2024 IEEE 63rd Conference on Decision and Control (CDC)
December 16-19, 2024. MiCo, Milan, Italy

Asymptotic properties of generalized maximum likelihood hyper-parameter estimator for regularized system identification

Meng Zhang, Tianshi Chen, and Biqiang Mu

Abstract— **Regularized system identification is one of the major advances in the field of system identification in the last decade. One key issue is the hyper-parameter estimation, for which the generalized maximum likelihood (GML) estimator is a popular one closely related to the empirical Bayes (EB) method. Considering the rich theoretical results on the EB estimator, the asymptotic properties of the GML estimator has not been studied before and is critical for understanding its efficacy when the sample size is large. In this paper, we investigate the asymptotic properties of the GML estimator and show that the GML estimator is asymptotically equivalent to the EB estimator. Furthermore, Monte-Carlo simulations verify their asymptotic equivalence and also indicates the GML estimator outperform the EB estimator for small sample sizes.**

Index Terms— **Regularized system identification, empirical Bayes estimators, generalized maximum likelihood estimators, asymptotic properties.**

I. Introduction

Over the past decade, the kernel-based regularization method (KRM) has achieved great success in system identification and become an emerging new system identification paradigm [1]–[4]. The recent progress of KRM are on the kernel design and analysis, input design, efficient implementation, asymptotic theory, and etc.

Kernel design refers to how to parameterize the kernel by hyper-parameters based on various prior knowledge of the system to be identified. For example, two systematic kernel design methods from a system theoretic perspective and a machine learning perspective for causal stable linear time-invariant system (LTI) identification were proposed in [5]. The machine learning perspective was then adopted in [6] for the harmonic analysis of kernels, and the system theoretic perspective was further extended to non-causal stable LTI system identification in [7] and led to the non-causal SI kernel.

Efficient implementation refers to how to develop efficient implementation algorithms by exploring the structure of the key components involved in the computation of KRM, including the structure of the kernel, the input signal, and the output kernel. For example, it was shown in [8] that the AMLS and SI kernels can be semi-separable, and moreover, for many frequently used test input signals in automatic control, and by exploring the semiseparable structure of a kernel and the corresponding output kernel, the computational complexity of KRM, without any approximations, can be lowered to $O(N)$, where N is the sample size, by making use of the implementation in [9]. Following [8], some latest results on efficient implementation of KRM can be found in [10]–[12].

Input design aims to design input signals such that for a chosen model structure, a scalar measure of the covariance matrix of the estimator is minimized subjects to certain kinds of constrains on the inputs. This problem was first studied in [13], [14] by maximizing the mutual information between the output and the impulse response subject to energy-constraint on the input. Then in [15], a two-step procedure was introduced to avoid the non-convex problem encountered in [13]. Following [15], some latest results on input design of KRM can be found in [16], [17].

The asymptotic theory studies asymptotic properties of model estimators as the sample size goes to infinity, which has been widely used to evaluate the quality of an estimator [18]. There have been many results on the asymptotic properties of some commonly used hyper-parameter estimators, such as empirical Bayes (EB), Stein's unbiased risk estimator (SURE), cross-validation (CV) [19]–[21], and etc. For instance, the asymptotic properties including almost sure convergence, convergence in distribution and their connection of the EB and SURE methods have been extensively studied in [21]–[25], the asymptotic optimality of the CV methods has been studied in [26]–[28].

Besides these hyper-parameter estimators, the generalized maximum likelihood (GML) estimator is also popular [29], [8]. A significant advantage of the GML estimator over the EB/SURE methods is that it does not involve the unknown noise variance in the estimation criterion. In contrast, the EB/SURE methods often estimate the noise variance using the least squares (LS) estimator of an FIR model or an ARX model [4], [18], [30]. However, this approach may not perform satisfactorily, especially when the data is short and/or has low signal-to-noise ratio, potentially affecting the accuracy of the EB/SURE methods. However, the asymptotic properties of the GML estimator have not been studied before, despite being crucial for understanding its efficacy when the sample size is sufficiently large. In this paper, we address this gap by investigating the GML estimator. To this end, we first demonstrate that the GML estimator is essentially equivalent to the EB estimator, with a noise variance estimator that is linked to the chosen hyperparameters. We then analyze the asymptotic properties of the GML estimator, showing that it behaves identically to the EB estimator

Meng Zhang and Tianshi Chen are with the School of Data Science and Shenzhen Research Institute of Big Data, The Chinese University of Hong Kong, Shenzhen, 518172, China, {mengzhang1,tschen}@link.cuhk.edu.cn.

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

asymptotically. Finally, we use Monte Carlo simulations to validate our theoretical findings. Interestingly, the simulation results also indicate that the GML estimator can outperform the EB hyper-parameter estimator in scenarios with small sample sizes.

The remaining parts of the paper are organized as follows. Section 2 provides a brief overview of regularized least squares for impulse response identification, while Section 3 reviews the EB and GML estimators. In Section 4, we present the theoretical properties of the GML estimator, followed by Monte Carlo simulations in Section 5. Finally, a concluding remark in Section 6.

II. Regularized least squares estimators

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

$$y_k = G_0(q^{-1})u_k + v_k,\ k = 1, \dots, N, \tag{1}$$

where k is the discrete time index, N is the sample size, $G_0(q^{-1})$ is a rational transfer function of the linear time-invariant system with q^{-1} being the backward shift operator ($q^{-1}u_k = u_{k-1}$), u_k and y_k are the input and the output corrupted by the measurement noise v_k independent of the input u_k, respectively. The identification problem is to estimate the transfer function

$$G_0(q^{-1}) = \sum_{k=1}^{\infty} g_k^0 q^{-k} \tag{2}$$

determined by the impulse response coefficients $\{g_k^0, k = 1, \cdots, \infty\}$ as well as possible based on the the available data $\{u_k, y_k\}_{k=1}^N$.

The exponential stability of $G_0(q^{-1})$ implies that it is possible to truncate the infinite impulse response at a sufficiently high order and obtain a finite impulse response (FIR) model:

$$G(q^{-1}) = \sum_{k=1}^{n} g_k^0 q^{-k},\quad \theta_0 = [g_1^0, \cdots, g_n^0]^T \in \mathbb{R}^n. \tag{3}$$

Thus the estimation of the infinite impulse response (2) is reduced to the linear FIR model:

$$y_k = \phi_k^T \theta_0 + v_k,\ k = 1, \dots, N$$

where $\phi_k = [u_{k-1}, \cdots, u_{k-n}]^T \in \mathbb{R}^n$ are the regressors, and its matrix-vector form is

$$Y = \Phi\theta_0 + V, \text{ where} \tag{4a}$$
$$Y = [y_1\ y_2\ \cdots\ y_N]^T \tag{4b}$$
$$\Phi = [\phi_1\ \phi_2\ \cdots\ \phi_N]^T \tag{4c}$$
$$V = [v_1\ v_2\ \cdots\ v_N]^T. \tag{4d}$$

We make the following assumptions on the above linear model.

Assumption 1: (i) The dimension n of parameters θ_0 is fixed as $N \to \infty$;

(ii) The noise sequence $\{v_k\}$ is a sequence of independent and identically distributed random variables with zero mean and variance $\sigma^2 > 0$.

The unknown parameters θ_0 are typically estimated using the LS estimator

$$\widehat{\theta}^{\text{ls}} = \arg\min_{\theta\in\mathbb{R}^m} \|Y - \Phi\theta\|^2 = (\Phi^T\Phi)^{-1}\Phi Y. \tag{5}$$

When the input is ill-conditioned and/or the dimension n is large, the estimator (5) usually encounters the big variance problem and the RLS estimator defined by

$$\widehat{\theta} = \arg\min_{\theta\in\mathbb{R}^m} \|Y - \Phi\theta\|^2 + \sigma^2\theta^T P^{-1}\theta \tag{6a}$$
$$= P\Phi^T Q^{-1} Y \tag{6b}$$
$$Q = \Phi P \Phi^T + \sigma^2 I_N \tag{6c}$$

is a proper method to cure the big variance problem and achieves a good bias-variance trade-off [2], [4], where P is a positive semi-definite kernel matrix to be specified later.

For a given kernel matrix P, the mean squared error (MSE) criterion is a reference to evaluate the performance of the RLS estimator (6). Here we introduce the MSE reflecting its output prediction ability defined by [20]

$$\text{MSE}(P) = E\|\Phi(\widehat{\theta} - \theta_0)\|^2 \tag{7}$$

where $\|\cdot\|$ for a vector represents Euclidean norm and $E(\cdot)$ is the mathematical expectation with respect to the noise distribution.

The choice of matrix P significantly impacts the performance of the RLS estimator (6). However, directly optimizing P based on an optimization criterion is often impractical. This is because the number of elements in P typically far exceeds the size of the parameter vector. To select an appropriate P, kernel-based regularization methods, initially introduced in [2] and subsequently refined in [4], have devised a two-step approach to identify a suitable candidate for P using data. It involves two consecutive steps: kernel design and hyper-parameter estimation.

Kernel design is to parameterize P by a few number of parameters η, called hyper-parameters, namely,

$$P(\eta),\quad \eta \in \Omega \subset \mathbb{R}^p, \tag{8}$$

in which available prior knowledge of the system to be identified, e.g., exponential stability and smoothness, is encoded by the parameterization of P. Taking into account various forms of prior knowledge from different angles, several parameterization strategies for P have been constructed. These include the stable spline (SS) kernel [2], the diagonal correlated (DC) kernel, and the tuned-correlated (TC) kernel [4]:

$$\text{SS}: \quad P_{kj}(\eta) = c\Big(\frac{\alpha^{k+j+\max(k,j)}}{2} - \frac{\alpha^{3\max(k,j)}}{6}\Big)$$
$$\eta = [c, \alpha] \in \Omega = \{c \geq 0, 0 \leq \alpha \leq 1\}; \tag{9a}$$
$$\text{DC}: \quad P_{kj}(\eta) = c\alpha^{(k+j)/2}\rho^{|j-k|}$$
$$\eta = [c, \alpha, \rho] \in \Omega = \{c \geq 0, 0 \leq \alpha \leq 1, |\rho| \leq 1\}; \tag{9b}$$
$$\text{TC}: \quad P_{kj}(\eta) = c\alpha^{\max(k,j)}$$
$$\eta = [c, \alpha] \in \Omega = \{c \geq 0, 0 \leq \alpha \leq 1\}. \tag{9c}$$

For convenience in the following, we also denote

$$P(\eta) = cK(\lambda) \tag{10}$$

with $\eta = [c, \lambda]$, where c is the scale hyper-parameter.

Hyper-parameter estimation is to estimate the hyper-parameters by the data based on certain criteria. Several commonly used hyper-parameter estimation methods have been developed in the literature, e.g., empirical Bayes (EB) estimator, Stein's unbiased risk estimator (SURE), generalized maximum likelihood (GML) estimator, generalized cross-validation (GCV) estimator, and among others [8], [17], [19]–[21], [29].

III. Review of the EB and GML hyper-parameter estimators

In this section, we recap the EB estimator, along with a detailed procedure of how the GML estimator is deduced from the maximum likelihood principle.

The EB estimator is derived from the Bayesian perspective by assuming $\theta_0 \sim \mathcal{N}(0, P)$. Then, the output Y is Gaussian with zero mean and covariance Q defined in (6c). As a consequence, the EB method tunes the hyper-parameters η by maximizing the marginal likelihood of Y, also called the marginal likelihood estimator, which is equivalent to minimizing $\mathscr{C}_{\rm eb}(P(\eta))$:

$$\text{EB}: \widehat{\eta}_{\rm eb} = \arg\min_{\eta\in\Omega} \mathscr{C}_{\rm eb}(P(\eta)), \tag{11a}$$

$$\mathscr{C}_{\rm eb}(P) = Y^T Q^{-1} Y + \log\det(Q), \tag{11b}$$

where $\det(\cdot)$ denotes the determinant of a square matrix.

Note that the EB criterion depends on the unknown noise variance σ^2, which is involved in the matrix Q in the way $Q = \Phi P \Phi^T + \sigma^2 I_N$. Therefore, it needs to be estimated from the data in practice. Right now, the state-of-the-art method for estimating σ^2 follows from [18], [30], which first estimates an ARX model [2], [3] or an FIR model [4], [20], [21] with least squares and then chooses the sample variance as the estimate of the unknown noise variance. This method is also used in recent work for investigating the asymptotics of the EB estimator [24]. However, one main drawback of this method is that the accuracy of the estimated noise variance might not be reliable when the sample size is small and the input is ill-conditioned.

To circumvent the estimation of σ^2, another way is to treat σ^2 as an extra hyper-parameter and estimate it along with hyper-parameter η by maximizing the marginal likelihood, which is used in [29], [8]. This procedure follows the derivation of the generalized maximum likelihood (GML) method for choosing the shape parameter of kernel functions in the curve estimation problem [31], [32]. Specifically, instead of estimating σ^2 directly, it first exposes σ^2 in (11b) by redefining the kernel matrix as

$$\overline{P}(\overline{\eta}) = \overline{c} K(\lambda) \tag{12}$$

with $\overline{\eta} = [\overline{c}, \lambda],\ \overline{c} = c/\sigma^2$, leading to

$$Q = \sigma^2 \overline{Q},\ \overline{Q} = \Phi \overline{P} \Phi^T + I_N,$$

where $\overline{Q}$ is independent of σ^2. Accordingly, the EB estimator defined in (11) equivalently becomes

$$\widehat{\overline{\eta}}_{\rm eb} = \arg\min_{\overline{\eta}\in\Omega} \mathscr{C}_{\rm eb}(\overline{\eta}, \sigma^2), \tag{13a}$$

$$\mathscr{C}_{\rm eb}(\overline{\eta}, \sigma^2) = \frac{1}{\sigma^2} Y^T \overline{Q}^{-1} Y + N \log\sigma^2 + \log\det(\overline{Q}). \tag{13b}$$

Clearly, the domain of the new hyper-parameter $\overline{\eta}$ is still Ω. Then by differentiating $\mathscr{C}_{\rm eb}(\overline{\eta}, \sigma^2)$ with respect to σ^2 and setting it to be zero, i.e.,

$$\frac{\partial \mathscr{C}_{\rm eb}(\overline{\eta}, \sigma^2)}{\partial \sigma^2} = -\frac{1}{\sigma^4} Y^T \overline{Q}^{-1} Y + \frac{N}{\sigma^2} = 0,$$

we obtain the optimal value of σ^2, which is given by

$$\widehat{\sigma^2} = \frac{Y^T \overline{Q}^{-1} Y}{N} \tag{14}$$

for a given $\overline{\eta}$. Replacing σ^2 in (13b) by the optimal value (14) leads to the GML estimator, which tunes the hyper-parameters $\overline{\eta}$ by

$$\text{GML}: \widehat{\overline{\eta}}_{\rm gml} = \arg\min_{\overline{\eta}\in\Omega} \mathscr{C}_{\rm gml}(\overline{P}(\overline{\eta})), \tag{15a}$$

$$\mathscr{C}_{\rm gml}(\overline{P}) = N \log\left(\frac{Y^T \overline{Q}^{-1} Y}{N}\right) + \log\det(\overline{Q}), \tag{15b}$$

where the term irrespective of $\overline{\eta}$ is neglected.

Building upon the redefined kernel matrix $\overline{P}(\overline{\eta})$, the EB estimator with the noise variance replaced by the LS is specified as

$$\widehat{\overline{\eta}}_{\rm ebls} = \arg\min_{\overline{\eta}\in\Omega} \mathscr{C}_{\rm ebls}(\overline{\eta}, \widehat{\sigma^2_{\rm ls}}), \tag{16a}$$

$$\mathscr{C}_{\rm ebls}(\overline{\eta}, \widehat{\sigma^2_{\rm ls}}) = \frac{1}{\widehat{\sigma^2_{\rm ls}}} Y^T \overline{Q}^{-1} Y + N \log\widehat{\sigma^2_{\rm ls}} + \log\det(\overline{Q}), \tag{16b}$$

where

$$\widehat{\sigma^2_{\rm ls}} = \frac{\|Y - \Phi\widehat{\theta}^{\rm ls}\|^2}{N - n}. \tag{17}$$

It has been shown in [24] that $\widehat{\sigma^2_{\rm ls}} = \sigma^2 + O_p(1/\sqrt{N})$, which means that $\widehat{\sigma^2_{\rm ls}}$ is a consistent estimator for σ^2 under mild assumptions. In the sequel, we call the hyper-parameter estimator $\widehat{\overline{\eta}}_{\rm ebls}$ as the EBLS hyper-parameter estimator for convenience.

Remark 1: To enable a clearer comparison with the GML, the EBLS (16) is introduced. In fact, it is important to note that the scale parameter estimate $\widehat{\overline{c}}$ in (16) equals to the scale parameter estimate $\widehat{c}$ in (11) divided by $\widehat{\sigma^2_{\rm ls}}$.

IV. Main results

In this section, we will establish the asymptotic properties of the GML estimator.

Note that the asymptotics of the EBLS estimator (16) has been reported in [21], [24]. And both of the derivations of the GML (15) and EBLS (16) are based on the EB (13) and thus the analysis of the GML should follow the

line used in [21], [24]. But the proof is not straightforward since we cannot directly tell their similarities and differences from the estimation criteria. Therefore, we turn to calculate their first-order derivatives, which can reveal the terms that really rely on hyper-parameter $\overline{\eta}$ in the estimation criteria. Before proceeding, we first introduce an assumption on the parameterized kernel matrix $\overline{P}(\overline{\eta})$.

Assumption 2: The set Ω is connected, the parameterized kernel matrix $\overline{P}(\overline{\eta})$ is symmetric and positive definite, continuous and twice differentiable with respect to $\overline{\eta}$, and $\|\overline{P}(\overline{\eta})\| < \infty$ for any interior point $\overline{\eta} \in \Omega$, where $\|\cdot\|$ for a square matrix denotes spectral norm.

Proposition 1: Suppose that Assumption 2 holds. Then, for $i = 1, \cdots, p$, the first-order derivatives of (15b) and (16b) with respect to $\overline{\eta}_i$ are, respectively,

$$\frac{\partial \mathscr{C}_{\rm ebls}(\overline{\eta}, \widehat{\sigma^2_{\rm ls}})}{\partial \overline{\eta}_i} = \frac{1}{\widehat{\sigma^2_{\rm ls}}} (\widehat{\theta}^{\rm ls})^T \frac{\partial \overline{S}(\overline{\eta})^{-1}}{\partial \overline{\eta}_i} \widehat{\theta}^{\rm ls} + {\rm Tr}\Big(\overline{S}(\overline{\eta})^{-1} \frac{\partial \overline{P}(\overline{\eta})}{\partial \overline{\eta}_i}\Big), \tag{18a}$$

$$\frac{\partial \mathscr{C}_{\rm gml}(\overline{P}(\overline{\eta}))}{\partial \overline{\eta}_i} = \frac{1}{\widehat{\sigma^2}(\overline{\eta})} (\widehat{\theta}^{\rm ls})^T \frac{\partial \overline{S}(\overline{\eta})^{-1}}{\partial \overline{\eta}_i} \widehat{\theta}^{\rm ls} + {\rm Tr}\Big(\overline{S}(\overline{\eta})^{-1} \frac{\partial \overline{P}(\overline{\eta})}{\partial \overline{\eta}_i}\Big), \tag{18b}$$

where $\overline{S} = \overline{P} + (\Phi^T\Phi)^{-1}$, and $\overline{\eta}_i$ is the ith element of $\overline{\eta}$.

Proposition 1 provides a clear overview on the same parts and the different parts of the GML and EBLS estimation criteria and shows that the only difference between the GML and EBLS lies in the different estimates for the noise variance. Provided that $\widehat{\sigma^2}(\overline{\eta})$ shares the same limit with $\widehat{\sigma^2_{\rm ls}}$, these observations suggest that $\frac{\partial \mathscr{C}_{\rm ebls}(\overline{\eta}, \widehat{\sigma^2_{\rm ls}})}{\partial \overline{\eta}}$ and $\frac{\partial \mathscr{C}_{\rm gml}(\overline{P}(\overline{\eta}))}{\partial \overline{\eta}}$ may have the same limit.

In the sequel, based on the insights from Proposition 1, we first investigate the convergence of the estimation criterion of the GML method and then that of the GML estimator. Before doing it, we need to give an assumption on the Gram matrix $\Phi^T\Phi$.

Assumption 3: $\Phi^T\Phi/N \to \Sigma$ almost surely as $N \to \infty$, where Σ is positive definite.

For convenience of notation, we introduce the averaged squared residuals by the LS estimator:

$$\chi_N \triangleq \frac{1}{N}\|Y - \Phi\widehat{\theta}^{\rm ls}\|^2 = \frac{1}{N} Y^T (I_N - \Phi(\Phi^T\Phi)^{-1}\Phi^T) Y. \tag{19}$$

Under Assumptions 1 and 3, there holds that

$$\chi_N = \sigma^2 + O_p\big(1/\sqrt{N}\big). \tag{20}$$

The following Proposition shows that each term within the estimation criterion of the GML method can be decomposed into one term relying on $\overline{\eta}$ and another term independent of $\overline{\eta}$.

Proposition 2: Suppose that Assumptions 1, 2 and 3 hold. Thus, we have the following decompositions for the estimation criterion of the GML:

$$\frac{Y^T\overline{Q}^{-1}Y}{N} = \underbrace{\chi_N}_{O_p(1)} + \underbrace{\frac{1}{N}(\widehat{\theta}^{\rm ls})^T\overline{S}^{-1}\widehat{\theta}^{\rm ls}}_{\gamma_N = O_p(1/N)}, \tag{21a}$$

$$N\log\Big(\frac{Y^T\overline{Q}^{-1}Y}{N}\Big) = \underbrace{N\log\Big(1 + \frac{\gamma_N}{\chi_N}\Big)}_{O_p(1)} + N\log(\chi_N), \tag{21b}$$

$$\log\det(\overline{Q}) = \underbrace{\log\det(\overline{S})}_{O_p(1)} + \log\det(\Phi^T\Phi); \tag{21c}$$

In light of the consistency (20), the equation (21a) in Proposition 2 shows that the estimator for the noise variance used in the GML estimator is consistent for σ^2. Together with Proposition 1, we speculate that the GML method shares the same asymptotic properties with the EBLS method. The following results confirm the conjecture.

Based on Proposition 2, the following proposition demonstrates that the affine transformations of the GML and EBLS estimation criteria, achieved by subtracting terms independent of the kernel matrix $\overline{P}$, converge to the same deterministic function almost surely as the sample size approaches infinity.

Proposition 3: Suppose that Assumptions 1 and 2 hold. Then, it entails that

$$\mathscr{C}_{\rm ebls}(\overline{\eta}, \widehat{\sigma^2_{\rm ls}}) - N\chi_N - \log\det(\Phi^T\Phi) \to \overline{W}(\overline{P}, \theta_0), \tag{22a}$$

$$\mathscr{C}_{\rm gml}(\overline{P}) - N\log(\chi_N) - \log\det(\Phi^T\Phi) \to \overline{W}(\overline{P}, \theta_0) \tag{22b}$$

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

$$\overline{W}(\overline{P}, \theta_0) = \frac{1}{\sigma^2}\theta_0^T\overline{P}^{-1}\theta_0 + \log\det(\overline{P}).$$

Based on the limiting loss function given in Proposition 3, we define the hyper-parameters that minimize $\overline{W}(\overline{P}, \theta_0)$ by

$$\overline{\eta}^* \triangleq \arg\min_{\overline{\eta}\in\Omega} \overline{W}(\overline{P}(\overline{\eta}), \theta_0). \tag{23}$$

For certain specific kernel matrices, the limiting hyper-parameters $\overline{\eta}^*$ have explicit expressions.

Corollary 1: Suppose that Assumptions 1 and 3 hold. Then,

(i) when the kernel matrix: $\overline{P} = \overline{\eta}K$, where $\overline{\eta} > 0$ and K is fixed and positive definite,

$$\overline{\eta}^* = \arg\min_{\overline{\eta}\geq 0} \frac{1}{\overline{\eta}\sigma^2}\theta_0^T K^{-1}\theta_0 + n\log\overline{\eta} + \log\det(K) = \frac{\theta_0^T K^{-1}\theta_0}{n\sigma^2}.$$

(ii) when the kernel matrix: $\overline{P} = {\rm diag}([\overline{\eta}_1, \cdots, \overline{\eta}_n])$ where $\overline{\eta}_i > 0,\ i = 1, \cdots, n$, and $\Sigma = cI_n$ with $c > 0$,

$$\overline{\eta}^* = \frac{1}{\sigma^2}[(g_1^0)^2, \cdots, (g_n^0)^2]^T.$$

However, regarding general cases, we require the following assumption on the location of $\overline{\eta}^*$ for further investigating its asymptotic.

Assumption 4: The set $\overline{\eta}^*$ consists of isolated interior points of Ω.

Then we have the following result according to Proposition 3.

Theorem 1: Suppose that Assumptions 1-4 hold. Then, the following limits hold

$$\widehat{\overline{\eta}}_{\text{ebls}} \to \overline{\eta}^*, \ \widehat{\overline{\eta}}_{\text{gml}} \to \overline{\eta}^* \tag{24a}$$

almost surely as $N \to \infty$.

Theorem 1 says that the hyper-parameters tuned by the GML and EBLS methods converge to the same limit $\overline{\eta}^*$ as $N \to \infty$, which is the minimizer of the limiting loss function $\overline{W}(\overline{\eta}, \theta_0)$. This means that the GML and EBLS estimators are asymptotically equivalent.

From the perspective of asymptotic properties, therefore, we do not need to distinguish the GML and EBLS estimators.

V. Numerical Illustration

In this section, we run Monte-Carlo simulations to illustrate the numerical performance of the GML estimator (15), the EBLS estimator (16), and the EB estimator (13).

A. Test data-bank

By using the method in [4], 1000 30th order test systems are generated randomly. And then each system is truncated to FIR system with order $n = 200$.

For each FIR system, we consider two different test input signals denoted by IT1 and IT2, which are white Gaussian noise of unit variance filtered by a second order transfer function $1/(1 - aq^{-1})^2$ with $a = 0.1, 0.9$, respectively. To generate data, we first obtain the noise-free output by simulating each FIR system with the test input signal and then corrupt the noise-free output by an additive white Gaussian noise with variance σ^2 such that the signal-to-noise ratio (SNR) is uniformly distributed in $[1, 10]$, which remains the same for two inputs. For each input signal, we use the sample sizes $N = 400$ and $N = 8000$ to show the performance of these estimators under the small and large sample sizes.

B. Simulation setup

Here, the unknown input is not used. Thus, the length of data used in each experiment is $N - n$. And the TC kernel is applied to parameterize the kernel matrix and the involved hyper-parameter becomes $\overline{\eta} = [\overline{c}, \lambda], \overline{c} = c/\sigma^2$ by adopting the noise variance σ^2 into the scale parameter c. Furthermore, the hyper-parameter $\overline{\eta}$ is estimated by the GML method (15), the EBLS method (16), and the EB method (13), where the true value of σ^2 is used for (13).

To evaluate the performance of the RLS estimates with these hyper-parameters, we use the fit defined in [33],

$$\text{Fit} = 100 \times \left(1 - \frac{\|\widehat{\theta} - \theta_0\|}{\|\theta_0 - \bar{\theta}_0\|}\right), \text{where } \bar{\theta}_0 = \frac{1}{n}\sum_{i=1}^{n} g_i^0.$$

C. Simulation results

The average fits of the RLS estimates over 1000 test systems are given in Table I. The boxplots of fits for IT1 and IT2 under the sample sizes $N = 400$ and $N = 8000$ are presented in Figs. 1 and 2.

TABLE I

Average fits for the RLS estimates over 1000 test systems and datasets for all the cases.

Inputs	Sample sizes	GML	EBLS	EB
IT1	$N = 400$	83.03	35.76	82.99
	$N = 8000$	96.31	96.31	96.31
IT2	$N = 400$	36.53	-5.59	35.67
	$N = 8000$	49.92	49.92	49.92

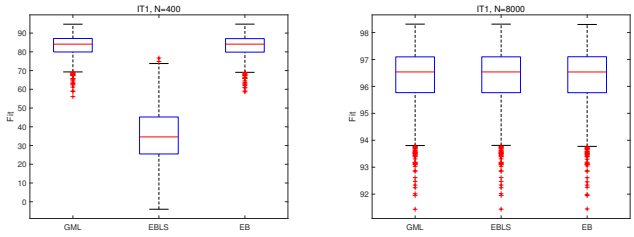

Fig. 1. Boxplots of fits over 1000 test systems for IT1 with sample size $N = 400, 8000$.

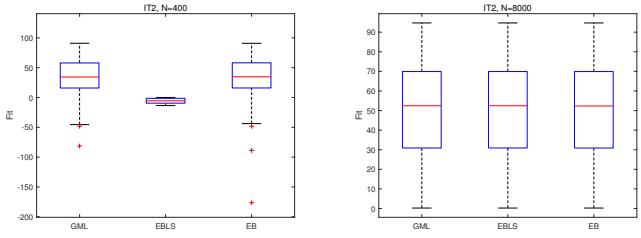

Fig. 2. Boxplots of fits over 1000 test systems for IT2 with sample size $N = 400, 8000$.

D. Findings

We obtained the following observations from the simulation results.

Firstly, for all sample sizes, the average fits of the three methods for IT1 is much larger compared to IT2. This is because $\Phi^T\Phi$ of IT2 is much more ill-conditioned than that of IT1, which can be deduced based on Lemma 2 in [24]. And for each input, the average fits get higher as the sample size N increases.

Secondly, for both IT1 and IT2 with $N = 400$, their average fits of the GML and EB hyper-parameter estimators are quite higher than that of the EBLS hyper-parameter estimator, and the average fit of the GML hyper-parameter estimator is slightly better than that of the EB. Actually, according to the simulation setup, the case $N = 400$ corresponds to the scenario where the sample size equals the dimension of the parameters. In this case, the noise variance

estimate given by the LS estimate is infinity, which makes the first term in the EB estimation criterion (16) tends to zero, leading to the ineffectiveness of the EBLS estimator.

Lastly, for both IT1 and IT2 with $N = 8000$, the average fits of the GML, EBLS and EB estimators are equal and their distributions of 1000 fits are almost the same, which verifies their asymptotic equivalence shown in Theorem 1.

VI. Concluding remark

In this paper, on the one hand, we studied the asymptotic properties of the GML estimator and demonstrated that it is asymptotically equivalent to the EBLS hyper-parameter estimator. On the other hand, Monte-Carlo simulations validate our theoretical findings that the GML estimator performs similarly with the EBLS when the sample size is sufficiently large and also indicate that the GML estimator behaves better than EBLS estimator for the scenario when the sample size is so small such that the noise variance estimator $\widehat{\sigma^2_{\text{ls}}}$ diverges. With this inspiration, we will continue our work to explore the distinction between the GML and EBLS estimators that adopt different noise variance estimators.

References

[1] 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.

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

[3] G. Pillonetto, A. Chiuso, and G. De Nicolao, "Prediction error identification of linear systems: A nonparametric Gaussian regression approach," *Automatica*, vol. 47, no. 2, pp. 291–305, 2011.

[4] 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.

[5] T. Chen, M. S. Andersen, B. Mu, L. Ljung, and S. J. Qin, "Regularized LTI system identification with multiple regularization matrix," in *Proceedings of the 18th IFAC Symposium on System Identification*, 2018, pp. 180–185.

[6] M. Zorzi and A. Chiuso, "The harmonic analysis of kernel functions," *Automatica*, vol. 94, pp. 125–137, 2018.

[7] X. Fang and T. Chen, "On kernel design for regularized non-causal system identification," *Automatica*, vol. 159, p. 111335, 2024.

[8] T. Chen and M. S. Andersen, "On semiseparable kernels and efficient implementation for regularized system identification and function estimation," *Automatica*, vol. 132, p. 109682, 2021.

[9] M. S. Andersen and T. Chen, "Smoothing splines and rank structured matrices: Revisiting the spline kernel," *SIAM Journal on Matrix Analysis and Applications*, vol. 41, no. 2, pp. 389–412, 2020.

[10] L. Chen, T. Chen, U. Detha, and M. S. Andersen, "Towards scalable kernel-based regularized system identification," in *2023 62nd IEEE Conference on Decision and Control (CDC)*. IEEE, 2023, pp. 1498–1504.

[11] Z. Shen, Y. Xu, M. S. Andersen, and T. Chen, "An efficient implementation for kernel-based regularized system identification with periodic input signals," in *2023 62nd IEEE Conference on Decision and Control (CDC)*. IEEE, 2023, pp. 1480–1485.

[12] J. Zhang, Y. Ju, B. Mu, R. Zhong, and T. Chen, "An efficient implementation for spatial–temporal gaussian process regression and its applications," *Automatica*, vol. 147, p. 110679, 2023.

[13] Y. Fujimoto and T. Sugie, "Informative input design for kernel-based system identification," in *Proceedings of IEEE Conference on Decision and Control*. IEEE, 2016, pp. 4636–4639.

[14] ——, "Informative input design for kernel-based system identification," *Automatica*, vol. 89, pp. 37–43, 2018.

[15] B. Mu and T. Chen, "On input design for regularized LTI system identification: Power-constrained input," *Automatica*, vol. 97, pp. 327–338, 2018.

[16] B. Mu, H. Kong, T. Chen, B. Jiang, L. Wang, and J. Wu, "Input design for regularized system identification: Stationary conditions and sphere preserving algorithm," *IEEE Transactions on Automatic Control*, vol. 68, no. 9, pp. 5714–5720, 2023.

[17] B. Mu, T. Chen, H. Kong, B. Jiang, L. Wang, and J. Wu, "On embeddings and inverse embeddings of input design for regularized system identification," *Automatica*, vol. 147, p. 110673, 2023.

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

[19] 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, March 2014.

[20] 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.

[21] 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.

[22] Y. Ju, T. Chen, B. Mu, and L. Ljung, "On convergence in distribution of Stein's unbiased risk hyper-parameter estimator for regularized system identification," in *Proceedings of the 41st Chinese Control Conference*, 2022, pp. 1491–1496.

[23] ——, "On the influence of ill-conditioned regression matrix on hyper-parameter estimators for kernel-based regularization methods," in *Proceedings of the 59th IEEE Conference on Decision and Control*, 2020, pp. 300–305.

[24] 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*, vol. 68, no. 12, pp. 7224–7239, 2023.

[25] M. Zhang, T. Chen, and B. Mu, "A family of hyper-parameter estimators linking eb and sure for kernel-based regularization methods," *IEEE Transactions on Automatic Control*, vol. 69, no. 12, p. 10.1109/TAC.2024.3416162, 2024.

[26] B. Mu, T. Chen, and L. Ljung, "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.

[27] ——, "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.

[28] 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*, vol. 69, no. 7, pp. 4352–4367, 2024.

[29] T. Chen and L. Ljung, "Constructive state space model induced kernels for regularized system identification," in *Proceedings of the 19th World Congress*, Cape Town, South Africa, 2014, pp. 1047–1052.

[30] G. C. Goodwin, M. Gevers, and B. Ninness, "Quantifying the error in estimated transfer functions with application to model order selection," *IEEE Transactions on Automatic Control*, vol. 37, no. 7, pp. 913–928, 1992.

[31] M. L. Stein, *Interpolation of spatial data: some theory for kriging*. Springer, 1999.

[32] G. Wahba, *Spline Models for Observational Data*. New York: SIAM, 1990.

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