Bayes and Biased Estimators Without Hyper-parameter Estimation: Comparable Performance to the Empirical-Bayes-Based Regularized Estimator

Yue Ju, Bo Wahlberg, Håkan Hjalmarsson

***Abstract*— Regularized system identification has become a significant complement to more classical system identification. Kernel-based regularized estimators often perform better than the maximum likelihood estimator in terms of minimizing mean squared error (MSE), but often require hyper-parameter estimation. This paper focuses on ridge regression and the regularized estimator by employing the empirical Bayes hyper-parameter estimator. We utilize the excess MSE expressions to develop a family of generalized Bayes estimators and a family of closed-form biased estimators. They have the same excess MSE as the empirical-Bayes-based regularized estimator, but eliminate the need for hyper-parameter estimation.**

I. Preliminaries and Problem Statement

We consider the following linear regression model,

$$\boldsymbol{Y} = \boldsymbol{\Phi}\boldsymbol{\theta} + \boldsymbol{E},$$

where $\boldsymbol{Y} \in \mathbb{R}^N$ is the measurement output vector with N being the sample size, $\boldsymbol{\Phi} \in \mathbb{R}^{N\times n}$ is a lower triangular matrix consisting of inputs, and $\boldsymbol{\theta} \in \mathbb{R}^n$ is the model parameter vector to be estimated. The measurement noise vector $\boldsymbol{E} \in \mathbb{R}^N$ is assumed to follow $\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_N)$. Given an estimator $\hat{\boldsymbol{\theta}} \in \mathbb{R}^n$ of the unknown parameter $\boldsymbol{\theta}$, we evaluate its average performance by its mean squared error (MSE): $\mathrm{MSE}(\hat{\boldsymbol{\theta}}) = \mathbb{E}(\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\|_2^2)$, where $\boldsymbol{\theta}_0 \in \mathbb{R}^n$ is the "true" value of $\boldsymbol{\theta}$ and the expectation $\mathbb{E}$ is with respect to the measurement noise $\boldsymbol{E}$. The smaller its MSE, the better its performance.

A. Maximum likelihood and regularized estimators

One classical estimator of $\boldsymbol{\theta}$ is the ML estimator given by $\hat{\boldsymbol{\theta}}^{\mathrm{ML}} = (\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^\top\boldsymbol{Y}$. It is well-known that $\hat{\boldsymbol{\theta}}^{\mathrm{ML}}$ is unbiased but may have a large variance, which will result in large $\mathrm{MSE}(\hat{\boldsymbol{\theta}}^{\mathrm{ML}}) = \sigma^2\,\mathrm{Tr}[(\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}]$. To achieve a better bias-variance trade-off, we consider the following regularized ridge regression estimator,

$$\hat{\boldsymbol{\theta}}^{\mathrm{R}}(\hat{\eta}) = [\boldsymbol{\Phi}^\top\boldsymbol{\Phi} + (\sigma^2/\hat{\eta})\mathbf{I}_n]^{-1}\boldsymbol{\Phi}^\top\boldsymbol{Y}, \tag{1}$$

where $\hat{\eta}$ is a hyper-parameter estimator. A commonly used hyper-parameter estimator is the EB one [4], given by

$$\begin{aligned}\hat{\eta}_{\mathrm{EB}} &= \arg\min_{\eta>0} \mathscr{F}_{\mathrm{EB}}(\eta),\\ \mathscr{F}_{\mathrm{EB}}(\eta) &= \boldsymbol{Y}^\top\mathbf{Q}(\eta)^{-1}\boldsymbol{Y} + \log\det(\mathbf{Q}(\eta))\end{aligned}$$

with $\mathbf{Q}(\eta) = \eta\boldsymbol{\Phi}\boldsymbol{\Phi}^\top + \sigma^2\mathbf{I}_N$. Correspondingly, $\hat{\boldsymbol{\theta}}^{\mathrm{R}}(\hat{\eta}_{\mathrm{EB}})$ in the form of (1) will be referred to as the EB-based regularized estimator in this paper.

B. Problem statement

Since $\mathrm{MSE}(\hat{\boldsymbol{\theta}}^{\mathrm{R}}(\hat{\eta}_{\mathrm{EB}}))$ is analytically intractable in finite-sample scenarios, we apply a high-order asymptotic quantity: the excess MSE (XMSE) [2]. It can be used to quantify the difference between $\mathrm{MSE}(\hat{\boldsymbol{\theta}}^{\mathrm{R}}(\hat{\eta}_{\mathrm{EB}}))$ and $\mathrm{MSE}(\hat{\boldsymbol{\theta}}^{\mathrm{ML}})$ for large sample sizes. Based on [2, Theorem 2], we can derive the XMSE of $\hat{\boldsymbol{\theta}}^{\mathrm{R}}(\hat{\eta}_{\mathrm{EB}})$.

Lemma 1: Assume that $\lim_{N\to\infty}\boldsymbol{\Phi}^\top\boldsymbol{\Phi}/N = \mathbf{I}_n$. We then have $\mathrm{XMSE}(\hat{\boldsymbol{\theta}}^{\mathrm{R}}(\hat{\eta}_{\mathrm{EB}})) = (-n^2+4n)(\sigma^2)^2/\|\boldsymbol{\theta}_0\|_2^2$.

In this work, we consider the following two problems.

1) Is it possible to design a generalized Bayes estimator[1] that has the same XMSE as $\hat{\boldsymbol{\theta}}^{\mathrm{R}}(\hat{\eta}_{\mathrm{EB}})$? It is free of hyper-parameters, thereby eliminating the computational cost associated with estimating such parameters.
2) Although a generalized Bayes estimator does not need any hyper-parameter, it often needs to be computed using sampling methods. The question thus arises whether it is possible to design a biased estimator in closed form that has the same XMSE as $\hat{\boldsymbol{\theta}}^{\mathrm{R}}(\hat{\eta}_{\mathrm{EB}})$.

II. Bayes and Biased Estimators

We first design generalized Bayes estimators that have the same XMSE as $\hat{\boldsymbol{\theta}}^{\mathrm{R}}(\hat{\eta}_{\mathrm{EB}})$.

Theorem 1: If the weighting function of $\hat{\boldsymbol{\theta}}^{\mathrm{Bayes,EB}}$ is

$$\pi(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_2^{2-n}(C_1\|\boldsymbol{\theta}\|_2 + C_2\|\boldsymbol{\theta}\|_2^{-1})^2, \tag{3}$$

where $C_1, C_2 \in \mathbb{R}$ are arbitrary constants, we then have $\mathrm{XMSE}(\hat{\boldsymbol{\theta}}^{\mathrm{Bayes,EB}}) = \mathrm{XMSE}(\hat{\boldsymbol{\theta}}^{\mathrm{R}}(\hat{\eta}_{\mathrm{EB}}))$.

We then design the following biased estimators that have the same XMSE as $\hat{\boldsymbol{\theta}}^{\mathrm{R}}(\hat{\eta}_{\mathrm{EB}})$ and $\hat{\boldsymbol{\theta}}^{\mathrm{Bayes,EB}}$.

Theorem 2: If $\hat{\boldsymbol{\theta}}^{\mathrm{Biased,EB}} = \hat{\boldsymbol{\theta}}^{\mathrm{ML}} + (1/N)\boldsymbol{b}_N(\hat{\boldsymbol{\theta}}^{\mathrm{ML}})$ and

$$\begin{aligned}\boldsymbol{b}_N(\hat{\boldsymbol{\theta}}^{\mathrm{ML}}) &= \sigma^2 N(\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}\\ &\times\left[2-n+\frac{2(C_1\|\hat{\boldsymbol{\theta}}^{\mathrm{ML}}\|_2 - C_2\|\hat{\boldsymbol{\theta}}^{\mathrm{ML}}\|_2^{-1})}{C_1\|\hat{\boldsymbol{\theta}}^{\mathrm{ML}}\|_2 + C_2\|\hat{\boldsymbol{\theta}}^{\mathrm{ML}}\|_2^{-1}}\right]\frac{\hat{\boldsymbol{\theta}}^{\mathrm{ML}}}{\|\hat{\boldsymbol{\theta}}^{\mathrm{ML}}\|_2^2},\end{aligned} \tag{4}$$

where $C_1, C_2 \in \mathbb{R}$ are arbitrary constants, then we have $\mathrm{XMSE}(\hat{\boldsymbol{\theta}}^{\mathrm{Biased,EB}}) = \mathrm{XMSE}(\hat{\boldsymbol{\theta}}^{\mathrm{Bayes,EB}}) = \mathrm{XMSE}(\hat{\boldsymbol{\theta}}^{\mathrm{R}}(\hat{\eta}_{\mathrm{EB}}))$.

III. Numerical Simulation

We generate 100 collections of test systems and input-output data. For each collection, 1) we generate $\tilde{\boldsymbol{\theta}}_0$ as a realization of $\mathcal{N}(\mathbf{0}, \mathbf{I}_n)$ and scale $\boldsymbol{\theta}_0 = m_\theta\tilde{\boldsymbol{\theta}}_0$ such that

[1]For a generalized Bayes estimator, its nonnegative weighting function $\pi(\boldsymbol{\theta})$ can be improper, i.e., $\int\pi(\boldsymbol{\theta})d\boldsymbol{\theta} = +\infty$.

$\|\boldsymbol{\theta}_0\|_2 = 1$; 2) generate $\{\tilde{u}(t)\}_{t=1}^N$ as independent realizations of $\mathcal{N}(0,1)$ and set $\sigma^2 = 1$; 3) scale $u(t) = m_u\tilde{u}(t)$ such that the sample SNR, which is the ratio between the sample variance of $\boldsymbol{\Phi}\boldsymbol{\theta}_0$ and the measurement noise variance σ^2, is 5; 4) corrupt the noise-free output $\boldsymbol{\Phi}\boldsymbol{\theta}_0$ with $N_{\rm MC} = 200$ additive independent noise realizations of $\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_n)$, to obtain $N_{\rm MC}$ measurement output sequences $\{y(t)\}_{t=1}^N$.

For each collection of test system and input-output data, we will perform $N_{\rm MC} = 200$ Monte Carlo (MC) simulations. The regularized estimator $\hat{\boldsymbol{\theta}}^{\rm R}(\hat{\eta}_{\rm EB})$ will be implemented using [1, Algorithm 2]. The Bayes estimator $\hat{\boldsymbol{\theta}}^{\rm Bayes,EB}$ will be approximated using the sampling method. The average performance of an estimator $\hat{\boldsymbol{\theta}}$ will be measured by the sample mean of $\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\|_2$ over 200 MC simulations, referred to as the sample MSE($\hat{\boldsymbol{\theta}}$). As its relative version, the average FIT($\hat{\boldsymbol{\theta}}$) [3] is given by the sample mean of FIT($\hat{\boldsymbol{\theta}}$) $= 100 \times (1 - \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_2/\|\boldsymbol{\theta}_0 - \bar{\theta}_0\|_2)$ with $\bar{\theta}_0 = \frac{1}{n}\sum_{k=1}^n [\boldsymbol{\theta}_0]_k$ over 200 MC simulations. The better $\hat{\boldsymbol{\theta}}$ performs, the smaller its sample MSE, while the larger its average FIT.

We consider the following two settings: 1) $n = 1$ $N = 5$, and $M_s = 200$, 2) $n = 5$, $N = 15$, and $M_s = 500$. From Fig. 1-2, we can observe that for $n = 1, N = 5$, $\hat{\boldsymbol{\theta}}^{\rm ML}$ outperforms $\hat{\boldsymbol{\theta}}^{\rm R}(\hat{\eta}_{\rm EB})$, while for $n = 5, N = 15$, the performance of $\hat{\boldsymbol{\theta}}^{\rm ML}$ is worse, which confirms the discussions after Lemma 1; for at least one combination of C_1 and C_2, $\hat{\boldsymbol{\theta}}^{\rm Bayes,EB}$ with (3) and $\hat{\boldsymbol{\theta}}^{\rm Biased,EB}$ with (4) perform similarly to $\hat{\boldsymbol{\theta}}^{\rm R}(\hat{\eta}_{\rm EB})$; among different combinations of C_1 and C_2, for $n = 1$, $\hat{\boldsymbol{\theta}}^{\rm Bayes,EB}$ and $\hat{\boldsymbol{\theta}}^{\rm Biased,EB}$ with $C_1 = 1, C_2 = 0$ perform the best; while for $n = 5$, $\hat{\boldsymbol{\theta}}^{\rm Bayes,EB}$ and $\hat{\boldsymbol{\theta}}^{\rm Biased,EB}$ with $C_1 = 0, C_2 = 1$ perform the best.

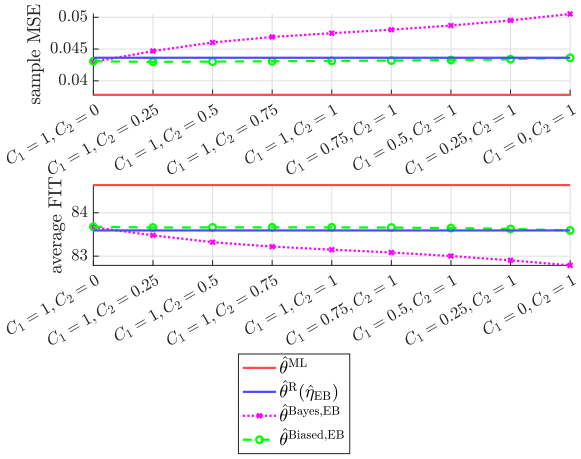

Fig. 1: Sample means of the sample MSE and the average FIT for $n = 1$ and $N = 5$.

For larger n and N, the influence of different ratios of C_1 and C_2 on the performance of $\hat{\boldsymbol{\theta}}^{\rm Bayes,EB}$ and $\hat{\boldsymbol{\theta}}^{\rm Biased,EB}$ becomes weaker. In Fig. 3 and Table[2] I, we consider $n = 80$, $N = 360$, and $M_s = 5 \times 10^3$. We can observe that the

[2]For $\hat{\boldsymbol{\theta}}^{\rm Bayes,EB}$ and $\hat{\boldsymbol{\theta}}^{\rm Biased,EB}$, we first calculate the total computing time, and the sample means of the average FIT and the sample MSE over 100 collections of test systems and data. Then, we calculate the sample means of these three statistics over different combinations of C_1 and C_2.

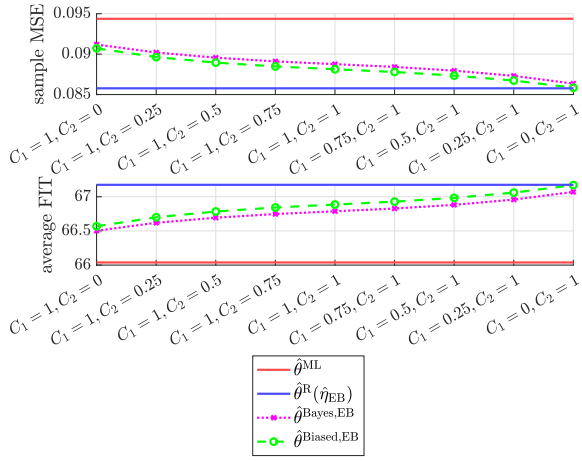

Fig. 2: Sample means of the sample MSE and the average FIT for $n = 5$ and $N = 15$.

performance of $\hat{\boldsymbol{\theta}}^{\rm R}(\hat{\eta}_{\rm EB})$ is quite close to that of $\hat{\boldsymbol{\theta}}^{\rm Bayes,EB}$ and $\hat{\boldsymbol{\theta}}^{\rm Biased,EB}$, while its computing time is over twice that of $\hat{\boldsymbol{\theta}}^{\rm Bayes,EB}$ and over 500 times that of $\hat{\boldsymbol{\theta}}^{\rm Biased,EB}$.

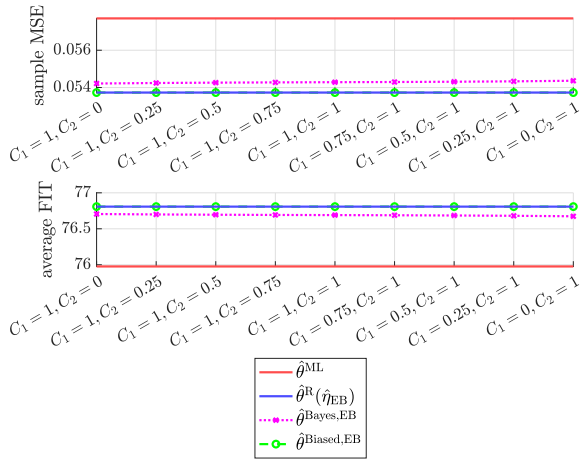

Fig. 3: Sample means of the sample MSE and the average FIT for $n = 80$ and $N = 360$.

TABLE I: Sample means of the sample MSE, the average FIT, and the total computing time for $n = 80$ and $N = 360$.

	$\hat{\boldsymbol{\theta}}^{\rm R}(\hat{\eta}_{\rm EB})$	$\hat{\boldsymbol{\theta}}^{\rm Bayes,EB}$	$\hat{\boldsymbol{\theta}}^{\rm Biased,EB}$
sample MSE	5.37×10^{-2}	5.43×10^{-2}	5.37×10^{-2}
average FIT	76.81	76.69	76.81
computing time (s)	1.81×10^3	6.68×10^2	3.51

REFERENCES

[1] T. Chen and L. Ljung. Implementation of algorithms for tuning parameters in regularized least squares problems in system identification. *Automatica*, 49:2213–2220, 2013.

[2] Y. Ju, B. Wahlberg, and H. Hjalmarsson. Excess mean squared error of empirical Bayes estimators. *arXiv e-prints: 2503.11863*, 2025.

[3] L. Ljung. *System Identification Toolbox for Use with MATLAB*. The Math Works, 1995.

[4] G. Pillonetto, T. Chen, A. Chiuso, G. De Nicolao, and L. Ljung. *Regularized System Identification: Learning Dynamic Models from Data*. Springer Nature, 2022.