# Regularized System Identification: A Hierarchical Bayesian Approach ⋆

**Mohammad Khosravi,** * **Andrea Iannelli,** * **Mingzhou Yin,** *
**Anilkumar Parsi,** * **Roy S. Smith** *

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

Abstract: In this paper, the hierarchical Bayesian method for regularized system identification is introduced. To this end, a hyperprior distribution is considered for the regularization matrix and then, the impulse response and the regularization matrix are jointly estimated based on a maximum a posteriori (MAP) approach. Toward introducing a suitable hyperprior, we decompose the regularization matrix using Cholesky decomposition and reduce the estimation problem to the cone of upper triangular matrices with positive diagonal entries. Following this, the hyperprior is introduced on a designed sub-cone of this set. The method differs from the current trend in regularized system identification from various aspect, e.g., the estimation is performed by solving a single stage problem. The MAP estimation problem reduces to a multi-convex optimization problem and a sequential convex programming algorithm is introduced for solving this problem. Consequently, the proposed method is a computationally efficient strategy specially when the regularization matrix has a large size. The method is numerically verified on benchmark examples. Owing to the employed full Bayesian approach, the estimation method shows a satisfactory bias-variance trade-off.

*Keywords:* System identification, hierarchical Bayesian, sequential convex programming.

## 1. INTRODUCTION

System identification address the problem of estimating appropriate mathematical models of dynamical systems using a set of measured data and methods of statistics (Ljung, 1999). The importance of models for prediction and control of physical systems is well recognized, for example thermal energy system in the buildings, network of agents, robotics, biological systems among many others.

Including regularization in the identification problem has received an extensive attention. The underlying idea is to integrate the prior knowledge, which is available in addition to the measurement data, into in the model estimation problem. For the purpose of obtaining a system featuring low complexity in McMillan sense, various types of regularization such as the Hankel matrix rank, the nuclear norm and the atomic norm of system are introduced in the literature (Fazel et al., 2013; Mohan and Fazel, 2010; Smith, 2014; Shah et al., 2012). Additionally, starting from the seminal work of Pillonetto and De Nicolao (2010), the problem of studying Tikhonov-type regularizations for imposing constraints on the latent system, like the smoothness and stability of the impulse response, has received significant attention (Pillonetto et al., 2014). In this framework, the system identification problem is formulated as a regularized regression with a regularization term coming from the norm of a reproducing kernel Hilbert space (RKHS) (Aronszajn, 1950) which penalizes the feasible solutions not agreeing with the prior knowledge. The main ingredient of these Hilbert spaces is the kernel function. Subsequently, designing a suitable regularization boils down to characterizing appropriate kernels. In the context of regularized system identification, the most common kernels are *tuned/correlated* (TC), *diagonal/correlated* (DC) and *stable spline* (SS) (Pillonetto et al., 2014). Moreover, many other kernels and also regularization matrices have been proposed following various approaches, e.g., inspired by machine learning and system theory (Chen, 2018; Khosravi et al., 2020), using the harmonic analysis of stochastic processes (Zorzi and Chiuso, 2018), based on filter design methods (Marconato et al., 2016). One can use various operation on given kernels or regularization matrices and introduce new ones (Hong et al., 2018; Chen et al., 2014, 2018). Once the type of kernel or regularization matrices is fixed, the characterizing hyperparameters are estimated using the given measurement data $\mathcal{D}$. In this regard, a number of approaches are introduced for hyperparameter estimation, e.g., empirical Bayes (EB) (Pillonetto et al., 2014), Stein unbiased risk estimator (SURE) (Hong et al., 2018) and cross-validation (CV) (Mu et al., 2018a,b). In summary, in the current trend of regularized system identification, first a prior or a regularization for the impulse response is considered and subsequently, the parameters of prior are tuned, and finally, the impulse response is estimated.

In this paper, the hierarchical Bayesian method for regularized system identification is investigated where a *joint estimation* for the impulse response and the regularization

⋆ 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.

matrix is proposed based on a maximum a posteriori (MAP) approach. To this end, appropriate hyperprior which suitably determines the structure of regularization matrix is introduced, and subsequently the form of impulse response. Moreover, the hyperprior is designed so that the estimation method is numerical tractable. The regularization matrix is indeed decomposed using Cholesky decomposition and the estimation problem is reduced to a subcone of the space of upper triangular matrices with positive diagonal entries. The final MAP estimation problem has a multi-convex programming form which can be efficiently solved using an introduced sequential convex programming algorithm. The method is finally numerically verified on benchmark examples.

## 2. NOTATIONS

The set of natural numbers, non-negative integers, integers, real numbers and non-negative real numbers are denoted here by $\mathbb{N}$, $\mathbb{Z}_+$, $\mathbb{Z}$, $\mathbb{R}$ and $\mathbb{R}_+$, respectively. Let $\mathbb{F}$ be either $\mathbb{R}$ or $\mathbb{R}_+$. Then, for any $n \in \mathbb{N}$, the space of $n$-dimensional vectors with entries in $\mathbb{F}$ is denoted by $\mathbb{F}^n$ and $\mathrm{h} \in \mathbb{F}^n$ is indicated by $\mathrm{h} = (h_k)_{k=1}^n$ to show explicitly in terms of its entries, i.e., $h_k$ is the $k^{\text{th}}$ entry of h. 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, $n$-by-$n$ upper triangular matrices with non-negative diagonal entries and $n$-by-$n$ upper triangular matrices with positive diagonal entries are denoted by $\mathbb{S}^n_{++}$, $\mathbb{S}^n_{+}$, $\mathbb{U}^n_{+}$ and $\mathbb{U}^n_{++}$, respectively. The zero vector, the zero matrix, and the identity matrix are indicated with $\mathbf{0}$, $\mathbb{0}$, and $\mathbb{I}$, respectively. For any vector x, the Euclidean norm of x is shown by $\|\cdot\|$. The interior and closure of set $\mathcal{S}$ is denoted by $\mathrm{int}\mathcal{S}$ and $\mathrm{cl}\mathcal{S}$, respectively. The convex cone of set $\mathcal{X}$ is denoted by $\mathrm{convcone}\mathcal{X}$. Given set $\mathcal{X}$, $\mathbf{1}_{\mathcal{X}}$ is a function which is one on $\mathcal{X}$ and zero elsewhere. 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 $\mathrm{u} \in \mathbb{R}^{\mathbb{Z}}$. By $X \sim \mathcal{N}(\mu, \Sigma)$, we mean that $X$ is a Gaussian random vector with mean $\mu$ and covariance $\Sigma$. The probability density functions are denoted by $p$. The *support* of distribution $\mu$ on the measurable space $(\mathscr{X}, \mathcal{A})$ is defined as $\mathrm{supp}(\mu) = \{\mathrm{x} \in \mathscr{X} \mid \mu(x) \neq 0\}$.

## 3. SYSTEM IDENTIFICATION

We consider a single-input-single-output causal and stable linear time-invariant (LTI) system described by transfer function $G(\mathrm{q})$. Let the input signal $\mathrm{u} \in \mathbb{R}^{\mathbb{Z}}$ be applied to the system and the output signal of system, denoted by $\mathrm{y} \in \mathbb{R}^{\mathbb{Z}}$, be measured. We assume that the measurement of the output is subject to additive white measurement noise. Accordingly, given that $\mathrm{u} = (u_t)_{t\in\mathbb{Z}}$ and $\mathrm{y} = (y_t)_{t\in\mathbb{Z}}$, at any measurement time instant $t$, we have

$$y_t = G(\mathrm{q})u_t + w_t, \tag{1}$$

where $\mathrm{w} := (w_t)_{t\in\mathbb{Z}} \in \mathbb{R}^{\mathbb{Z}}$ is the measurement noise, a zero mean white Gaussian signal with variance $\sigma_w^2$. More precisely, w is a sequence of independent and identically distributed (i.i.d.) random variables with distribution $\mathcal{N}(0, \sigma_w^2)$. Given a finite set of measurement data, it is desired to estimate the LTI system.

### *3.1 Maximum Likelihood and Prediction Error Method*

In the classical system identification (Ljung, 1999), besides nonparametric approaches such as *empirical transfer function estimation* (ETFE), the system is represented in a parametric form and then, the parameters of the system are estimated using available measurement data. This procedure is done principally using a *maximum likelihood* (ML) approach. The system is assumed stable, one possible parametric representation of the system is in form of a finite impulse response (FIR). More precisely, the transfer function of the system $G(\mathrm{q})$ is approximated by

$$G(q) = \textstyle\sum_{k=0}^{n_{\mathrm{g}}-1} g_k \mathrm{q}^{-k}, \tag{2}$$

where $\mathrm{g} := (g_k)_{k=0}^{n_{\mathrm{g}}-1} \in \mathbb{R}^{n_{\mathrm{g}}}$ is the corresponding FIR of the system. Now, let $\mathcal{D}$ be a set of measurement data

$$\mathcal{D} = \{(u_t, y_t) \mid t = 0, 1, \ldots, n_{\mathcal{D}} - 1\}. \tag{3}$$

Define vector $\varphi_t$ as

$$\varphi_t = \begin{bmatrix} u_t & u_{t-1} & \ldots & u_{t-n_{\mathrm{g}}+1} \end{bmatrix}^{\mathsf{T}}, \tag{4}$$

for $t = 0, \ldots, n_{\mathcal{D}} - 1$. Also, define vector y, vector w and matrix $\Phi$ respectively as

$$\mathrm{y} = \begin{bmatrix} y_0 \\ \vdots \\ y_{n_{\mathcal{D}}-1} \end{bmatrix}, \quad \mathrm{w} = \begin{bmatrix} w_0 \\ \vdots \\ w_{n_{\mathcal{D}}-1} \end{bmatrix}, \quad \Phi = \begin{bmatrix} \varphi_0^{\mathsf{T}} \\ \vdots \\ \varphi_{n_{\mathcal{D}}-1}^{\mathsf{T}} \end{bmatrix}. \tag{5}$$

From (1) and (2), one has

$$\mathrm{y} = \Phi\mathrm{g} + \mathrm{w}. \tag{6}$$

Since, for any $t$, we have $w_t \sim \mathcal{N}(0, \sigma_w^2)$, it holds $\mathrm{y} - \Phi\mathrm{g} \sim \mathcal{N}(0, \sigma_w^2\mathbb{I})$. Therefore, it follows that

$$p(\mathrm{y}|\Phi, \mathrm{g}) = (2\pi\sigma_w^2)^{-\frac{n}{2}} \exp(-\frac{1}{2\sigma_w^2}\|\mathrm{y} - \Phi\mathrm{g}\|^2), \tag{7}$$

and the ML estimation of g is defined as

$$\mathrm{g}^{\mathrm{ML}} = \mathrm{argmax}_{\mathrm{g}\in\mathbb{R}^{n_{\mathrm{g}}}}\ p(\mathrm{y}|\Phi, \mathrm{g}). \tag{8}$$

Since, we have

$$-\ln p(\mathrm{y}|\Phi, \mathrm{g}) = \frac{n}{2}\ln(2\pi\sigma_w^2) + \frac{1}{2\sigma_w^2}\|\mathrm{y} - \Phi\mathrm{g}\|^2, \tag{9}$$

it is easily deduced that

$$\mathrm{g}^{\mathrm{ML}} = \mathrm{argmin}_{\mathrm{g}\in\mathbb{R}^{n_{\mathrm{g}}}}\ \|\mathrm{y} - \Phi\mathrm{g}\|^2. \tag{10}$$

From (10), one can see that the ML estimation, $\mathrm{g}^{\mathrm{ML}}$, is obtained by solving a least square (LS) problem which is the minimization of the prediction error and the solution of prediction error method (PEM). Hence, one can also denote the ML estimation by $\mathrm{g}^{\mathrm{LS}}$. One should note that when $\Phi$ is full column rank, then $\mathrm{g}^{\mathrm{ML}}$ or $\mathrm{g}^{\mathrm{LS}}$ equals to $(\Phi^{\mathsf{T}}\Phi)^{-1}\Phi^{\mathsf{T}}\mathrm{y}$.

Though the ML or the LS estimation is an unbiased method, it has a number of drawbacks. The main disadvantage of this approach is the requirement of having a large set of high quality data for obtaining an improved decent estimation $\mathrm{g}^{\mathrm{ML}}$ by solving problem (10). More precisely, the condition number of $\Phi^{\mathsf{T}}\Phi$ should not be large. Otherwise, the estimation can be non-unique or subject to high variance, specifically for large $n_{\mathrm{g}}$. Moreover, choosing a suitable model order, here the length of impulse response g, is a critical and difficult procedure. This is commonly performed by cross validation, model validation techniques or model complexity criteria such as Akaike information criterion (AIC) or Bayesian information criterion (BIC) (Ljung, 1999). These classical strategies are not desirably reliable (Pillonetto et al., 2014; Chen, 2018).

### 3.2 Maximum A Posterior and Regularized Methods

In order to estimate the impulse response, one may use a *maximum a posteriori* (MAP) approach by introducing a prior for the impulse response. More precisely, we set a Gaussian prior for the impulse response as

$$\mathrm{g} \sim \mathcal{N}(0, \mathrm{R}^{-1}), \tag{11}$$

where $\mathrm{R}^{-1}$ is a positive definite matrix. The covariance $\mathrm{R}^{-1}$ can encode available prior knowledge and desired features for the latent impulse response, such as smoothness and stability. Once the prior is set, using Bayes rule, one has

$$p(\mathrm{g}|\Phi, \mathrm{y}) = \frac{p(\mathrm{y}|\Phi, \mathrm{g})p(\mathrm{g})}{p(\Phi, \mathrm{y})}, \tag{12}$$

and the MAP estimation of g is derived by

$$\mathrm{g}^{\mathrm{MAP}} = \underset{\mathrm{g}\in\mathbb{R}^{n_\mathrm{g}}}{\operatorname{argmax}}\ p(\mathrm{g}|\Phi, \mathrm{y}) = \underset{\mathrm{g}\in\mathbb{R}^{n_\mathrm{g}}}{\operatorname{argmax}}\ p(\mathrm{y}|\Phi, \mathrm{g})p(\mathrm{g}). \tag{13}$$

Since $\mathrm{y} - \Phi\mathrm{g} \sim \mathcal{N}(0, \sigma_w^2\mathbb{I})$ and given prior, equations (7) and

$$p(\mathrm{g}) = (2\pi)^{-\frac{n}{2}} \det(\mathrm{R})^{\frac{1}{2}} \exp(-\frac{1}{2}\mathrm{g}^\mathsf{T}\mathrm{R}\mathrm{g}), \tag{14}$$

hold. Subsequently, it is concluded that

$$\mathrm{g}^{\mathrm{MAP}} = \underset{\mathrm{g}\in\mathbb{R}^{n_\mathrm{g}}}{\operatorname{argmax}}\ \exp(-\frac{1}{2\sigma_w^2}\|\mathrm{y} - \Phi\mathrm{g}\|^2 - \frac{1}{2}\mathrm{g}^\mathsf{T}\mathrm{R}\mathrm{g}). \tag{15}$$

Due to the monotonicity of logarithmic function, it follows that

$$\begin{aligned} \mathrm{g}^{\mathrm{MAP}} &= \underset{\mathrm{g}\in\mathbb{R}^{n_\mathrm{g}}}{\operatorname{argmin}}\ \frac{1}{2\sigma_w^2}\|\mathrm{y} - \Phi\mathrm{g}\|^2 + \frac{1}{2}\mathrm{g}^\mathsf{T}\mathrm{R}\mathrm{g}, \\ &= (\Phi^\mathsf{T}\Phi + \sigma_w^2\mathrm{R})\Phi^\mathsf{T}\mathrm{y}. \end{aligned} \tag{16}$$

From (16), one can see that $\mathrm{g}^{\mathrm{MAP}}$ is the solution of a *regularized least square* (RLS) problem. The method is therefore called *regularized system identification* and one may alternatively denote this estimation by $\mathrm{g}^{\mathrm{RLS}}$. For this reason, the covariance matrix R is also called the *regularization matrix*.

The main role of regularization is to include the additionally available prior information in the estimation. In addition, it can also alleviate the issue of possible high variance. As a result, choosing a suitable regularization matrix R has a significant impact on the estimation. This is commonly done in two steps. In the first step, a family of regularization matrices, denoted by $\mathcal{R}$, is chosen. Usually, each element of $\mathrm{R}_\eta$ is a regularization matrix parameterized with a vector of scalars called *hyperparameters*, denoted by $\eta$, which belongs to a given set $\mathscr{E} \subset \mathbb{R}^{n_\eta}$. In other words, we have $\mathcal{R} = \{\mathrm{R}_\eta \mid \eta \in \mathscr{E}\} \subseteq \mathbb{S}_+^{n_\eta}$. In the literature (Pillonetto et al., 2014), the set of regularization matrices are generally defined based on the notion of the kernels,

$$\begin{aligned} \mathcal{R}_{\boldsymbol{k}} = \{\mathrm{R}_\eta \in \mathbb{S}_+^{n_\mathrm{g}} \mid\ & \mathrm{R}_{\eta,ij} = \boldsymbol{k}_\eta(i-1, j-1), \\ & 1 \le i, j \le n_\mathrm{g},\ \eta \in \mathscr{E}\}, \end{aligned} \tag{17}$$

where $\boldsymbol{k}_\eta : \mathbb{Z}_+ \times \mathbb{Z}_+$ is a *positive-definite kernel* and $\eta \in \mathscr{E}$ is the vector of hyperparameters (Pillonetto et al., 2014). The most common kernels are *tuned/correlated* (TC), *diagonal/correlated* (DC) and *stable spline* (SS) (Pillonetto et al., 2014). In the literature, many other kernels and also regularization matrices are designed in various approaches, e.g., inspired by machine learning and system theory (Chen, 2018), using the harmonic analysis of stochastic processes (Zorzi and Chiuso, 2018), based on filter-design methods (Marconato et al., 2016). Given a set of regularization matrices $\mathcal{R}$, one can build a larger set of regularization matrices by considering the convex cone of $\mathcal{R}$. Also, Minkowski sum of sets of regularization matrices produces set of regularization matrices. Based on these ideas and similar ones, algorithms for designing the kernel and regularization have been developed (Hong et al., 2018; Chen et al., 2014, 2018). Once the set of regularization matrices is fixed, one should estimate the hyperparameters using the given data $\mathcal{D}$. In this regard, a number of approaches are introduced for hyperparameter estimation, e.g. empirical Bayes (EB) (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).

Compared to the ML or the LS estimation method, the regularized method have a satisfactory bias-variance trade-off and requires less amount of data. Moreover, the issue of model order selection is alleviated, specially when the regularization matrix is determined only by a few number of hyperparameters. However, when the number of hyperparameters is significantly large, the hyperparameters estimation methods are prone to high-variance or become computationally intractable. Motivated by this issue, we introduce a new method framed in a full Bayesian setting.

## 4. HIERARCHICAL BAYESIAN APPROACH

In this section, we introduce the *hierarchical Bayesian* method for joint estimation of the impulse response and the regularization matrix.

Let the noise model be the same as in Section 3. Moreover, given positive definite matrix R, take the prior distribution of impulse response as

$$\mathrm{g} \sim \mathcal{N}(0, \mathrm{R}^{-1}), \tag{18}$$

Now, let $p_\mathrm{R}$ be a given probability distribution defined over $\mathbb{S}_+^{n_\mathrm{g}}$, called the *hyperprior*, and $\mathcal{R}$ be the support of $p_\mathrm{R}$, i.e.,

$$\mathcal{R} := \operatorname{supp}(p_\mathrm{R}) = \{\mathrm{R} \in \mathbb{S}_+^{n_\mathrm{g}} \mid p_\mathrm{R}(\mathrm{R}) > 0\}. \tag{19}$$

Assume that almost surely the realizations of $p_\mathrm{R}$ are in $\mathbb{S}_{++}^{n_\mathrm{g}}$. This ensures that the distribution (18) is well defined. From the Bayes rule, one has

$$p(\mathrm{g}, \mathrm{R}|\Phi, \mathrm{y}) = \frac{p(\mathrm{y}|\Phi, \mathrm{g})p(\mathrm{g}|\mathrm{R})p_\mathrm{R}(\mathrm{R})}{p(\Phi, \mathrm{g})}. \tag{20}$$

Consequently, the joint maximum a posteriori estimation of g and R is derived as

$$\begin{aligned} (\mathrm{g}^{\mathrm{HB}}, \mathrm{R}^{\mathrm{HB}}) &= \operatorname{argmax}_{\mathrm{g}\in\mathbb{R}^{n_\mathrm{g}}, \mathrm{R}\in\mathcal{R}}\ p(\mathrm{g}, \mathrm{R}|\Phi, \mathrm{y}) \\ &= \operatorname{argmax}_{\mathrm{g}\in\mathbb{R}^{n_\mathrm{g}}, \mathrm{R}\in\mathcal{R}}\ p(\mathrm{y}|\Phi, \mathrm{g})p(\mathrm{g}|\mathrm{R})p_\mathrm{R}(\mathrm{R}). \end{aligned}$$

From (7), (18), the prior distribution $p_\mathrm{R}$ and the monotonicity of logarithm function, it follows that

$$\begin{aligned} (\mathrm{g}^{\mathrm{HB}}, \mathrm{R}^{\mathrm{HB}}) = \underset{\mathrm{g}\in\mathbb{R}^{n_\mathrm{g}}, \mathrm{R}\in\mathcal{R}}{\operatorname{argmin}}\ & \frac{1}{2\sigma_w^2}\|\mathrm{y} - \Phi\mathrm{g}\|^2 - \frac{1}{2}\ln\det\mathrm{R} \\ & + \frac{1}{2}\mathrm{g}^\mathsf{T}\mathrm{R}\mathrm{g} - \ln\, p_\mathrm{R}(\mathrm{R}). \end{aligned} \tag{21}$$

A suitable choice of $p_\mathrm{R}$ has two significant impacts: the correctness of the outcome of the estimation problem and also the numerical tractability of optimization problem (21). Regarding the former issue, one may suggest using a Wishart distribution with suitably chosen *scale matrix*

and *degree of freedom* (Wishart, 1928). However, the latter issue is still a case of concern since of the large number of tuning parameters in the definition of $p_{\mathrm{R}}$.

Tractability of problem (21) depends on the structure of $-\ln p_{\mathrm{R}}$ and therefore, it is desired to introduce a suitable structure for this distribution. The elements of $\mathcal{R}$ are positive semi-definite matrices. Therefore, they have a Cholesky decomposition (Horn and Johnson, 2012). Motivated by this fact, we propose parameterization of the matrix factors in Cholesky decomposition. More precisely, let $n_\eta \in \mathbb{N}$ and $\mathrm{U}_1, \ldots, \mathrm{U}_{n_\eta} \in \mathbb{U}_{++}^{n_{\mathrm{g}}}$ be linearly independent matrices. Define set $\mathcal{U}$ as the convex cone of these matrices, i.e., we have

$$\mathcal{U} = \Big\{ \sum_{i=1}^{n_\eta} \eta_i \mathrm{U}_i \mid \eta_i \geq 0, \quad \forall i = 1, \ldots, n_\eta \Big\}. \tag{22}$$

Denote by $\eta$ as the vector defined by $[\eta_1, \ldots, \eta_{n_\eta}]^{\mathsf{T}}$ in (22). Since $\mathrm{U}_1, \ldots, \mathrm{U}_{n_\eta}$ are linearly independent, it can be seen from (22) that each element of $\mathcal{U}$ uniquely determines a vector $\eta$ in $\mathbb{R}_+^{n_\eta}$ and each $\eta \in \mathbb{R}_+^{n_\eta}$ uniquely identify an element of $\mathcal{U}$. Moreover, we know that $\{\mathrm{U}^{\mathsf{T}}\mathrm{U} |\ \mathrm{U} \in \mathcal{U}\} \backslash \{\mathbb{0}\} \subseteq \mathbb{S}_{++}^{n_{\mathrm{g}}}$. Let assume $p_\eta$ be a probability distribution on $\mathbb{R}_+^{n_\eta}$, i.e., $\mathrm{supp}(p_\eta) \subseteq \mathbb{R}_+^{n_\eta}$. Consequently, one can introduce a corresponding distribution $p_{\mathrm{R}}$ on $\mathcal{R} := \{\mathrm{U}^{\mathsf{T}}\mathrm{U} |\ \mathrm{U} \in \mathcal{U}\}$ as

$$p_{\mathrm{R}}(\mathrm{R}) = p_{\mathrm{R}}(\mathrm{U}^{\mathsf{T}}\mathrm{U}) = p_\eta(\eta), \tag{23}$$

where $\mathrm{R} = \mathrm{U}^{\mathsf{T}}\mathrm{U}$ and $\mathrm{U} = \sum_{i=1}^{n_\eta} \eta_i \mathrm{U}_i$. Note that these decompositions are well-defined and unique due to properties of Cholesky decomposition and linear independency of $\mathrm{U}_1, \ldots, \mathrm{U}_{n_\eta}$. From the Bayes rule, we can derive a MAP estimation as

$$\begin{aligned}(\mathrm{g}^{\mathrm{HB}}, \eta^{\mathrm{HB}}) &= \mathrm{argmax}_{\mathrm{g}\in\mathbb{R}^{n_{\mathrm{g}}}, \eta\in\mathbb{R}_+^{n_\eta}} \ \ p(\mathrm{g}, \eta | \Phi, \mathrm{y}), \\ &= \mathrm{argmax}_{\mathrm{g}\in\mathbb{R}^{n_{\mathrm{g}}}, \eta\in\mathbb{R}_+^{n_\eta}} \ p(\mathrm{y}|\Phi, \mathrm{g}) p(\mathrm{g}|\mathrm{U}) p_\eta(\eta).\end{aligned}$$

Define function $\tilde{J} : \mathbb{R}^{n_{\mathrm{g}}} \times \mathbb{R}_+^{n_\eta} \to \mathbb{R}$ as

$$\tilde{J}(\mathrm{g}, \eta) = J(\mathrm{g}, \sum_{i=1}^{n_\eta} \eta_i \mathrm{U}_i). \tag{24}$$

Since, $\mathrm{U}_1, \ldots, \mathrm{U}_{n_\eta}$ are upper triangular, one can see that

$$\ln \det \sum_{i=1}^{n_\eta} \eta_i \mathrm{U}_i = \ln \prod_{j=1}^{n_{\mathrm{g}}} \sum_{i=1}^{n_\eta} \eta_i \mathrm{U}_{i,jj} = \sum_{j=1}^{n_{\mathrm{g}}} \ln \sum_{i=1}^{n_\eta} \eta_i \mathrm{U}_{i,jj},$$

where $\mathrm{U}_{i,jj}$ is the $j$-th element on the diagonal of $\mathrm{U}_i$, for $i = 1, \ldots, n_\eta$ and $j = 1, \ldots, n_{\mathrm{g}}$. Therefore, we have

$$\begin{aligned}\tilde{J}(\mathrm{g}, \eta) = &\tfrac{1}{2\sigma_w^2} \|\mathrm{y} - \Phi \mathrm{g}\|^2 - \textstyle\sum_{j=1}^{n_{\mathrm{g}}} \ln \sum_{i=1}^{n_\eta} \eta_i \mathrm{U}_{i,jj} \\ &+ \tfrac{1}{2} \| \textstyle\sum_{i=1}^{n_\eta} \eta_i \mathrm{U}_i \mathrm{g} \|^2 - \ln p_\eta(\eta).\end{aligned} \tag{25}$$

Subsequently, one can see that

$$(\mathrm{g}^{\mathrm{HB}}, \eta^{\mathrm{HB}}) = \mathrm{argmin}_{\mathrm{g}\in\mathbb{R}^{n_{\mathrm{g}}}, \eta\in\mathbb{R}_+^{n_\eta}} \ \tilde{J}(\mathrm{g}, \eta). \tag{26}$$

Potential candidates for $p_\eta$ are distributions from family of exponential distributions like

$$p_\eta(\eta) = A(\lambda_1, \lambda_2) \exp(-\lambda_1 \sum_{i=1}^{n_\eta} \eta_i - \lambda_2 \sum_{i=1}^{n_\eta} \eta_i^2) \mathbf{1}_{\mathbb{R}_+^{n_\eta}}(\eta), \tag{27}$$

where $\lambda_2 \geq 0$ and $A(\lambda_1, \lambda_2)$ is the normalizing coefficient. Therefore, for $\eta \in \mathbb{R}_+^{n_\eta}$, it follows that

$$-\ln p_\eta(\eta) = -\ln A(\lambda_1, \lambda_2) - \lambda_1 \sum_{i=1}^{n_\eta} \eta_i - \lambda_2 \sum_{i=1}^{n_\eta} \eta_i^2. \tag{28}$$

**Algorithm 1** Sequential Convex Programming for (29)

1: **Input:** y, $\Phi$, $\mathrm{U}_1, \ldots, \mathrm{U}_{n_\eta}$, $\lambda_1$, $\lambda_2$ and initial $\eta^{(0)}$.
2: $k \leftarrow 0$.
3: **while** stopping condition is not met **do**
4: $\quad \mathrm{U}^{(k)} \leftarrow \sum_{i=1}^{n_\eta} \eta_i^{(k)} \mathrm{U}_i$.
5: $\quad$ Compute
$$\mathrm{g}^{(k)} = (\Phi^{\mathsf{T}}\Phi + \sigma_w^2 \mathrm{U}^{(k)\mathsf{T}}\mathrm{U}^{(k)})^{-1} \Phi^{\mathsf{T}} \mathrm{y}.$$
6: $\quad$ Solve the following convex program:
$$\begin{aligned}\eta^{(k+1)} = \mathrm{argmin}_{\eta\in\mathbb{R}_+^{n_\eta}} \ &\frac{1}{2} \| \sum_{i=1}^{n_\eta} \eta_i \mathrm{U}_i \mathrm{g}^{(k+1)} \|^2 \\ &- \sum_{j=1}^{n_{\mathrm{g}}} \ln \sum_{i=1}^{n_\eta} \eta_i \mathrm{U}_{i,jj} + \lambda_1 \sum_{i=1}^{n_\eta} \eta_i + \lambda_2 \sum_{i=1}^{n_\eta} \eta_i^2.\end{aligned}$$
7: $\quad k \leftarrow k + 1$.
8: **end**
9: **Output:** $(\mathrm{g}^{\mathrm{HB}}, \eta^{\mathrm{HB}}, \mathrm{U}^{\mathrm{HB}}, \mathrm{R}^{\mathrm{HB}})$.

It is worth mentioning that for this choice of $p_\eta$, we have a lasso penalty function if $\lambda_1 > 0$ and $\lambda_2 = 0$, a ridge penalty function if $\lambda_1 = 0$ and $\lambda_2 > 0$, and an elastic-net penalty function if $\lambda_1 > 0$ and $\lambda_2 > 0$. Also, for particular choices of $\lambda_1$ and $\lambda_2$, one has a *maximum entropy* prior on $\eta$ (Dowson and Wragg, 1973). Moreover, since the distribution $p_\eta$ is integrable, one should have $\lambda_2 \geq 0$, and therefore, the loss function is convex. When $\lambda_2 > 0$, the loss function is strongly convex.

It is noteworthy that (26) reduces to

$$\begin{aligned}(\mathrm{g}^{\mathrm{HB}}, \eta^{\mathrm{HB}}) = \mathrm{argmin}_{\mathrm{g}\in\mathbb{R}^{n_{\mathrm{g}}}, \eta\in\mathbb{R}_+^{n_\eta}} \ &\frac{1}{2\sigma_w^2} \|\mathrm{y} - \Phi \mathrm{g}\|^2 - \sum_{j=1}^{n_{\mathrm{g}}} \ln \sum_{i=1}^{n_\eta} \eta_i \mathrm{U}_{i,jj} \\ &+ \frac{1}{2} \| \sum_{i=1}^{n_\eta} \eta_i \mathrm{U}_i \mathrm{g} \|^2 + \lambda_1 \sum_{i=1}^{n_\eta} \eta_i + \lambda_2 \sum_{i=1}^{n_\eta} \eta_i^2.\end{aligned} \tag{29}$$

*Theorem 4.1.* i) Let $\eta^{(0)}$ be a non-zero vector in $\mathbb{R}_+^{n_\eta}$. Then, $\tilde{J}(\cdot, \mathrm{U}_0) : \mathbb{R}^{n_{\mathrm{g}}} \to \mathbb{R}$ is a strongly convex quadratic function and

$$\mathrm{argmin}_{\mathrm{g}\in\mathbb{R}^{n_{\mathrm{g}}}} \ J(\mathrm{g}, \eta^{(0)}) = (\Phi^{\mathsf{T}}\Phi + \sigma_w^2 \mathrm{U}_0^{\mathsf{T}}\mathrm{U}_0)^{-1} \Phi^{\mathsf{T}} \mathrm{y}, \tag{30}$$

where $\mathrm{U}_0 = \sum_{i=1}^{n_\eta} \eta_i^{(0)} \mathrm{U}_i$.
ii) Let $p_{\mathrm{U}}$ be given as in (27). Then, for any $\mathrm{g}_0 \in \mathbb{R}^{n_{\mathrm{g}}}$, the function $\tilde{J}(\mathrm{g}_0, \cdot) : \mathbb{R}_+^{n_\eta} \to \mathbb{R}$ is a proper strictly convex function and the optimization problem

$$\mathrm{min}_{\eta\in\mathbb{R}_+^{n_\eta}} \ \tilde{J}(\mathrm{g}_0, \eta) \tag{31}$$

is a convex optimization problem with a unique solution. Moreover, if $\lambda_2 > 0$, it is a strongly convex function.

*Corollary 4.2.* The optimization problem (29) is a bi-convex (multi-convex) programming.

*Remark 4.3.* Due to the Theorem 4.1, we can introduce a *sequential convex programming* approach for solving (29). The details of this procedure presented in Algorithm 1.

*Remark 4.4.* The stopping condition in Algorithm 1 can be a combination of maximum number of iterations, a cross-validation error and reaching to a preset minimum step length or reaching a predetermined minimum improvement of objective function.

## 5. NUMERICAL EXPERIMENT

In this section, the approach is numerically assessed, with main focus on Algorithm 1. To this end, it is required to introduce the settings of optimization, the test data-bank and the numerical setup.

### *5.1 Generics of Optimization Settings*

As the first step, it is required to choose $\mathrm{U}_1, \ldots, \mathrm{U}_{n_\eta}$. In this regard, we consider TC-kernels, denoted by $\boldsymbol{k}^{(\mathrm{TC})}$ : $\mathbb{Z}_+ \times \mathbb{Z}_+ \to \mathbb{R}_+$ and defined as

$$\boldsymbol{k}_\alpha^{(\mathrm{DC})}(t_1, t_2) := \alpha^{\max(t_1, t_2)}, \quad \forall\, t_1, t_2 \in \mathbb{Z}_+, \tag{32}$$

where $\alpha$ is a real scalar in $(0, 1)$. Accordingly, one can define a regularization matrix $\mathrm{R}_\alpha \in \mathbb{S}_{++}^{n_\mathrm{g}}$ such that the entry of $\mathrm{R}_\alpha^{-1}$ at location $(t_1, t_2)$ is $\boldsymbol{k}_\alpha^{(\mathrm{DC})}(t_1 - 1, t_2 - 1)$, for any $t_1, t_2 = 1, \ldots, n_\mathrm{g}$. Since $\mathrm{R}_\alpha^{-1}$ is positive definite, it has a unique Cholesky decomposition (Marconato et al., 2016) as $\mathrm{R}_\alpha^{-1} = \mathrm{U}_\alpha^\mathsf{T} \mathrm{U}_\alpha$ where $\mathrm{U}_\alpha$ is defined entry-wise by

$$\mathrm{U}_{\alpha,st} = \begin{cases} \alpha^{\frac{1-s}{2}}(1-\alpha)^{-\frac{1}{2}}, & \text{if } s = t, s < n_\mathrm{g}, \\ \alpha^{\frac{1-s}{2}}, & \text{if } s = t = n_\mathrm{g}, \\ -\alpha^{\frac{1-s}{2}}(1-\alpha)^{-\frac{1}{2}}, & \text{if } s = t-1, \\ 0, & \text{otherwise.} \end{cases} \tag{33}$$

Given $\alpha_1, \ldots, \alpha_{n_\eta}$, we can set $\mathrm{U}_i$ as $\mathrm{U}_{\alpha_i}$, for $i = 1, \ldots, n_\eta$. Regarding $\lambda_1$ and $\lambda_2$, we consider two scenarios:

- $\lambda_1 = 0$, i.e., $\ell_2$ or ridge penalty function, and,
- $\lambda_2 = 0$, i.e., $\ell_1$ or lasso penalty function.

In each of these scenarios, the non-zero parameter can be tuned based on a cross-validation procedure. The initial $\eta$, i.e. $\eta^{(0)}$, can be chosen either randomly or by solving (31), where $\mathrm{g}^{(0)}$ is taken as the estimated impulse response obtained with another regularized identification method like TC-kernel regularization. The latter approach is used here. The final step for utilizing Algorithm 1 is specifying $\Phi$ and y which are discussed in the following section.

### *5.2 Test Data-Bank*

Define transfer functions $G(z)$ as

$$G(z) = \frac{z^3 + 0.5z^2}{z^4 - 2.2z^3 + 2.42z^2 - 1.87z + 0.7225}. \tag{34}$$

which are initially introduced in (Wahlberg and Ljung, 1986, Example 5.1) and also utilized in (Pillonetto and De Nicolao, 2010) as benchmark systems. To have a fair comparison, the systems are normalized by their $\mathcal{H}_2$ norm. The input signal consists of a pseudo-random binary signal (PRBS) of length $n_\mathcal{D} = 63$ for time instants $t = 0, \ldots, n_\mathcal{D} - 1$. Moreover, it is assumed that the systems are initially at rest. At the output of each system, a white additive Gaussian noise with distribution $\mathcal{N}(0, \sigma_w^2)$ is added. Three values of $\sigma_w^2$ are considered, namely 0.001, 0.01, 0.1. For each of them 100 realizations of noise are generated and subsequently, 100 data set (3) are obtained by simulation.

### *5.3 Numerical Experiment Setup*

In the numerical experiment, three impulse response estimation methods are tested:

1) regularized impulse response estimation using TC kernel,
2) hierarchical Bayes with $\ell_1$ penalty function, and
3) hierarchical Bayes with $\ell_2$ penalty function.

For the cases 2) and 3), we take 51 upper triangular matrices, $\{\mathrm{U}_i \mid i = 1, \ldots, 51\}$, defined according to (33). Here, $\alpha_1, \ldots, \alpha_{51}$ are taken equidistantly from interval $[0.8, 0.995]$. The hyperparameters $\lambda_1$ and $\lambda_2$ are chosen based on a 85%-15% cross-validation scheme. The Algorithm 1 stops when either the number of iterations reaches 50 or when $\|(\mathrm{g}^{(k)}, \eta^{(k)}) - (\mathrm{g}^{(k-1)}, \eta^{(k-1)})\| \leq 10^{-4}$.

The measure of fit is defined (Ljung, 2012) as

$$\mathrm{Fit} = 100 \times \left(1 - \frac{\|\mathrm{g} - \hat{\mathrm{g}}\|}{\|\mathrm{g} - \bar{\mathrm{g}}\|}\right), \tag{35}$$

where g and $\hat{\mathrm{g}}$ are respectively the true and the estimated impulse responses, and $\bar{\mathrm{g}}$ is a vector of the same length as g and with entries identically equal to $n_\mathrm{g}^{-1} \sum_{k=0}^{n_\mathrm{g}-1} g_k$. The results of the estimation is shown in Figure 1 and the statistics are shown in Table 1.

### *5.4 Discussion*

From the numerical results, one can see that the algorithm shows an improved performance, especially when the variance of noise is not significantly high. Additionally, one should note that in all of the cases, the hierarchical Bayes method has smaller mean square error. The performance of $\ell_1$ penalty is slightly better comparing to $\ell_2$ penalty, however, more numerical analysis is required for providing a more solid argument.

It is worth noting that the optimization problem (29) is a non-convex program and thus can be hampered by the known issue of local minima. This revealed itself in the numerical results in form of outliers (Figure 1, red crosses). However, its effect can be alleviated by performing suitable cross-validation and invalidating the local minima. In fact, when the variance of noise is high, the cross validation does not have enough fidelity and the local minima issue becomes more sever in this case. One possible solution can be utilizing global optimization heuristics.

## 6. CONCLUSION

A new hierarchical Bayesian method is proposed in this paper. This essentially consists of a maximum a posteriori (MAP) approach for joint estimation of impulse response and regularization matrix. The crucial step is choosing an appropriate hyperprior. To this end, the Cholesky decomposition is used for decomposing the regularization matrix and then the prior is introduced on a suitable sub-cone of the cone of upper triangular matrices with positive diagonal entries. By doing so, one can specify the prior simply on the positive orthant of the space of hyperparameters. Utilizing suitable distributions from the exponential family, the estimation is reduced to a tractable optimization problem with multi-convex structure. Efficient sequential convex programming algorithms are proposed to find optimal solutions. Due to the non-convexity, reaching global optimal solution cannot be guaranteed. The approach is verified numerically on benchmark examples and compared, as an example, with the results

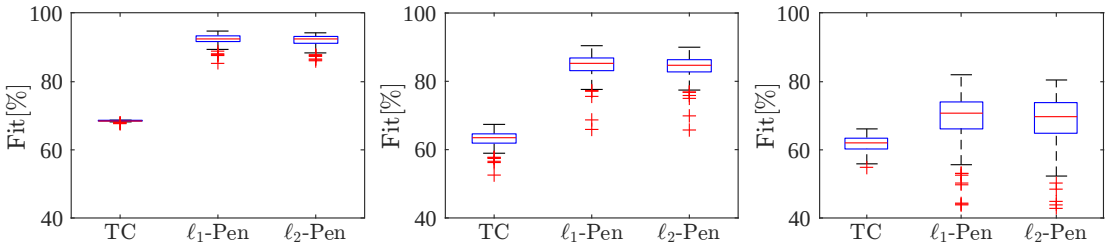


Figure 1. Comparison of fitting performance under different noise levels of $\sigma_w^2 = 0.001$ (left), $\sigma_w^2 = 0.01$ (middle) and $\sigma_w^2 = 0.1$ (right), for system $G(z)$ given in (34).

of TC-kernel. The numerical results show the significant performance of the proposed method.

## REFERENCES


Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3), 337–404.

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.

Dowson, D. and Wragg, A. (1973). Maximum-entropy distributions having prescribed first and second moments (corresp.). *IEEE Transactions on Information Theory*, 19(5), 689–693.

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.

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.

Horn, R.A. and Johnson, C.R. (2012). *Matrix analysis*. Cambridge university press.

Khosravi, M., Yin, M., Iannelli, A., Parsi, A., and Smith, R.S. (2020). Low-complexity identification by sparse hyperparameter estimation. *IFAC-papersonline*.

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

Ljung, L. (2012). System identification toolbox: for use with matlab.

Marconato, A., Schoukens, M., and Schoukens, J. (2016). Filter-based regularisation for impulse response modelling. *IET Control Theory & Applications*, 11(2), 194–204.

Mohan, K. and Fazel, M. (2010). Reweighted nuclear norm minimization with application to system identification. In *Proceedings of the 2010 American Control Conference*, 2953–2959. IEEE.

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 *2018 IEEE Conference on Decision and Control (CDC)*, 644–649. IEEE.

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 *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, 6265–6270. IEEE.

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.

Wahlberg, B. and Ljung, L. (1986). Design variables for bias distribution in transfer function estimation. *IEEE Transactions on Automatic Control*, 31(2), 134–144.

Wishart, J. (1928). The generalised product moment distribution in samples from a normal multivariate population. *Biometrika*, 32–52.

Zorzi, M. and Chiuso, A. (2018). The harmonic analysis of kernel functions. *Automatica*, 94, 125–137.


| $\sigma_w^2 = 0.001$ | TC | $\ell_1$ | $\ell_2$ |
|---|---|---|---|
| Bias$^2$ [$\times 10^4$] | 916 | 16.2 | 12.7 |
| Var [$\times 10^4$] | 9.34 | 40.3 | 49.1 |
| MSE [$\times 10^4$] | 925 | **56.5** | **61.8** |
| $\sigma_w^2 = 0.01$ | TC | $\ell_1$ | $\ell_2$ |
| Bias$^2$ [$\times 10^3$] | 120.8 | 3.33 | 2.88 |
| Var [$\times 10^3$] | 6.68 | 20.1 | 22.0 |
| MSE [$\times 10^3$] | 127 | **23.4** | **24.9** |
| $\sigma_w^2 = 0.1$ | TC | $\ell_1$ | $\ell_2$ |
| Bias$^2$ [$\times 10^3$] | 114 | 11.9 | 11.9 |
| Var [$\times 10^3$] | 20.9 | 85.3 | 94.2 |
| MSE [$\times 10^3$] | 135 | **97.2** | **106** |

Table 1. The table provides the statistics of estimation performance. One can see that the proposed method shows better bias-variance trade-off.