieve a good performance without regularization. Furthermore, the bias and the variance of the estimates are calculated in Table 1.

It can be seen, as discussed in (Pillonetto et al., 2014), that the bias-variance trade-off is controlled by the amount of regularization imposed. The mean square error (MSE) of LS is dominated by the variance since no regularization is imposed. Note that there is an inherent bias induced by the impulse response truncation. *Atom*, conversely, induces the highest amount of bias with l_1-norm regularization. The proposed method REB imposes more regularization than EB with an additional sparsity regularization in hyperparameter tuning, but less regularization compared to the direct sparsity regularization on the impulse response estimation as in *Atom*. This characteristic leads to an appropriate balance in terms of the bias-variance trade-off as can be seen from the MSE values. This result agrees with the discussion in Remark 3 and Section 5.2.

Table 1. Bias-variance trade-off

	LS	**TCK**	**Atom**	**EB**	**REB**
$\sigma^2 = 0.01$					
Bias^2 $[\times 10^{-4}]$	1.6	8.5	11.2	5.6	8.4
Var $[\times 10^{-4}]$	61.0	23.6	17.1	20.8	12.3
MSE $[\times 10^{-4}]$	62.6	32.1	28.3	26.4	**20.8**
$\sigma^2 = 0.001$					
Bias^2 $[\times 10^{-4}]$	0.15	0.96	1.24	0.37	0.68
Var $[\times 10^{-4}]$	6.23	3.79	2.97	2.98	2.52
MSE $[\times 10^{-4}]$	6.38	4.75	4.21	3.35	**3.20**

7. CONCLUSIONS

In this paper, we have presented a novel regularized system identification method using a low-complexity kernel design. The main characteristic of the proposed approach is that it promotes low complexity in terms of the McMillan degree of the identified system with a satisfactory bias-variance trade-off. In this method, the regularization matrix is framed as a linear combination of low-rank matrices with unknown coefficients as hyperparameters, which are then estimated using an MAP approach with a sparse prior. This design upper-bounds the McMillan degree of the identified model with the cardinality of the hyperparameters. The hyperparameter tuning problem is formulated as a difference of convex programming problem which can be efficiently solved to a local minimum. Numerical experiments confirm the effectiveness of the proposed approach, in terms of a better bias-variance trade-off as well as a better fit to the model comparing to both the empirical Bayes method and the atomic-norm regularization.

REFERENCES

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

Chen, T. (2018). On kernel design for regularized LTI system identification. *Automatica*, 90, 109–122.

Chen, T., Andersen, M.S., Ljung, L., Chiuso, A., and Pillonetto, G. (2014). System identification via sparse multiple kernel-based regularization using sequential convex optimization techniques. *IEEE Transactions on Automatic Control*, 59(11), 2933–2945.

Chen, T., Andersen, M.S., Mu, B., Yin, F., Ljung, L., and Qin, S.J. (2018). Regularized LTI system identification with multiple regularization matrix. *IFAC-PapersOnLine*, 51(15), 180–185.

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

Chiuso, A. (2016). Regularization and Bayesian learning in dynamical systems: past, present and future. *Annual Reviews in Control*, 41, 24–38.

Fazel, M., Pong, T.K., Sun, D., and Tseng, P. (2013). Hankel matrix rank minimization with applications to system identification and realization. *SIAM Journal on Matrix Analysis and Applications*, 34(3), 946–977.

Good, I.J. (1966). The estimation of probabilities. *Journal of the Institute of Mathematics and its Applications*, 2, 364–383.

Hong, S., Mu, B., Yin, F., Andersen, M.S., and Chen, T. (2018). Multiple kernel based regularized system identification with SURE hyper-parameter estimator. *IFAC-PapersOnLine*, 51(15), 13–18.

Khosravi, M., Iannelli, A., Yin, M., Parsi, A., and Smith, R.S. (2020). Regularized system identification: A hierarchical Bayesian approach. *IFAC-papersonline*.

Ljung, L. (1999). *System identification: theory for the user*. Prentice Hall.

Marconato, A., Schoukens, M., and Schoukens, J. (2016). Filter-based regularisation for impulse response modelling. *IET Control Theory & Applications*, 11, 194–204.

Mu, B., Chen, T., and Ljung, L. (2018a). Asymptotic properties of generalized cross validation estimators for regularized system identification. *IFAC-PapersOnLine*, 51(15), 203–208.

Mu, B., Chen, T., and Ljung, L. (2018b). Asymptotic properties of hyperparameter estimators by using cross-validations for regularized system identification. In *Conference on Decision and Control*, 644–649.

Pillonetto, G., Chen, T., Chiuso, A., Nicolao, G.D., and Ljung, L. (2016). Regularized linear system identification using atomic, nuclear and kernel-based norms: The role of the stability constraint. *Automatica*, 69, 137–149.

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.

Shah, P., Bhaskar, B.N., Tang, G., and Recht, B. (2012). Linear system identification via atomic norm regularization. In *Conference on Decision and Control*, 6265–6270.

Shen, X., Diamond, S., Gu, Y., and Boyd, S. (2016). Disciplined convex-concave programming. In *Conference on Decision and Control*, 1009–1014.

Smith, R.S. (2014). Frequency domain subspace identification using nuclear norm minimization and Hankel matrix realizations. *IEEE Transactions on Automatic Control*, 59(11), 2886–2896.

Yuille, A.L. and Rangarajan, A. (2002). The concave-convex procedure (CCCP). In *Advances in Neural Information Processing Systems*, 1033–1040.