2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

A Family of Hyper-parameter Estimators for Regularized Linear System Identification

Meng Zhang, Tianshi Chen, and Biqiang Mu

Abstract— **Hyper-parameter estimation is one of the fundamental issues for kernel-based regularized system identification methods. Empirical Bayes (EB) estimator and Stein's unbiased risk estimator (SURE) are two popular hyper-parameter estimators, but they both have advantages and disadvantages. Specifically, EB is not asymptotically optimal in the mean squared error (MSE) sense but SURE is, while SURE is more sensitive to ill-conditioned regression matrix but EB is more robust. In this paper, to find a better estimator by combining their strength and mitigating their weakness, we propose a family of hyper-parameter estimators by linking EB and SURE estimators together through an index. The finite sample and asymptotic properties of this family of estimators have been established. The Monte Carlo simulation results show that there does exist a 'middle' hyper-parameter estimator in this family that is superior to the EB and SURE.**

Index Terms— **Linear system identification, Kernel-based regularization method, hyper-parameter estimator, empirical Bayes estimator, Stein's unbiased risk estimator.**

I. Introduction

Kernel-based regularization methods (KRMs) for system identification, which were initially introduced in [1] and further developed in [2]–[6], have attracted increasing interest in the system identification community during the past decade. It has been a supplement to standard maximum likelihood/prediction error methods (ML/PEM) [7] and has gradually evolved into a new paradigm for system identification [8]. The essential idea of the KRM is to incorporate prior knowledge of the system to be identified into the kernel. It therefore consists of two successive procedures: kernel design and hyper-parameter estimation, which are analogous to model structure design and model order selection in the ML/PEM, respectively.

For the kernel design, it involves parameterizing the kernel with hyper-parameter based on the prior knowledge of the system. Various kernels with different kinds of prior knowledge have been developed for the KRM , such as the stable spline kernel (SS) [1], and the diagonal correlated (DC) kernel, and the tuned-corrected (TC) kernel [2]. In addition, in terms of the system theory and the machine learning, two systematic techniques have been proposed in [9], respectively. For hyper-parameter estimation, it involves estimating the hyper-parameters based on the observed data. Comparing to the ML/PEM, it tunes the model complexity in a continuous way, which can achieve a better bias-variance trade-off. Many methods have been provided for hyper-parameter estimation in [3], such as empirical Bayes (EB) estimator, Stein's unbiased risk estimator (SURE) [3], [4], [6]. In order to understand the behavior of these estimators, there have been many results for hyper-parameter estimation, such as the robustness and the mean squared error (MSE) properties of the hyper-parameter estimator [4], [10], the asymptotic properties of hyper-parameter estimators [6], [11]–[15], the influence of ill-conditioned regression matrix on hyper-parameter estimators [16].

The motivation of this paper stems from the theoretical and empirical results of the EB and SURE estimators, which can be found in [2], [4]–[6], [10]–[13], [17], [18]. In theory, SURE is asymptotically optimal in MSE sense but EB is generally not. In practice, SURE is more sensitive to ill-conditioned regression matrix and/or short data but EB is more robust. In fact, the C_p statistics and the generalized maximum likelihood (GML) for function smoothing also exhibits this kind of behavior, see e.g. [19]. To find a better estimator by combining the strength and mitigating the weakness of the EB and SURE estimators, we propose a family of hyper-parameter estimators by linking EB and SURE estimators together through an index, where the EB and SURE estimators correspond to the two end points of this family. To understand this family of estimators, we first show the decomposition of each term of the estimation criterion and then derive the first-order optimality conditions for a special case, which sheds light on how the index influences this family of estimators. Then we investigate the almost sure convergence of the estimation criterion and the corresponding hyper-parameter estimator. Finally, we run Monte Carlo simulations to show that there does exist a 'middle' hyper-parameter estimator in this family that is superior to the EB and SURE estimators.

The rest of the paper is organized as follows. In Section II, we introduce the KRM for linear system identification. In Section III, we introduce a family of hyper-parameter estimators by linking EB and SURE estimators together

This work was supported in part by the National Key R&D Program of China under contract No. 2022YFA1004700, the NSFC under contract No. 62273287, by the Shenzhen Science and Technology Innovation Council under contract No. JCYJ20220530143418040 and JCY20170411102101881, the Thousand Youth Talents Plan funded by the central government of China.

Meng Zhang was with the Key Laboratory of Systems and Control of CAS, Institute of Systems Science, Academy of Mathematics and System Science, Chinese Academy of Sciences, Beijing 100190, China. She is now with the School of Data Science and Shenzhen Research Institute of Big Data, The Chinese University of Hong Kong, Shenzhen, 518172, China (email: zhangmeng2020@amss.ac.cn)

Tianshi Chen is with the School of Data Science and Shenzhen Research Institute of Big Data, The Chinese University of Hong Kong, Shenzhen, 518172, China (email: tschen@cuhk.edu.cn)

Biqing Mu is with the Key Laboratory of Systems and Control of CAS, Institute of Systems Science, Academy of Mathematics and System Science, Chinese Academy of Sciences, Beijing 100190, China (email: bqmu@amss.ac.cn)

through an index and study their properties including finite sample and asymptotic properties. In Section IV, a Monte Carlo simulation is provided to illustrate the existence of a superior estimator in this family than both the EB and SURE estimators. In Section V, we conclude this paper.

II. Regularized Linear System Identification

A. Regularized Least Squares

Consider a single-input single-output, stable, causal and discrete-time linear system

$$y(t) = G_0(q)u(t) + v(t),\ t = 1, \cdots, n, \tag{1}$$

where t is the time index, $y(t)$, $u(t)$ and $v(t)$ are the output, input and measurement noise of the system, respectively, q is the forward shift operator($qu(t) = u(t+1)$), $G_0(q)$ is the transfer function of the system, and n is the number of data. We represent the transfer function $G_0(q)$ in the impulse response form:

$$G_0(q) = \sum_{k=1}^{\infty} g_k^0 q^{-k},$$

where $\{g_k^0\}_{k=1}^{\infty}$ are the impulse response coefficients of the system. Thus, the identification problem is to estimate the impulse response coefficients $\{g_k^0\}_{k=1}^{\infty}$ as well as possible based on the available data $\{u(t), y(t)\}_{t=1}^{n}$. Due to the stability of $G_0(q)$, we can truncate the impulse response sequence at a sufficiently high order and then acquire a finite impulse response (FIR) model:

$$G(q) = \sum_{k=1}^{p} g_k q^{-k},\ \theta = [g_1, \cdots, g_p]^T. \tag{2}$$

With FIR model (2), the system (1) can be written as follows:

$$y(t) = \phi^T(t)\theta + v(t),\ t = 1, \cdots, n,$$

where $\phi(t) = [u(t-1), \cdots, u(t-p)]^T$. And it has the following linear regression form:

$$\begin{aligned} &Y = \Phi\theta + V, \\ &Y = [y(1), \cdots, y(n)]^T, \\ &\Phi = [\phi(1), \cdots, \phi(n)]^T, \\ &V = [v(1), \cdots, v(n)]^T. \end{aligned} \tag{3}$$

Hence, our goal is transformed into the estimation of the parameter θ in the linear regression model (3). We make the following assumptions on model (3).

Assumption 1:
1) The dimension p of parameters is fixed.
2) The input sequence $\{u(t)\}_{t=1}^{n}$ is deterministic and $\Phi^T\Phi/n = O(1)$.
3) The noise sequence $\{v(t)\}_{t=1}^{n}$ is a sequence of independent and identically distributed (i.i.d.) random variables with zero mean and finite variance $\sigma^2 > 0$ and is independent of $u(t)$.

The classic method to estimate the parameter θ in linear regression is the least squares (LS), which is

$$\widehat{\theta}^{\mathrm{ls}} = \arg\min_{\theta \in \mathbb{R}^p} \|Y - \Phi\theta\|^2 \tag{4a}$$

$$= (\Phi^T\Phi)^{-1}\Phi^T Y, \tag{4b}$$

where $\|\cdot\|$ is the Euclidean norm. The LS estimator is unbiased, but it might suffer from the large variance problem under some unfavorable conditions, e.g., the input signal is low-passed and the data has low signal-to-noise ratio. To improve it, we introduce the regularized least squares (RLS) estimator by adding a regularization term in the least squares criterion

$$\widehat{\theta}^{\mathrm{rls}} = \arg\min_{\theta \in \mathbb{R}^p} \|Y - \Phi\theta\|^2 + \sigma^2\theta^T K^{-1}\theta \tag{5a}$$

$$= (\Phi^T\Phi + \sigma^2 K^{-1})^{-1}\Phi^T Y \tag{5b}$$

$$= K\Phi^T(\Phi K\Phi^T + \sigma^2 I_n)^{-1}Y, \tag{5c}$$

where matrix K, called the kernel matrix, is symmetric and positive semidefinite. From a Bayesian perspective, estimator (5) is exactly the maximum a posterior mean of θ under the assumption that $\theta \sim \mathcal{N}(0, K)$, $V \sim \mathcal{N}(0, \sigma^2 I_n)$, and θ is independent of V. For a given kernel matrix K, the performance of RLS estimator (5b) is usually evaluated by the following mean squared error (MSE) criteria [2]

$$\mathrm{MSE}(K) = E\|\Phi(\widehat{\theta}^{\mathrm{rls}} - \theta_0)\|^2, \tag{6}$$

where $E(\cdot)$ is the mathematical expectation with respect to the noise distribution and θ_0 is the true parameter vector.

B. Kernel Design

Kernel design is the first procedure of the KRM, which is to parameterize kernel matrix K by a few number of parameters, called the hyper-parameters, denoted by η, based on the prior knowledge, e.g., the stability and smoothness. So far, several common kernels are introduced, e.g., the stable spline (SS) kernel [1] , the diagonal correlated (DC) kernel and the tuned-correlated (TC) kernel [2]:

$$\begin{aligned} &\mathrm{SS}: \quad K_{ij}(\eta) = c\Big(\frac{\lambda^{i+j+\max(i,j)}}{2} - \frac{\lambda^{3\max(i,j)}}{6}\Big) \\ &\quad \eta = [c, \lambda] \in \Omega = \{c \geq 0, 0 \leq \lambda \leq 1\}; \\ &\mathrm{DC}: \quad K_{kj}(\eta) = c\lambda^{(k+j)/2}\rho^{|j-k|} \\ &\quad \eta = [c, \lambda, \rho] \in \Omega = \{c \geq 0, 0 \leq \lambda \leq 1, |\rho| \leq 1\}; \\ &\mathrm{TC}: \quad K_{ij}(\eta) = c\lambda^{\max(i,j)} \\ &\quad \eta = [c, \lambda] \in \Omega = \{c \geq 0, 0 \leq \lambda \leq 1\}. \end{aligned}$$

C. Hyper-parameter Estimation

Once the parameterization of $K(\eta)$ is given, the next procedure is to estimate hyper-parameters η using data $\{u(t), y(t)\}_{t=1}^{n}$. Several popular hyper-parameter estimation methods are Empirical Bayes (EB), Stein's unbiased risk estimators (SURE) and Cross Validation (CV) [1], [3], [4], [12], [13]. Under the assumption that $\theta \sim \mathcal{N}(0, K)$, $V \sim \mathcal{N}(0, \sigma^2 I_n)$, and θ is independent of V, the EB estimator

aims to maximize the marginal likelihood of Y, which is equivalent to minimizing the optimization problem:

$$\text{EB}: \widehat{\eta}_{\text{eb}} = \arg\min_{\eta\in\Omega} \mathscr{C}_{\text{eb}}(K(\eta)), \tag{7a}$$

$$\mathscr{C}_{\text{eb}}(K) = Y^T Q^{-1} Y + \log\det(Q), \tag{7b}$$

where $Q = \Phi K \Phi^T + \sigma^2 I_n$ and $\det(\cdot)$ denotes the determinant of a square matrix. While, the SURE aims to minimize an unbiased estimator of the MSE criterion (6) as follows:

$$\begin{aligned}\mathscr{C}_{\text{sure}}(K) &= \|Y - \Phi\widehat{\theta}^{\text{rls}}\|^2 + 2\sigma^2\text{Tr}(I - \sigma^2 Q^{-1}) \\ &= \sigma^4 Y^T Q^{-2} Y + 2\sigma^2\text{Tr}(I - \sigma^2 Q^{-1}).\end{aligned} \tag{8}$$

Accordingly, the SURE for tuning η is defined by

$$\text{SURE}: \widehat{\eta}_{\text{sure}} = \arg\min_{\eta\in\Omega} \mathscr{C}_{\text{sure}}(K(\eta)). \tag{9}$$

In particular, $\|Y - \Phi\widehat{\theta}^{\text{rls}}\|^2$ characterizes the model fit of the estimated model (5b), and $\text{Tr}(I - \sigma^2 Q^{-1})$ characterizes its model complexity, which is called the degrees of freedom of (5b) [20]. It means that the SURE estimator (9) is to balance the model fit and model complexity of (5b).

III. A Family of hyper-parameter Estimators

In this section, we propose a family of hyper-parameter estimators that links the EB and SURE estimators in a unified way through an index and further investigate the properties of this family including finite sample and asymptotic properties.

First, we introduce the key idea to link the EB and SURE estimators in an unified way by the following lemma.

Lemma 1: Suppose that $n \times n$ matrix A is diagonalizable and has positive eigenvalues. Thus, there holds that

$$\frac{1}{p}\text{Tr}(A^p - I_n) \to \log\det(A) \quad \text{as } p \to 0,$$

where $\text{Tr}(\cdot)$ denotes the trace of a square matrix.

According to Lemma 1, as $p \to 0$ we have

$$\frac{1}{p}\text{Tr}(I_n - (\sigma^2 Q^{-1})^p) \to \log\det(Q) + N\log\sigma^2, \tag{10}$$

which is because the matrix Q^{-1} is symmetric and positive definite. Thus, motivated by (7b) and (8) of the EB and SURE estimators as well as (10), we define the family of hyper-parameter estimators with respect to an index $\alpha \in [1, 2]$ by

$$\widehat{\eta}_\alpha = \arg\min_{\eta\in\Omega} \mathscr{C}_\alpha(K(\eta)), \tag{11a}$$

$$\mathscr{C}_\alpha(K) = \sigma^{2\alpha} Y^T Q^{-\alpha} Y + \frac{\alpha\sigma^2}{\alpha - 1}\text{Tr}\Big(I_n - (\sigma^2 Q^{-1})^{\alpha-1}\Big). \tag{11b}$$

In this paper, we assume that σ^2 is known. By a straightforward calculation, it can be verified that

$$\begin{cases} \mathscr{C}_\alpha(K) = \mathscr{C}_{\text{sure}}(K) & \text{for } \alpha = 2; \\ \mathscr{C}_\alpha(K) \to \sigma^2\mathscr{C}_{\text{eb}}(K) - n\sigma^2\log\sigma^2 & \text{as } \alpha \to 1^+. \end{cases}$$

Therefore, the family of hyper-parameter estimators (11) unifies the EB and SURE estimators in a continuous way by index α and the EB and SURE methods correspond to the cases $\alpha = 1$ and $\alpha = 2$, respectively.

A. Finite Sample Properties

Along with the technique for deriving the properties of the SURE and EB estimators in [6], we investigate the properties of the estimation criterion $\mathscr{C}_\alpha(K)$ and its first-order optimality condition with respect to η, which helps us to understand the properties of the family. We first show that both of the two terms of the estimation criterion (11b) can be decomposed into two terms: one term dependent on K and the other one independent of K.

Proposition 1: Suppose that Assumption 1 holds. Thus, the two terms of the estimation criterion (11b) have the following decomposition:

$$\begin{aligned}\sigma^{2\alpha} Y^T Q^{-\alpha} Y = &\underbrace{\|Y - \Phi\widehat{\theta}^{\text{ls}}\|^2}_{O_p(n)} \\ &+ \underbrace{\sigma^{2\alpha} Y^T Q^{1-\alpha}\Phi(\Phi^T\Phi)^{-1}\Phi^T Q^{-1} Y}_{O_p(1/n^{\alpha-1})},\ \alpha \in [1, 2],\end{aligned} \tag{12a}$$

$$\begin{aligned}&\text{Tr}\big(I_n - (\sigma^2 Q^{-1})^{\alpha-1}\big) \\ &= \underbrace{p}_{O(1)} - \underbrace{\sigma^{2(\alpha-1)}\text{Tr}\big((\sigma^2 I_p + \Phi^T\Phi K)^{-(\alpha-1)}\big)}_{O(1/n^{\alpha-1})},\ \alpha \in (1, 2],\end{aligned} \tag{12b}$$

$$\begin{aligned}&\frac{\alpha}{\alpha - 1}\text{Tr}\Big(I_n - (\sigma^2 Q^{-1})^{\alpha-1}\Big) \\ &\qquad \to \log\det Q - n\log\sigma^2,\ \alpha \to 1^+.\end{aligned} \tag{12c}$$

Remark 1: The first term of (11b) consists of the prediction error of the LS estimate independent of K and another smaller term in scale but dependent of K, which can be understood as the prediction error of the RLS estimate $\widehat{\theta}^{\text{rls}}$ with the hyper-parameter $\widehat{\eta}_\alpha$ in some sense. In particular, if we take $K^{-1} = 0$ (no regularization corresponding to the LS estimate), then the smaller term is equal to zero. Therefore, the smaller term can be thought of as the price paid using regularization for fidelity to the data. The second term of (11) is decomposed into a constant term p (the dimension of parameters) and a smaller term for $\alpha \in (1, 2]$. In particular, the second term tends to the term $\log\det Q$ plus the term $-n\log\sigma^2$ independent of K as $\alpha \to 1^+$.

In the following, we study the first-order optimality condition of the family of hyper-parameters. In order to derive a clear and simple expression, however, we focus on the special case: $K = \eta K_1$ with $\eta > 0$ and a fixed positive definite matrix K_1.

Proposition 2: Suppose that $K = \eta K_1$ with $\eta > 0$ and a fixed positive definite matrix K_1. Then for $\alpha \in [1, 2]$, the first-order derivative of (11b) with respect to η is

$$\frac{\partial\mathscr{C}_\alpha(K(\eta))}{\partial\eta} = \alpha\sigma^{2\alpha}\text{Tr}\Big(Q^{-\alpha}\big(YY^T Q^{-1} - I_N\big)\Phi K_1\Phi^T\Big),$$

where $Q = \eta\Phi K_1\Phi^T + \sigma^2 I_N$. Thus, the family of hyper-parameters $\widehat{\eta}_\alpha$ are the roots of the system of equations

$$\text{Tr}\Big(Q^{-\alpha}\big(YY^T Q^{-1} - I_N\big)\Phi K_1\Phi^T\Big) = 0 \tag{13}$$

over η.

Proposition 2 tell us that index α influences the family of hyper-parameters through the matrix $(\sigma^2 Q^{-1})^\alpha$.

B. Asymptotic Properties: Convergence

In this subsection, we investigate the asymptotic properties of $\mathscr{C}_\alpha(K)$ and hyper-parameter estimators $\widehat{\eta}_\alpha$ for $\alpha \in [1,2]$. Before moving forward, we make assumptions on the Gram matrix $\Phi^T\Phi$ and kernel matrix $K(\eta)$.

Assumption 2: The Gram matrix $\Phi^T\Phi/n \to \Sigma$ as $n \to \infty$, where Σ is positive definite.

Assumption 3: The kernel parameterization $K(\eta)$ is positive definite for any $\eta \in \Omega$.

Then based on the decomposition derived in Proposition 1, we first derive the convergence result for (11b).

Proposition 3: Suppose that Assumptions 1 and 2 hold and the kernel matrix K is positive definite. Thus, an affine transform of $\mathscr{C}_\alpha(K)$ converges to a deterministic function almost surely, i.e., for $1 \le \alpha \le 2$,

$$\overline{\mathscr{C}}_\alpha(K) \to W_\alpha(K,\Sigma,\theta_0) \tag{14}$$

almost surely as $n \to \infty$, where

$$\overline{\mathscr{C}}_\alpha(K) = \begin{cases} n^{\alpha-1}\Big(\mathscr{C}_\alpha(K) - \|Y-\Phi\widehat{\theta}^{\rm ls}\|^2 - \dfrac{\alpha p\sigma^2}{\alpha-1}\Big) + \dfrac{\alpha p\sigma^{2\alpha}}{\alpha-1}, \alpha \in (1,2], \\ \mathscr{C}_\alpha(K) - \|Y-\Phi\widehat{\theta}^{\rm ls}\|^2 + \sigma^2 p\log\big(\sigma^2/n\big),\ \alpha = 1, \end{cases}$$

$$W_\alpha(K,\Sigma,\theta_0) = \sigma^{2\alpha}\Big(\theta_0^T K^{-1}(K\Sigma)^{1-\alpha}\theta_0 - \frac{\alpha}{\alpha-1}{\rm Tr}\big((K\Sigma)^{1-\alpha}\big)\Big) + \underbrace{\sigma^{2\alpha}\frac{\alpha}{\alpha-1}p}_{\text{independent of } K},\ \alpha \in [1,2].$$

Remark 2: Limiting function $W_\alpha(K,\Sigma,\theta_0)$ is a unified form for $\alpha \in [1,2]$ since its limit as $\alpha \to 1^+$ is

$$\sigma^2\big(\theta_0^T K^{-1}\theta_0 + \log\det(K)\big) + \underbrace{\sigma^2\log\det(\Sigma)}_{\text{independent of } K},$$

which is also the limit of $\overline{\mathscr{C}}_\alpha(K)$ for $\alpha = 1$. This unified expression discloses that index α enters the limiting function $W_\alpha(K,\Sigma,\theta_0)$ by the matrix $(K\Sigma)^{1-\alpha}$. In particular, as α goes from 1 to 2, the limiting function $W_\alpha(K,\Sigma,\theta_0)$ is influenced by the more and more ill-conditioned matrix $(K\Sigma)^{1-\alpha}$ if at least one of K and Σ is ill-conditioned.

We define the global minima of $W_\alpha(K,\Sigma,\theta_0)$ by

$$\eta_\alpha^* = \arg\min_{\eta\in\Omega} W_\alpha(K(\eta),\Sigma,\theta_0),\ \alpha \in [1,2]. \tag{16}$$

Then we show the closed-form expressions of η_α^* for two special kernel matrices.

Corollary 1: Suppose that Assumptions 1 and 2 hold.

1) If $K = \eta K_1$ with $\eta > 0$ and a fixed positive definite matrix K_1, then for $\alpha \in [1,2]$,

$$\eta_\alpha^* = \arg\min_{\eta\in\Omega}\Big(\eta^{-\alpha}\theta_0^T K_1^{-1}(K_1\Sigma)^{1-\alpha}\theta_0 - \frac{\alpha}{\alpha-1}\eta^{1-\alpha}{\rm Tr}((\Sigma K_1)^{1-\alpha})\Big) \tag{17}$$

$$= \frac{\theta_0^T K_1^{-1}(K_1\Sigma)^{1-\alpha}\theta_0}{{\rm Tr}((\Sigma K_1)^{1-\alpha})}. \tag{18}$$

2) If $K = {\rm diag}([\eta_1,\cdots,\eta_p])$ with $\eta_i \ge 0,\ 1 \le i \le p$ and $\Sigma = dI_p$ with $d > 0$, then for $\alpha \in [1,2]$,

$$\eta_\alpha^* = [(g_1^0)^2,\cdots,(g_p^0)^2]^T, \tag{19}$$

where g_i^0 is the ith element of θ_0.

Lastly, in order to derive the convergence of $\widehat{\eta}_\alpha$, we need the following assumption on η_α^*.

Assumption 4: For each $\alpha \in [1,2]$, the global minima η_α^* of $W_\alpha(K,\Sigma,\theta_0)$ exist and moreover, are isolated interior points of Ω.

Theorem 1: Suppose that Assumptions 1-4 hold. Thus, for each $\alpha \in [1,2]$, we have $\widehat{\eta}_\alpha \to \eta_\alpha^*$ almost surely as $n \to \infty$.

Remark 3: Theorem 1 embraces the convergence of the EB and SURE methods derived in [6] as the special cases for $\alpha = 1, 2$.

Clearly, it is interesting to investigate what value of $\alpha \in [1,\ 2]$ can yield the best hyper-parameter estimator in this family. In the next section, we use numerical simulation to show that there does exist an estimator in this family that is superior to the EB and SURE estimators. As for how to find out this estimator, i.e., the corresponding α, we will report the results in the journal version of this paper.

IV. Numerical Simulation

In this section, we test hyper-parameter estimators (11) with α taking values equidistantly within the interval $[1,2]$ at intervals of 0.1.

A. Test data-bank

We generate 1000 test system of order 30 using the method in [2] and [4]. To generate data, we first feed each test system with two different test inputs, which are the bandlimited white Gaussian noise with band [0,0.6] and [0,1], denoted by IT1 and IT2, respectively. Then we corrupt the noise-free output by an additive white Gaussian noise such that the signal-to-noise ratio (SNR), i.e., the ratio between the variance of the noise-free output and the noise, is uniformly distributed over [1, 10], and is unchanged for the two test inputs. We limit the sample sizes at 500 and 8000 to illustrate the finite sample and large sample performance of this family of hyper-parameter estimators, respectively.

B. Simulation set-up

The measure of fit [21] is used to evaluate the performance of RLS estimators (5b), which is defined as follows:

$${\rm Fit} = 100 \times \Big(1 - \frac{\|\hat{\theta}^{\rm rls} - \theta_0\|}{\|\theta_0 - \bar{\theta}_0\|}\Big),\quad \bar{\theta}_0 = \frac{1}{p}\sum_{i=1}^{p} g_i^0$$

where p is set to 200.

The TC kernel is employed and hyper-parameters η involved in kernel matrix are estimated by using this family of estimators defined in (11a) with $\alpha = \{1, 1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7,\ 1.8, 1.9, 2\}$, in which $\alpha = 1, 2$ correspond to EB (7) and SURE (9), respectively.

C. Simulation results

The average fits are given in Table I. The boxplots of 1000 fits are showed in Figs. 1-2 with the largest median being marked by a green circle.

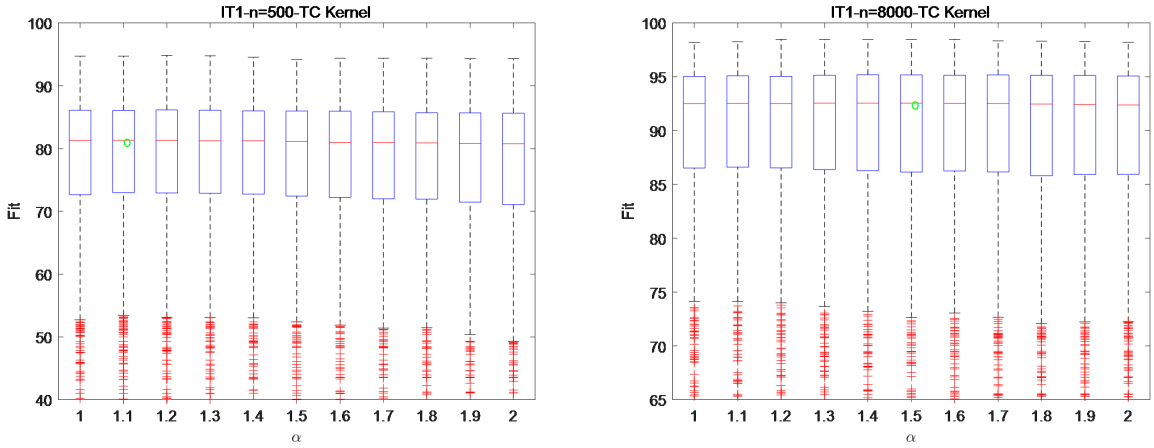

Fig. 1. Boxplots of the 1000 fits for the bandlimited white Gaussian noise with band [0,0.6]: $n = 500$ (left) and $n = 8000$ (right).

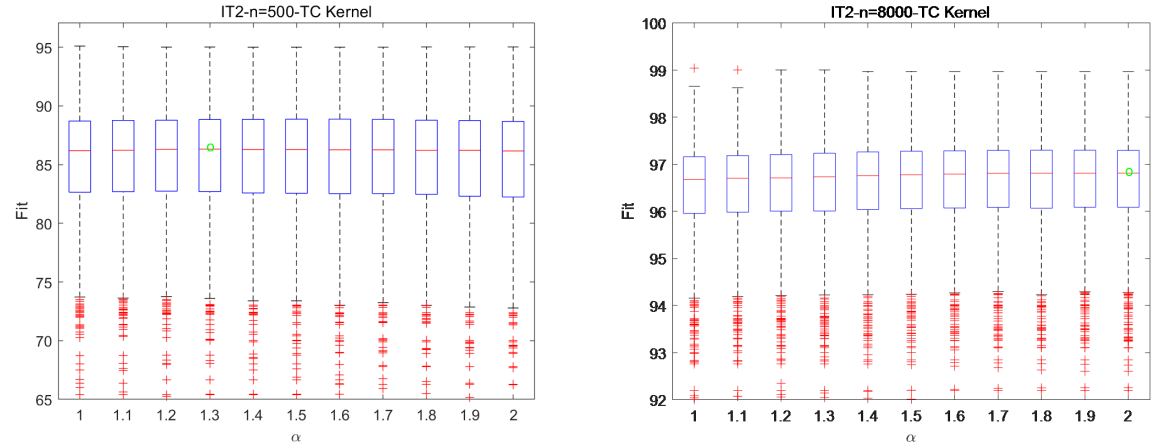

Fig. 2. Boxplots of the 1000 fits for the bandlimited white Gaussian noise with band [0,1]: $n = 500$ (left) and $n = 8000$ (right).

TABLE I

AVERAGE FITS FOR 1000 TEST SYSTEMS AND FOUR TEST INPUTS.

α	1	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2	optimal
IT1												
$n = 500$	77.1107	77.1640	77.1529	77.1257	77.0656	76.9476	76.7034	76.3690	76.2262	76.0577	75.1287	1.1
$n = 8000$	88.2759	88.2944	88.2774	88.2511	88.2216	87.8352	87.6543	87.5159	87.2918	87.1950	86.1877	1.1
IT2												
$n = 500$	85.1490	85.2093	85.2443	85.2625	85.2262	85.2220	85.1889	85.1699	85.1093	85.0503	84.9733	1.3
$n = 8000$	96.4006	96.4305	96.4549	96.4755	96.4963	96.5137	96.5272	96.5384	96.5439	96.5468	96.5443	1.9

D. Findings

It has been shown in [6] that input IT1 is bad, which make matrix $\Phi^T\Phi$ ill-conditioned, while input IT2 is the completely opposite and has a well-conditioned $\Phi^T\Phi$. Then we conclude the findings as follows.

Firstly, it can be observed in all the boxplots that the medians varies a little when α changes from 1 to 2 for the two inputs and sample sizes. However, it can be observed in Table I that the average fits differ slightly for input IT2, but considerably for input IT1. This indicates that this family of hyperparamter estimators possess different robustness, displayed by variance, to ill-conditioned inputs. Furthermore, the discrepancy of the average fits among this family is shortened when the sample size increases from 500 to 8000, which indicates that the variance will decrease as the number of data grows.

Secondly, for all the cases, it can be observed from Table I that the α values giving the maximum fits lie in $(1, 2)$. This confirms that there does exist a "middle" estimator in this family that is superior to the EB and SURE method. For input IT1, the optimal value α is around 1 though the sample size reaches 8000. In particular, when the sample size $n = 500$, there is a huge gap between the average fits of $\alpha = 1$ and 2, i.e., the EB and SURE method, and the average fit roughly decreases as α varying from 1 to 2.

V. Conclusion

The paper has proposed a family of hyper-parameter estimators linking the EB and SURE estimators in a unified way. The finite sample and asymptotic properties of this family have been established to further understand this family. And the Monte Carlo simulations suggest that there does exists a "middle" estimator in this family that is superior to the EB and SURE method no matter the input is bad or not. But how to choose α is still a problem to be solved and deserves to investigate in the future. Moreover, further theoretical grounds are necessary to support the simulation results.

References

[1] G. Pillonetto and G. De Nicolao, "A new kernel-based approach for linear system identification," *Automatica*, vol. 46, no. 1, pp. 81–93, 2010.

[2] T. Chen, H. Ohlsson, and L. Ljung, "On the estimation of transfer functions, regularizations and gaussian processes–revisited," *Automatica*, vol. 48, no. 8, pp. 1525–1535, 2012.

[3] G. Pillonetto, F. Dinuzzo, T. Chen, G. De Nicolao, and L. Ljung, "Kernel methods in system identification, machine learning and function estimation: A survey," *Automatica*, vol. 50, no. 3, pp. 657–682, 2014.

[4] G. Pillonetto and A. Chiuso, "Tuning complexity in regularized kernel-based regression and linear system identification: The robustness of the marginal likelihood estimator," *Automatica*, vol. 58, pp. 106–117, 2015.

[5] T. Chen, M. S. Andersen, L. Ljung, A. Chiuso, and G. Pillonetto, "System identification via sparse multiple kernel-based regularization using sequential convex optimization techniques," *IEEE Transactions on Automatic Control*, vol. 59, no. 11, pp. 2933–2945, 2014.

[6] B. Mu, T. Chen, and L. Ljung, "On asymptotic properties of hyper-parameter estimators for kernel-based regularization methods," *Automatica*, vol. 94, pp. 381–395, 2018.

[7] L. Ljung, *System Identification: Theory for the User*. Upper Saddle River, NJ: Prentice-Hall, 1999.

[8] L. Ljung, T. Chen, and B. Mu, "A shift in paradigm for system identification," *International Journal of Control*, vol. 93, no. 2, pp. 173–180, 2020.

[9] T. Chen, "On kernel design for regularized lti system identification," *Automatica*, vol. 90, pp. 109–122, 2018.

[10] A. Aravkin, J. V. Burke, A. Chiuso, and G. Pillonetto, "On the mse properties of empirical bayes methods for sparse estimation," in *Proceeding of the IFAC Symposium on System Identification*, Brussels, Belgium, 2012, pp. 965–970.

[11] B. Mu, T. Chen, and L. Ljung, "Tuning of hyperparameters for fir models–an asymptotic theory," in *Proceedings of the 20th IFAC World Congress*, Toulouse, France, 2017, pp. 2818–2823.

[12] ——, "Asymptotic properties of generalized cross validation estimators for regularized system identification," in *Proceedings of the IFAC Symposium on System Identification*, Stockholm, Sweden, 2018, pp. 203–205.

[13] ——, "Asymptotic properties of hyperparameter estimators by using cross-validations for regularized system identification," in *Proceedings of the 57th IEEE Conference on Decision and Control*, 2018, pp. 644–649.

[14] B. Mu and T. Chen, "On asymptotic optimality of cross-validation based hyper-parameter estimators for kernel-based regularized system identification," *IEEE Transactions on Automatic Control*, pp. 1–16, 2024.

[15] Y. Ju, B. Mu, L. Ljung, and T. Chen, "Asymptotic theory for regularized system identification Part I: Empirical Bayes hyper-parameter estimator," *IEEE Transactions on Automatic Control*, pp. 1–16, 2023.

[16] Y. Ju, T. Chen, B. Mu, and L. Ljung, "On the influence of ill-conditioned regression matrix on hyper-parameter estimators for kernel-based regularization methods," in *2020 59th IEEE Conference on Decision and Control (CDC)*, Dec 2020, pp. 300–305.

[17] A. Aravkin, J. V. Burke, A. Chiuso, and G. Pillonetto, "On the estimation of hyperparameters for empirical bayes estimators: Maximum marginal likelihood vs minimum mse," in *Proceeding of the IFAC Symposium on System Identification*, Brussels, Belgium, 2012, pp. 125–130.

[18] ——, "Convex vs non-convex estimators for regression and sparse estimation: the mean squared error properties of ard and glasso." *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 217–252, 2014.

[19] B. Efron, "Selection criteria for scatterplot smoothers," *The Annals of Statistics*, vol. 29, no. 2, pp. 470 – 505, 2001. [Online]. Available: https://doi.org/10.1214/aos/1009210549

[20] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. New York: Springer, 2009.

[21] L. Ljung, *System Identification Toolbox for Use with MATLAB*, 8th ed. Natick, MA: The MathWorks, Inc., 2012.