Automatica 94 (2018) 381–395

Contents lists available at ScienceDirect

Automatica

journal homepage: www.elsevier.com/locate/automatica

On asymptotic properties of hyperparameter estimators for kernel-based regularization methods☆

Biqiang Mu [a], Tianshi Chen [b,c,*], Lennart Ljung [a]

[a] Division of Automatic Control, Department of Electrical Engineering, Linköping University, Linköping SE-58183, Sweden
[b] School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, China
[c] Shenzhen Research Institute of Big Data, The Chinese University of Hong Kong, Shenzhen, China

ARTICLE INFO

Article history:
Received 2 July 2017
Received in revised form 13 January 2018
Accepted 4 April 2018

Keywords:
Kernel-based regularization
Empirical Bayes
Stein's unbiased risk estimator
Asymptotic analysis

ABSTRACT

The kernel-based regularization method has two core issues: kernel design and hyperparameter estimation. In this paper, we focus on the second issue and study the properties of several hyperparameter estimators including the empirical Bayes (EB) estimator, two Stein's unbiased risk estimators (SURE) (one related to impulse response reconstruction and the other related to output prediction) and their corresponding Oracle counterparts, with an emphasis on the asymptotic properties of these hyperparameter estimators. To this goal, we first derive and then rewrite the first order optimality conditions of these hyperparameter estimators, leading to several insights on these hyperparameter estimators. Then we show that as the number of data goes to infinity, the two SUREs converge to the best hyperparameter minimizing the corresponding mean square error, respectively, while the more widely used EB estimator converges to another best hyperparameter minimizing the expectation of the EB estimation criterion. This indicates that the two SUREs are asymptotically optimal in the corresponding MSE senses but the EB estimator is not. Surprisingly, the convergence rate of two SUREs is slower than that of the EB estimator, and moreover, unlike the two SUREs, the EB estimator is independent of the convergence rate of $\Phi^T\Phi/N$ to its limit, where Φ is the regression matrix and N is the number of data. A Monte Carlo simulation is provided to demonstrate the theoretical results.

1. Introduction

The kernel-based regularization methods (KRM) from machine learning and statistics were first introduced to the system identification community in Pillonetto and De Nicolao (2010) and then further developed in Chen, Andersen, Ljung, Chiuso, and Pillonetto (2014), Chen, Ohlsson, and Ljung (2012) and Pillonetto, Chiuso, and De Nicolao (2011). These methods attract increasing attention in the community and have become a complement to the classical maximum likelihood/prediction error methods (ML/PEM) (Chen et al., 2012; Ljung, Singh, & Chen, 2015; Pillonetto & Chiuso, 2015). In particular, KRM may have better average accuracy and robustness than ML/PEM when the data is short and/or has low signal-to-noise ratio (SNR).

There are two core issues for KRM: kernel design and hyperparameter estimation. The former is regarding how to parameterize the kernel matrix with a parameter vector, called hyperparameter, to embed the prior knowledge of the system to be identified, and the latter is regarding how to estimate the hyperparameter based on the data such that the resulting model estimator achieves a good bias–variance trade-off or equivalently, suitably balances the adherence to the data and the model complexity.

The kernel design plays a similar role as the model structure design for ML/PEM and determines the underlying model structure for KRM. In the past few years, many efforts have been spent on this issue and several kernels have been invented to embed various types of prior knowledge, e.g., Carli, Chen, and Ljung (2017), Chen (2018a), Chen et al. (2014), Chen et al. (2016), Chen et al. (2012), Chen and Pillonetto (2018), Dinuzzo (2015), Marconato, Schoukens, and Schoukens (2016), Pillonetto, Chen, Chiuso, Nicolao, and Ljung (2016), Pillonetto et al. (2011), Pillonetto and De Nicolao (2010) and Zorzi and Chiuso (2017). In particular, two systematic kernel design methods (one is from a machine learning perspective and the other one is from a system theory perspective) were developed in Chen (2018b) by embedding the corresponding type of prior knowledge.

☆ The material in this paper was partially presented at the 20th World Congress of the International Federation of Automatic Control, July 9–14, 2017, Toulouse, France. This paper was recommended for publication in revised form by Associate Editor Thomas Bo Schön under the direction of Editor Torsten Söderström.

* Corresponding author: Tianshi Chen, School of Science and Engineering and Shenzhen Research Institute of Big Data, The Chinese University of Hong Kong, Shenzhen, China.
E-mail addresses: biqiang.mu@liu.se (B. Mu), tschen@cuhk.edu.cn (T. Chen), ljung@isy.liu.se (L. Ljung).

https://doi.org/10.1016/j.automatica.2018.04.035
0005-1098/

The hyperparameter estimation plays a similar role as the model order selection in ML/PEM and its essence is to determine a suitable model complexity based on the data. As mentioned in the survey of KRM (Pillonetto, Dinuzzo, Chen, De Nicolao, & Ljung, 2014), many methods can be used for hyperparameter estimation, such as the cross-validation (CV), empirical Bayes (EB), C_p statistics and Stein's unbiased risk estimator (SURE) and etc. In contrast with the numerous results on kernel design, there are however few results on hyperparameter estimation except Aravkin, Burke, Chiuso, & Pillonetto (2012a, b, 2014), Chen et al. (2014) and Pillonetto & Chiuso (2015). In Aravkin et al. (2012a, b, 2014), two types of diagonal kernel matrices are considered. When $\Phi^T\Phi/N$ is an identity matrix, where Φ is the regression matrix and N is the number of data, the optimal hyperparameter estimate of the EB estimator has explicit form and is shown to be consistent in terms of the mean square error (MSE). When $\Phi^T\Phi/N$ is not an identity matrix, the EB estimator is shown to asymptotically minimize a weighted MSE. In Chen et al. (2014), the EB with linear multiple kernel is shown to be a difference of convex programming problem and moreover, the optimal hyperparameter estimate is sparse. In Pillonetto and Chiuso (2015), the robustness of the EB estimator is analysed.

In this paper, we study the properties of the EB estimator and two SUREs in Pillonetto and Chiuso (2015) with an emphasis on the asymptotic properties of these hyperparameter estimators. In particular, we are interested in the following questions: When the number of data goes to infinity,

(1) what will be the best kernel matrix, or equivalently, the best value of the hyperparameter?
(2) which estimator (method) shall be chosen such that the hyperparameter estimate tends to this best value in the given sense?
(3) what will be the convergence rate of that the hyperparameter estimate tends to this best value? and what factors does this rate depend on?

In order to answer these questions, we employ the regularized least squares method for FIR model estimation in Chen et al. (2012). As a motivation, we first show that the regularized least squares estimate can have smaller MSE than the least squares estimate for any data length if the kernel matrix is chosen carefully. We then derive the first order optimality conditions of these hyperparameter estimators and their corresponding Oracle counterparts (relying on the true impulse response, see Section 3.2 for details). These first order optimality conditions are then rewritten in a way to better expose their relations, leading to several insights on these hyperparameter estimators. For instance, one insight is that for the Oracle estimators, for any data length, and without structure constraints on the kernel matrix, the optimal kernel matrices are same as the one in Chen et al. (2012) and equal to the outer product of the vector of the true impulse response and its transpose. Moreover, explicit solutions of the optimal hyperparameter estimate for two special cases are derived accordingly. Then we turn to the asymptotic analysis of these hyperparameter estimators. Regardless of the parameterization of the kernel matrix, we first show that the two SUREs actually converge to the best hyperparameter minimizing the corresponding MSE, respectively, as the number of data goes to infinity, while the more widely used EB estimator converges to the best hyperparameter minimizing the expectation of the EB estimation criterion. In general, these best hyperparameters are different from each other except for some special cases. This means that the two SUREs are asymptotically optimal in the corresponding MSE senses but the EB estimator is not. We then show that the convergence rate of two SUREs is slower than that of the EB estimator, and moreover, unlike the two SUREs, the EB estimator is independent of the convergence rate of $\Phi^T\Phi/N$ to its limit.

The remaining parts of the paper is organized as follows. In Section 2, we recap the regularized least squares method for FIR model estimation and introduce two types of MSE. In Section 3, we introduce six hyperparameter estimators, including the EB estimator, two SUREs, and their corresponding Oracle counterparts. In Section 4, we derive the first order optimality conditions of these hyperparameter estimators and put them in a form that clearly shows their relation, leading to several insights. In Section 5, we give the asymptotic analysis of these hyperparameter estimators, including the asymptotic convergence and the corresponding convergence rate. In Section 6, we illustrate our theoretical results with Monte Carlo simulations. Finally, we conclude this paper in Section 7. All proofs of the theoretical results are postponed to Appendix A.

2. Regularized least squares approach for FIR model estimation

2.1. Regularized least squares and two types of MSEs

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

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

where t is the time index, $y(t), u(t), v(t)$ are the output, input and disturbance of the system at time t, respectively, $G_0(q)$ is the rational transfer function of the system and q is the forward shift operator: $qu(t) = u(t+1)$. Assume that the input $u(t)$ is known (deterministic) and the input–output data are collected at time instants $t = 1, \dots, N$, and moreover, the disturbance $v(t)$ is a zero mean white noise with finite variance $\sigma^2 > 0$. The problem is to estimate a model for $G_0(q)$ as well as possible based on the available data $\{u(t-1), y(t)\}_{t=1}^N$.

The transfer function $G_0(q)$ can be written as

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

where $g_k^0, k = 1, \dots, \infty$ form the impulse response of the system. Since the impulse response coefficients $\{g_k^0\}$ of the stable rational transfer function $G_0(q)$ decay exponentially, it is possible to truncate the infinite impulse response at a sufficiently high order, leading to the finite impulse response (FIR) model:

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

With the FIR model (3), system (1) is now written as

$$y(t) = \phi^T(t)\theta + v(t),\ t = 1, \dots, N$$

where $\phi(t) = [u(t-1), \dots, u(t-n)]^T \in \mathbb{R}^n$, and its matrix–vector form is

$$Y = \Phi\theta + V,\ \text{where} \tag{4}$$
$$Y = [y(1)\, y(2) \cdots y(N)]^T$$
$$\Phi = [\phi(1)\, \phi(2)\ \cdots\ \phi(N)]^T$$
$$V = [v(1)\, v(2)\ \cdots\ v(N)]^T.$$

The well-known least squares (LS) estimator

$$\widehat{\theta}^{\text{LS}} = \operatorname*{argmin}_{\theta\in\mathbb{R}^n} \|Y - \Phi\theta\|^2 \tag{5a}$$
$$= (\Phi^T\Phi)^{-1}\Phi^T Y \tag{5b}$$

where $\|\cdot\|$ is the Euclidean norm, is unbiased with respect to the FIR model (4) but may have large variance and mean square error

(MSE) (e.g., when the input is low-pass filtered white noise). The large variance can be mitigated if some bias is allowed and traded for smaller variance and smaller MSE.

One possible way to achieve this goal is to add a regularization term $\sigma^2\theta^T P^{-1}\theta$ in the LS criterion (5a), leading to the regularized least squares (RLS) estimator:

$$\widehat{\theta}^{\mathrm{R}} = \underset{\theta\in\mathbb{R}^n}{\operatorname{argmin}} \|Y-\Phi\theta\|^2 + \sigma^2\theta^T P^{-1}\theta \tag{6a}$$

$$= P\Phi^T(\Phi P\Phi^T + \sigma^2 I_N)^{-1}Y \tag{6b}$$

where P is symmetric and positive semidefinite and is called the kernel matrix ($\sigma^2 P^{-1}$ is often called the regularization matrix), and I_N is the N-dimensional identity matrix.

Remark 1. As is well known, the RLS estimator (6b) has a Bayesian interpretation. Specifically, assume that θ and $v(t)$ are independent and Gaussian distributed with

$$\theta \sim \mathcal{N}(0, P), \quad v(t) \sim \mathcal{N}(0, \sigma^2) \tag{7}$$

where P is the prior covariance matrix. Then θ and Y are jointly Gaussian distributed and moreover, the posterior distribution of θ given Y is

$$\theta|Y \sim \mathcal{N}(\widehat{\theta}^{\mathrm{R}}, \widehat{P}^{\mathrm{R}})$$
$$\widehat{\theta}^{\mathrm{R}} = P\Phi^T(\Phi P\Phi^T + \sigma^2 I_N)^{-1}Y$$
$$\widehat{P}^{\mathrm{R}} = P - P\Phi^T(\Phi P\Phi^T + \sigma^2 I_N)^{-1}\Phi P.$$

Two types of MSE could be used to evaluate the performance of the RLS estimator (6b). The first one is the MSE related to the impulse response reconstruction, see e.g., Chen et al. (2012) and Pillonetto and Chiuso (2015),

$$\mathrm{MSEg}(P) = E\|\widehat{\theta}^{\mathrm{R}} - \theta_0\|^2 \tag{8}$$

where $\theta_0 = [g_1^0, \ldots, g_n^0]^T$ with g_i^0, $i = 1, \ldots, n$, defined in (2) and $E(\cdot)$ is the mathematical expectation with respect to the noise distribution. The second one is the MSE related to output prediction, see e.g., Pillonetto and Chiuso (2015),

$$\mathrm{MSEy}(P) = E\left[\sum_{t=1}^{N}\left(\phi^T(t)\theta_0 + v^*(t) - \widehat{y}(t)\right)^2\right] \tag{9}$$

where $\widehat{y}(t) = \phi^T(t)\widehat{\theta}^{\mathrm{R}}$ and $v^*(t)$ is an independent copy of the noise $v(t)$. Interestingly, the two MSEs (8) and (9) are related with each other through

$$\mathrm{MSEy}(P) = \mathrm{Tr}\left(E(\widehat{\theta}^{\mathrm{R}}-\theta_0)(\widehat{\theta}^{\mathrm{R}}-\theta_0)^T\Phi^T\Phi\right) + N\sigma^2 \tag{10}$$

where $\mathrm{Tr}(\cdot)$ is the trace of a square matrix. Moreover, they have explicit expressions, which are given in the following proposition.

Proposition 1. *For a given kernel matrix P, the two MSEs (8) and (9) take the following form*

$$\mathrm{MSEg}(P) = \|P\Phi^T Q^{-1}\Phi\theta_0 - \theta_0\|^2 + \sigma^2\mathrm{Tr}(P\Phi^T Q^{-1}Q^{-T}\Phi P^T) \tag{11a}$$

$$\mathrm{MSEy}(P) = \|\Phi P\Phi^T Q^{-1}\Phi\theta_0 - \Phi\theta_0\|^2 + N\sigma^2 + \sigma^2\mathrm{Tr}(\Phi P\Phi^T Q^{-1}Q^{-T}\Phi P^T\Phi^T) \tag{11b}$$

where A^{-T} means $(A^{-1})^T$ for a non-singular matrix A and

$$Q = \Phi P\Phi^T + \sigma^2 I_N. \tag{12}$$

2.2. RLS estimator can outperform LS estimator

It is interesting to investigate whether the RLS estimator (6b) with a suitable choice of the kernel matrix P can have smaller MSEs (8) and (9) than the LS estimator (5b). The answer is affirmative for MSEg (8) and for the ridge regression case, where $P^{-1} = (\beta/\sigma^2)I_n$ with $\beta > 0$, Hoerl and Kennard (1970) and Theobald (1974). In what follows, we further show that this property also holds for more general P for MSEg (8) and MSEy (9).

Proposition 2. *Consider the RLS estimator (6b) and the LS estimator (5b). Suppose that $P^{-1} = \beta A/\sigma^2$, where $\beta > 0$ and A is symmetric and positive semidefinite. Then for the given A, there exists $\beta > 0$ such that (6b) has a smaller MSEg (8) and MSEy (9) than (5b). Moreover, if A is positive definite, then (6b) has a smaller MSEg (8) and MSEy (9) than (5b) whenever $0 < \beta < 2\sigma^2/(\theta_0^T A\theta_0)$.*

Proposition 2 shows that for any data length N, the RLS estimator (6b) can have smaller MSEg (8) and MSEy (9) than the LS estimator (5b) with a sufficiently small regularization "in any direction" and this merit motivates to further explore the potential of the RLS estimator (6b) by careful design of the kernel matrix P. It is worth to note the paper (Zorzi, 2017) also shows the Bayesian estimator is still optimal when a priori information is known with some uncertainty.

3. Design of kernel matrix and hyperparameter estimation

The regularization method has two core issues: kernel matrix design, namely parameterization of the kernel matrix by a parameter vector, called hyperparameter, and hyperparameter estimation.

3.1. Parameterization of kernel matrix

For efficient regularization, the symmetric and positive semidefinite kernel matrix P has to be chosen carefully. It is typically done by postulating a parameterized family of matrices

$$P(\eta), \quad \eta \in \Omega \subset \mathbb{R}^p \tag{13}$$

where η is called the *hyperparameter* and the feasible set Ω of η is assumed to be compact. The choice of parameterization is a trade-off of the same kind as the choice of model class in identification: On one hand it should be a large and flexible class to allow as much benefits from regularization as possible. On the other hand, a large set requires larger dimensions of η, and the estimation of these comes with their own penalties (much in the spirit of the Akaike's criterion). Since P is the prior covariance of the true impulse response, the prior knowledge of the underlying system to be identified, e.g., exponential stability and smoothness, should be embedded in the parameterized matrix $P(\eta)$.

A popular way to achieve this goal is through a parameterized positive semidefinite kernel function. So far, several kernels have been invented, such as the stable spline (SS) kernel (Pillonetto & De Nicolao, 2010), the diagonal correlated (DC) kernel and the tuned-correlated (TC) kernel (Chen et al., 2012), which are defined as follows:

$$\text{SS}: \quad P_{kj}(\eta) = c\left(\frac{\alpha^{k+j+\max(k,j)}}{2} - \frac{\alpha^{3\max(k,j)}}{6}\right)$$
$$\eta = [c, \alpha] \in \Omega = \{c \geq 0, 0 \leq \alpha \leq 1\}; \tag{14a}$$

$$\text{DC}: \quad P_{kj}(\eta) = c\alpha^{(k+j)/2}\rho^{|j-k|}$$
$$\eta = [c, \alpha, \rho] \in \Omega = \{c \geq 0, 0 \leq \alpha \leq 1, |\rho| \leq 1\}; \tag{14b}$$

$$\text{TC}: \quad P_{kj}(\eta) = c\alpha^{\max(k,j)}$$
$$\eta = [c, \alpha] \in \Omega = \{c \geq 0, 0 \leq \alpha \leq 1\}. \tag{14c}$$

Remark 2. For nonlinearly parameterized kernels, e.g., the kernels in (14), the optimal hyperparameter should be located in the interior of the set Ω. To justify this, we take the DC kernel as an example. For the case where either $c = 0$ or $\alpha = 0$, $P(\eta) = 0$ and thus (6b) is trivially 0. For the case where $\alpha = 1$, this violates the stability of the system. For the case where $|\rho| = 1$, the coefficients of the impulse response are perfectly positive or negative correlated, but this is impossible for a stable system. In fact, more formal justification regarding this issue can be found in Pillonetto and Chiuso (2015, p. 115), which shows that the measure of the set containing all optimal estimates lying on the boundary of Ω is zero and thus can be neglected when making almost sure convergence statement.

3.2. Hyperparameter estimation

Once a parameterized family of the kernel matrix $P(\eta)$ has been chosen, the task is to estimate, or "tune", the hyperparameter η based on the data.

Several methods are suggested in the literature, see e.g., Section 14 of Pillonetto et al. (2014), including the empirical Bayes (EB) and SURE methods. The EB method uses the Bayesian interpretation in Remark 1. It follows from the assumption (7) that Y is Gaussian with mean zero and covariance matrix Q. As a result, it is possible to estimate the hyperparameter η by maximizing the (marginal) likelihood of Y, i.e.,

$$\text{EB}: \widehat{\eta}_{\text{EB}} = \operatorname*{argmin}_{\eta\in\Omega} \mathscr{F}_{\text{EB}}(P(\eta)) \tag{15a}$$

$$\mathscr{F}_{\text{EB}}(P) = Y^T Q^{-1} Y + \log\det(Q) \tag{15b}$$

where Q is defined in (12) and $\det(\cdot)$ denotes the determinant of a square matrix. The SURE method first constructs a Stein's unbiased risk estimator (SURE) of the MSE and then estimates the hyperparameter by minimizing the constructed estimator. Two variants of the SURE method were considered in Pillonetto and Chiuso (2015), which construct the SUREs for MSEg(P) in (11a) and MSEy(P) in (11b), and are referred to as SUREg and SUREy, respectively:

$$\begin{aligned}\mathscr{F}_{\text{Sg}}(P) &= \|\widehat{\theta}^{\text{LS}} - \widehat{\theta}^{\text{R}}\|^2 + \sigma^2\text{Tr}\big(2R^{-1} - (\Phi^T\Phi)^{-1}\big)\\ &= \sigma^4 Y^T Q^{-T}\Phi(\Phi^T\Phi)^{-2}\Phi^T Q^{-1}Y\\ &\quad + \sigma^2\text{Tr}\big(2R^{-1} - (\Phi^T\Phi)^{-1}\big)\end{aligned} \tag{16a}$$

$$\begin{aligned}\mathscr{F}_{\text{Sy}}(P) &= \|Y - \Phi\widehat{\theta}^{\text{R}}\|^2 + 2\sigma^2\text{Tr}\big(\Phi P\Phi^T Q^{-1}\big)\\ &= \sigma^4 Y^T Q^{-T} Q^{-1} Y + 2\sigma^2\text{Tr}\big(\Phi P\Phi^T Q^{-1}\big)\end{aligned} \tag{16b}$$

where

$$R = \Phi^T\Phi + \sigma^2 P^{-1}. \tag{17}$$

Then the hyperparameter η is estimated by minimizing the SUREg (16a) and SUREy (16b):

$$\text{SUREg}: \widehat{\eta}_{\text{Sg}} = \operatorname*{argmin}_{\eta\in\Omega} \mathscr{F}_{\text{Sg}}(P(\eta)) \tag{18a}$$

$$\text{SUREy}: \widehat{\eta}_{\text{Sy}} = \operatorname*{argmin}_{\eta\in\Omega} \mathscr{F}_{\text{Sy}}(P(\eta)). \tag{18b}$$

In the following sections, we will study the properties of the above three estimators EB, SUREg and SUREy. To set reference for these estimators, we introduce their corresponding Oracle counterparts that depend on the true impulse response θ_0:

$$\begin{aligned}\text{MSEg}: \widehat{\eta}_{\text{MSEg}} &= \operatorname*{argmin}_{\eta\in\Omega} E[\mathscr{F}_{\text{Sg}}(P(\eta))]\\ &= \operatorname*{argmin}_{\eta\in\Omega} \text{MSEg}(P(\eta))\end{aligned} \tag{19a}$$

$$\begin{aligned}\text{MSEy}: \widehat{\eta}_{\text{MSEy}} &= \operatorname*{argmin}_{\eta\in\Omega} E[\mathscr{F}_{\text{Sy}}(P(\eta))]\\ &= \operatorname*{argmin}_{\eta\in\Omega} \text{MSEy}(P(\eta))\end{aligned} \tag{19b}$$

$$\begin{aligned}\text{EEB}: \quad \widehat{\eta}_{\text{EEB}} &= \operatorname*{argmin}_{\eta\in\Omega} E[\mathscr{F}_{\text{EB}}(P(\eta))]\\ &= \operatorname*{argmin}_{\eta\in\Omega} \text{EEB}(P(\eta))\end{aligned} \tag{19c}$$

where MSEg(P) and MSEy(P) are defined in (11a) and (11b), respectively, and

$$\text{EEB}(P) = \theta_0^T\Phi^T Q^{-1}\Phi\theta_0 + \sigma^2\text{Tr}(Q^{-1}) + \log\det(Q). \tag{20}$$

The hyperparameter estimators (19a) and (19b) give the optimal hyperparameter estimates for any data length in the corresponding MSE sense and thus provide reference when evaluating the performance of hyperparameter estimators.

Remark 3. Among these hyperparameter estimators, only SUREg (16a) depends on $(\Phi^T\Phi)^{-1}$. When $(\Phi^T\Phi)^{-1}$ is ill-conditioned, SUREg (16a) should be avoided for hyperparameter estimation. One may also note that $(\Phi^T\Phi)^{-1}$ in the second term is independent of P and thus can actually be removed in the calculation.

Remark 4. It is interesting to note that the first terms of $\mathscr{F}_{\text{Sg}}(P)$, $\mathscr{F}_{\text{Sy}}(P)$, and $\mathscr{F}_{\text{EB}}(P)$ given in (16a), (16b), and (15b) contain the same factors Y and Q^{-1}. Moreover, similar to (10), $\mathscr{F}_{\text{Sg}}(P)$ and $\mathscr{F}_{\text{Sy}}(P)$ are related with each other through

$$\begin{aligned}\mathscr{F}_{\text{Sy}}(P) = \text{Tr}\Big\{&\big[(\widehat{\theta}^{\text{LS}} - !\widehat{\theta}^{\text{R}})(\widehat{\theta}^{\text{LS}} - \widehat{\theta}^{\text{R}})^T\\ &+ \sigma^2(2R^{-1} - (\Phi^T\Phi)^{-1})\big]\Phi^T\Phi\Big\}\\ &+ \underbrace{Y^T\Phi(\Phi^T\Phi)^{-1}\Phi^T Y - Y^T Y - n\sigma^2}_{\text{independent of the kernel matrix } P}.\end{aligned} \tag{21}$$

In what follows, we will investigate the properties of the hyperparameter estimators EB, SUREg, and SUREy and their corresponding Oracle estimators EEB, MSEg and MSEy. Before proceeding to the details, we make, without loss of generality, the following assumption.

Assumption 1. The hyperparameter estimates $\widehat{\eta}_{\text{Sg}}$, $\widehat{\eta}_{\text{Sy}}$, $\widehat{\eta}_{\text{EB}}$, $\widehat{\eta}_{\text{MSEg}}$, $\widehat{\eta}_{\text{MSEy}}$ and $\widehat{\eta}_{\text{EEB}}$ are interior points of Ω.

4. Properties of hyperparameter estimators: finite data case

In this section, focusing on the finite data case we first give the first order optimality conditions of the hyperparameter estimators and then we consider two special cases for which closed-form expressions of the optimal hyperparameter estimates are available.

4.1. First order optimality conditions

The hyperparameter estimates $\widehat{\eta}_{\text{Sg}}$, $\widehat{\eta}_{\text{Sy}}$, and $\widehat{\eta}_{\text{EB}}$ in (18a), (18b), and (15a) should satisfy the first order optimality conditions if they are interior points of Ω. For convenience, we let $\mathscr{C}$ to denote one of the following estimation criteria $\mathscr{F}_{\text{Sg}}$, $\mathscr{F}_{\text{Sy}}$, $\mathscr{F}_{\text{EB}}$, MSEg, MSEy or EEB. Then the corresponding hyperparameter estimate is a root of the system of equations:

$$\frac{\partial\mathscr{C}(P(\eta))}{\partial\eta} = 0. \tag{22}$$

By the chain rule of compound functions, we have

$$\text{Tr}\left(\frac{\partial\mathscr{C}(P)}{\partial P}\left(\frac{\partial P(\eta)}{\partial\eta_i}\right)^T\right) = 0,\ \ 1 \le i \le p \tag{23}$$

where, for convenience, when calculating $\frac{\partial\mathscr{C}(P)}{\partial P}$ the symmetry of P is ignored according to Lemma B1, that is, the elements of P are treated independently. One merit of using $\frac{\partial\mathscr{C}(P)}{\partial P}$ without imposing the structure information on P is that explicit expressions for the estimation criteria (16a), (16b), and (15b) are available.

Proposition 3. *The first order partial derivatives of (16a), (16b), and (15b) with respect to P are, respectively,*

$$\frac{\partial \mathscr{F}_{\mathrm{Sg}}(P)}{\partial P} = -2\sigma^4\Phi^T Q^{-T}\Phi(\Phi^T\Phi)^{-2}\Phi^T Q^{-1}YY^T Q^{-T}\Phi + 2\sigma^4 H^{-T}\overline{H}^{-T} \tag{24a}$$

$$\frac{\partial \mathscr{F}_{\mathrm{Sy}}(P)}{\partial P} = -2\sigma^4\Phi^T Q^{-T}Q^{-1}YY^T Q^{-T}\Phi + 2\sigma^4\Phi^T Q^{-T}Q^{-T}\Phi \tag{24b}$$

$$\frac{\partial \mathscr{F}_{\mathrm{EB}}(P)}{\partial P} = -\Phi^T Q^{-T}YY^T Q^{-T}\Phi + \Phi^T Q^{-T}\Phi \tag{24c}$$

where

$$H = P\Phi^T\Phi + \sigma^2 I_n,\ \overline{H} = \Phi^T\Phi P + \sigma^2 I_n. \tag{25}$$

Similarly, the partial derivatives of MSEg(P), MSEy(P), and EEB(P) with respect to P are also available.

Proposition 4. *The first order partial derivatives of (11a), (11b), and (20) with respect to P are, respectively,*

$$\frac{\partial \mathrm{MSEg}(P)}{\partial P} = -2\sigma^4 H^{-T}H^{-1}\theta_0\theta_0^T\Phi^T Q^{-T}\Phi + 2\sigma^4 H^{-T}H^{-1}P\Phi^T Q^{-T}\Phi \tag{26a}$$

$$\frac{\partial \mathrm{MSEy}(P)}{\partial P} = -2\sigma^4\Phi^T Q^{-T}Q^{-1}\Phi\theta_0\theta_0^T\Phi^T Q^{-T}\Phi + 2\sigma^4\Phi^T Q^{-T}Q^{-1}\Phi P\Phi^T Q^{-T}\Phi \tag{26b}$$

$$\frac{\partial \mathrm{EEB}(P)}{\partial P} = -\Phi^T Q^{-T}\Phi\theta_0\theta_0^T\Phi^T Q^{-T}\Phi + \Phi^T Q^{-T}\Phi P^T\Phi^T Q^{-T}\Phi \tag{26c}$$

where H is defined in (25).

In order to better expose the relation among the partial derivatives derived in Propositions 3 and 4, we define

$$S = P + \sigma^2(\Phi^T\Phi)^{-1}. \tag{27}$$

With the use of (27) and the identities (B.11)–(B.13) in the Appendix, we rewrite the partial derivatives derived in Propositions 3 and 4 as follows.

Corollary 1. *The partial derivatives derived in Propositions 3 and 4 can be rewritten as follows:*

$$\frac{\partial \mathrm{MSEg}(P)}{\partial P} = 2\sigma^4 S^{-T}(\Phi^T\Phi)^{-2}S^{-1}(P-\theta_0\theta_0^T)S^{-T} \tag{28a}$$

$$\frac{\partial \mathscr{F}_{\mathrm{Sg}}(P)}{\partial P} = 2\sigma^4 S^{-T}(\Phi^T\Phi)^{-2}S^{-1}\big(S-\widehat{\theta}^{\mathrm{LS}}(\widehat{\theta}^{\mathrm{LS}})^T\big)S^{-T} \tag{28b}$$

$$\frac{\partial \mathrm{MSEy}(P)}{\partial P} = 2\sigma^4 S^{-T}(\Phi^T\Phi)^{-1}S^{-1}(P-\theta_0\theta_0^T)S^{-T} \tag{28c}$$

$$\frac{\partial \mathscr{F}_{\mathrm{Sy}}(P)}{\partial P} = 2\sigma^4 S^{-T}(\Phi^T\Phi)^{-1}S^{-1}\big(S-\widehat{\theta}^{\mathrm{LS}}(\widehat{\theta}^{\mathrm{LS}})^T\big)S^{-T} \tag{28d}$$

$$\frac{\partial \mathrm{EEB}(P)}{\partial P} = S^{-T}(P^T-\theta_0\theta_0^T)S^{-T} \tag{28e}$$

$$\frac{\partial \mathscr{F}_{\mathrm{EB}}(P)}{\partial P} = S^{-T}\big(S^T-\widehat{\theta}^{\mathrm{LS}}(\widehat{\theta}^{\mathrm{LS}})^T\big)S^{-T}. \tag{28f}$$

It follows from Corollary 1 that the difference between the partial derivatives of $\mathscr{F}_{\mathrm{Sg}}(P)$, $\mathscr{F}_{\mathrm{Sy}}(P)$, $\mathscr{F}_{\mathrm{EB}}(P)$ and that of their Oracle counterparts is that the factor $S-\widehat{\theta}^{\mathrm{LS}}(\widehat{\theta}^{\mathrm{LS}})^T$ is replaced by $P-\theta_0\theta_0^T$. Moreover, the difference between the partial derivative of $\mathscr{F}_{\mathrm{Sg}}(P)$ and that of $\mathscr{F}_{\mathrm{Sy}}(P)$ is that there is one extra factor $(\Phi^T\Phi)^{-1}$. The difference between the first order derivative of $\mathscr{F}_{\mathrm{Sy}}(P)$ and that of $\mathscr{F}_{\mathrm{EB}}(P)$ is that there is one extra factor $2\sigma^4(\Phi^T\Phi)^{-1}S^{-1} = 2\sigma^4 H^{-1}$. The above relations extend to the partial derivatives of their Oracle counterparts.

Remark 5. It is important to note from Propositions 3 and 4 that only the first term of $\frac{\partial \mathscr{F}_{\mathrm{Sg}}(P)}{\partial P}$ depends on the possibly ill-conditioned $(\Phi^T\Phi)^{-1}$. With the use of S in (27), all partial derivatives of the hyperparameter estimators seemingly depend on the possibly ill-conditioned term $(\Phi^T\Phi)^{-1}$. However, it should be stressed that the partial derivatives derived in Corollary 1 are not intended for numerical calculation but for theoretical analysis and for better exposition of the relation among the partial derivatives derived in Propositions 3 and 4.

Remark 6. We keep the transpose notation for symmetric matrices in Propositions 3 and 4 and Corollary 1 because they are derived by using Lemma B1 without imposing the symmetric assumption on P. After we have derived $\frac{\partial \mathscr{C}(P)}{\partial P}$ and made the symmetric assumption on P, the first optimality condition (23) can be written as

$$\mathrm{Tr}\left(\frac{\partial \mathscr{C}(P)}{\partial P}\frac{\partial P(\eta)}{\partial \eta_i}\right) = 0,\ 1 \le i \le p$$

where the transpose notation appearing in $\frac{\partial \mathscr{C}(P)}{\partial P}$ can be dropped for symmetric matrices, e.g. $S^T = S$, and $Q^{-T}Q^{-1}$ can be written as Q^{-2}.

Setting $\frac{\partial \mathrm{MSEg}(P)}{\partial P} = 0$, $\frac{\partial \mathrm{MSEg}(P)}{\partial P} = 0$, and $\frac{\partial \mathrm{EEB}(P)}{\partial P} = 0$ in Corollary 1 leads to the next proposition.

Proposition 5. *The optimal kernel matrix that minimizes* MSEg(P), MSEy(P), *and* EEB(P) *without structure constraints on P is*

$$P = \theta_0\theta_0^T. \tag{29}$$

It was found in Chen et al. (2012) that (29) minimizes the MSE matrix $E(\widehat{\theta}^{\mathrm{R}}-\theta_0)(\widehat{\theta}^{\mathrm{R}}-\theta_0)^T$ in the matrix sense. Here we further find that (29) is optimal for MSEg(P), MSEy(P) and EEB(P), and for any data length N.

4.2. Two special cases

In general, there is no explicit expression of these hyperparameter estimators. However, there exist some special cases, for which it is possible to derive the explicit solution based on Corollary 1. In the following, we consider two special cases.

4.2.1. Ridge regression with $\Phi^T\Phi = NI_n$

Let $P(\eta) = \eta I_n$ with $\eta \ge 0$ and assume $\Phi^T\Phi = NI_n$. Then we have the following result.

Proposition 6. *Consider $P(\eta) = \eta I_n$ with $\eta \ge 0$. Further assume that $\Phi^T\Phi = NI_n$. Then we have*

$$\widehat{\eta}_{\mathrm{Sg}} = \widehat{\eta}_{\mathrm{Sy}} = \widehat{\eta}_{\mathrm{EB}} = \max\Big(0, \frac{(\widehat{\theta}^{\mathrm{LS}})^T\widehat{\theta}^{\mathrm{LS}}}{n} - \frac{\sigma^2}{N}\Big). \tag{30}$$

Moreover,

$$\widehat{\eta}_{\mathrm{MSEg}} = \widehat{\eta}_{\mathrm{MSEy}} = \widehat{\eta}_{\mathrm{EEB}} = \theta_0^T\theta_0/n. \tag{31}$$

Remark 7. It is worth noting that the optimal hyperparameter $\theta_0^T\theta_0/n$ holds for any N. Moreover, one has

$$\mathrm{MSEg}(\theta_0^T\theta_0/nI_n) = \frac{n\sigma^2}{N + n\sigma^2/(\theta_0^T\theta_0)} < \frac{n\sigma^2}{N}$$

where $n\sigma^2/N$ is the MSEg of the LS estimator (5b). This means that the ridge regression with $P = \theta_0^T\theta_0/nI_n$ has a smaller MSEg than the LS estimator (5b) when $\Phi^T\Phi = NI_n$. Finally, (30) is a consistent estimate of $\theta_0^T\theta_0/n$ if $\widehat{\theta}^{\mathrm{LS}} \to \theta_0$ as $N \to \infty$.

4.2.2. Diagonal kernel matrix with $\Phi^T\Phi = NI_n$

Let $P(\eta)$ be a diagonal kernel matrix (in this case we have $p = n$.), i.e.,

$$P(\eta) = \text{diag}[\eta_1, \ldots, \eta_n] \text{ with } \eta_i \geq 0,\ 1 \leq i \leq n \tag{32}$$

where $\eta_1, \ldots, \eta_n$ are the main diagonal elements of the diagonal matrix $\text{diag}[\eta_1, \ldots, \eta_n]$. Then under the assumption $\Phi^T\Phi = NI_n$, we have the following result.

Proposition 7. *Consider $P(\eta)$ in (32) and assume that $\Phi^T\Phi = NI_n$. Then we have*

$$\widehat{\eta}_{\text{Sg}} = \widehat{\eta}_{\text{Sy}} = \widehat{\eta}_{\text{EB}} = \left[\max\{0, \widehat{g}_1^2 - \sigma^2/N\}, \cdots, \max\{0, \widehat{g}_n^2 - \sigma^2/N\}\right]^T \tag{33}$$

where $\widehat{g}_i$ is the ith element of the LS estimate (5b), $i = 1, \ldots, n$. Moreover,

$$\widehat{\eta}_{\text{MSEg}} = \widehat{\eta}_{\text{MSEy}} = \widehat{\eta}_{\text{EEB}} = \left[(g_1^0)^2, \ldots, (g_n^0)^2\right]^T. \tag{34}$$

Remark 8. In the papers Aravkin et al. (2012b, 2014), the linear model (4) but with a slightly different setting is considered, where the parameter θ is partitioned into m sub-vectors $\theta = [\theta^{(1)^T}, \ldots, \theta^{(m)^T}]^T$ and the dimension of $\theta^{(i)}$ is n_i so that $n = \sum_{i=1}^m n_i$. In addition, the prior distribution of $\theta^{(i)}$ is set to be $\mathcal{N}(0, \eta_i I_{n_i})$ and η_i is an independent and identically distributed exponential random variable with probability density $p_\gamma(\eta_i) = \gamma \exp(-\gamma\eta_i)\chi(\eta_i)$ where γ is a positive scalar and $\chi(t) = 1$ for $t \geq 0$ and 0 otherwise. Under the setting given above, the solution maximizing the marginal posterior of η given the data and the optimal solution of the MSEg are derived in Aravkin et al. (2012b, 2014) when $\Phi^T\Phi = NI_n$. When $m = 1, n_1 = n, \gamma = 0$, their estimates become (30) and (31), respectively. When $n = m, n_i = 1$ for $i = 1, \ldots, n$, and $\gamma = 0$, their solutions become (33) and (34), respectively. In contrast, we study here the SUREg, SUREy, MSEy, and EEB estimators besides the EB and MSEg estimators and find their solutions are the same under the simplified setting. Clearly, $\max\{0, \widehat{g}_i^2 - \sigma^2/N\}$ is a consistent estimator of $(g_i^0)^2, i = 1, \ldots, n$.

5. Properties of hyperparameter estimators: when the data length goes to infinity

In this section, we investigate the asymptotic properties of these hyperparameter estimators. For this purpose, it is useful to first consider the asymptotic property of the partial derivatives derived in Corollary 1. Noting the finding of Corollary 1 and that $S - \widehat{\theta}^{\text{LS}}(\widehat{\theta}^{\text{LS}})^T$ converges to $P - \theta_0\theta_0^T$ under proper conditions, we can derive the following proposition.

Proposition 8. *Consider the partial derivatives derived in Corollary 1. Assume that P is nonsingular and $\Phi^T\Phi/N \to \Sigma$ almost surely as $N \to \infty$, where Σ is positive definite. Then we have as $N \to \infty$*

$$N^2\frac{\partial \text{MSEg}(P)}{\partial P} \to 2\sigma^4 P^{-T}\Sigma^{-2}P^{-1}(P - \theta_0\theta_0^T)P^{-T} \tag{35a}$$

$$N^2\frac{\partial \mathscr{F}_{\text{Sg}}(P)}{\partial P} \to 2\sigma^4 P^{-T}\Sigma^{-2}P^{-1}(P - \theta_0\theta_0^T)P^{-T} \tag{35b}$$

$$N\frac{\partial \text{MSEy}(P)}{\partial P} \to 2\sigma^4 P^{-T}\Sigma^{-1}P^{-1}(P - \theta_0\theta_0^T)P^{-T} \tag{35c}$$

$$N\frac{\partial \mathscr{F}_{\text{Sy}}(P)}{\partial P} \to 2\sigma^4 P^{-T}\Sigma^{-1}P^{-1}(P - \theta_0\theta_0^T)P^{-T} \tag{35d}$$

$$\frac{\partial \text{EEB}(P)}{\partial P} \to P^{-T}(P^T - \theta_0\theta_0^T)P^{-T} \tag{35e}$$

$$\frac{\partial \mathscr{F}_{\text{EB}}(P)}{\partial P} \to P^{-T}(P^T - \theta_0\theta_0^T)P^{-T} \tag{35f}$$

almost surely.

Proposition 8 shows that the three pairs, $N^2\frac{\partial \text{MSEg}(P)}{\partial P}$ and $N^2\frac{\partial \mathscr{F}_{\text{Sg}}(P)}{\partial P}$, $N\frac{\partial \text{MSEy}(P)}{\partial P}$ and $N\frac{\partial \mathscr{F}_{\text{Sy}}(P)}{\partial P}$, and $\frac{\partial \text{EEB}(P)}{\partial P}$ and $\frac{\partial \mathscr{F}_{\text{EB}}(P)}{\partial P}$, have respectively the same limit as N goes to ∞. This observation motivates to explore if this property also holds for the estimation criteria of these hyperparameter estimators. The answer is affirmative and we have the following result.

Proposition 9. *Consider the hyperparameter estimation criteria SUREg (16a), SUREy (16b), and EB (15b), and their corresponding Oracle counterparts MSEg (11a), MSEy (11b), and EEB (20). Assume that P is nonsingular and $\Phi^T\Phi/N \to \Sigma$ almost surely as $N \to \infty$, where Σ is positive definite. Then we have as $N \to \infty$*

$$N^2(\text{MSEg}(P) - \sigma^2\text{Tr}((\Phi^T\Phi)^{-1})) \to W_g(P, \Sigma, \theta_0) \tag{36a}$$

$$N^2(\mathscr{F}_{\text{Sg}}(P) - \sigma^2\text{Tr}((\Phi^T\Phi)^{-1})) \to W_g(P, \Sigma, \theta_0) \tag{36b}$$

$$N(\text{MSEy}(P) - (n+N)\sigma^2) \to W_y(P, \Sigma, \theta_0) \tag{36c}$$

$$N(\mathscr{F}_{\text{Sy}}(P) + Y^T\Phi(\Phi^T\Phi)^{-1}\Phi^TY - Y^TY - 2n\sigma^2) \to W_y(P, \Sigma, \theta_0) \tag{36d}$$

$$\text{EEB}(P) - (N-n) - (N-n)\log\sigma^2 - \log\det(\Phi^T\Phi) \to W_{\text{B}}(P, \theta_0) \tag{36e}$$

$$\mathscr{F}_{\text{EB}}(P) + Y^T\Phi(\Phi^T\Phi)^{-1}\Phi^TY/\sigma^2 - Y^TY/\sigma^2 - (N-n)\log\sigma^2 - \log\det(\Phi^T\Phi) \to W_{\text{B}}(P, \theta_0) \tag{36f}$$

almost surely, where

$$W_g(P, \Sigma, \theta_0) = \sigma^4\theta_0^TP^{-1}\Sigma^{-2}P^{-1}\theta_0 - 2\sigma^4\text{Tr}\left(\Sigma^{-1}P^{-1}\Sigma^{-1}\right) \tag{37a}$$

$$W_y(P, \Sigma, \theta_0) = \sigma^4\theta_0^TP^{-1}\Sigma^{-1}P^{-1}\theta_0 - 2\sigma^4\text{Tr}\left(\Sigma^{-1}P^{-1}\right) \tag{37b}$$

$$W_{\text{B}}(P, \theta_0) = \theta_0^TP^{-1}\theta_0 + \log\det(P). \tag{37c}$$

Remark 9. For these hyperparameter estimation criteria, $W_g(P, \Sigma, \theta_0)$, $W_y(P, \Sigma, \theta_0)$ and $W_B(P, \theta_0)$ contain all information about the asymptotic benefits of regularization: how it depends on any kernel matrix P, any true impulse response vector θ_0 and any stationary properties of the input covariance matrix Σ.

Proposition 9 enables us to derive asymptotic properties of these hyperparameter estimators for any parameterization $P(\eta)$ of the kernel matrix. Moreover, it also implies that the estimators $\widehat{\eta}_{\text{Sg}}$, $\widehat{\eta}_{\text{Sy}}$, and $\widehat{\eta}_{\text{EB}}$ possibly share the same limits with their corresponding Oracle counterparts $\widehat{\eta}_{\text{MSEg}}$, $\widehat{\eta}_{\text{MSEy}}$, and $\widehat{\eta}_{\text{EEB}}$, respectively.

To state the result, we need an extra assumption. It is worth to note that the limit functions $W_g(P(\eta), \Sigma, \theta_0)$, $W_y(P(\eta), \Sigma, \theta_0)$ and $W_{\text{B}}(P(\eta), \theta_0)$ may not have a unique global minimum, respectively. In this case, the analysis of how minimizing elements of a sequence of functions $M_N(\eta)$ converge to the minimizing element of the limit function $\lim M_N(\eta)$, i.e.,

$$\text{“}\lim \arg\min M_N(\eta) = \arg\min \lim M_N(\eta)\text{”} \tag{38}$$

where $M_N(\eta)$ denotes any function on the left hand side of “$\to$” in (36), follows the same idea as for prediction error identification methods, see, e.g. Lemma 8.2 and Theorem 8.2 in Ljung (1999). Accordingly, it is useful in this context to let “arg min” denote the set of minimizing arguments in case where $W_g(P(\eta), \Sigma, \theta_0)$, $W_y(P(\eta), \Sigma, \theta_0)$ and $W_{\text{B}}(P(\eta), \theta_0)$ do not have a unique global minimum, respectively,:

$$\arg\min_{\eta\in\Omega} M(\eta) = \left\{\eta | \eta \in \Omega, M(\eta) = \min_{\eta'\in\Omega} M(\eta')\right\} \tag{39}$$

where $M(\eta)$ could be any one of $W_g(P(\eta), \Sigma, \theta_0)$, $W_y(P(\eta), \Sigma, \theta_0)$ and $W_{\text{B}}(P(\eta), \theta_0)$.

Now we define

$$\eta_g^* = \arg\min_{\eta\in\Omega} W_g(P(\eta),\ \Sigma, \theta_0) \tag{40}$$

$$\eta_y^* = \operatorname*{argmin}_{\eta\in\Omega} W_y(P(\eta),\ \Sigma, \theta_0) \tag{41}$$

$$\eta_B^* = \operatorname*{argmin}_{\eta\in\Omega} W_B(P(\eta), \theta_0). \tag{42}$$

Remark 10. The optimal hyperparameter η_B^* has the following interpretation: under the assumption $\theta_0 \sim \mathcal{N}(0, P(\eta))$, η_B^* maximizes the value of the probability density function of θ_0.

The following assumption is also needed.

Assumption 2. The sets η_g^*, η_y^* and η_B^* consist of interior points of Ω, and are discrete, i.e., made up of only isolated points, respectively.

Then we have the following theorem.

Theorem 1. *Assume that $P(\eta)$ is any parameterization of the kernel matrix such that $P(\eta)$ is positive definite and moreover, $\Phi^T\Phi/N \to \Sigma$ almost surely as $N \to \infty$, where Σ is positive definite. Then we have as $N \to \infty$*

$$\widehat{\eta}_{\mathrm{MSEg}} \to \eta_g^*,\ \ \widehat{\eta}_{\mathrm{Sg}} \to \eta_g^* \tag{43a}$$

$$\widehat{\eta}_{\mathrm{MSEy}} \to \eta_y^*,\ \ \widehat{\eta}_{\mathrm{Sy}} \to \eta_y^* \tag{43b}$$

$$\widehat{\eta}_{\mathrm{EEB}} \to \eta_B^*,\ \ \widehat{\eta}_{\mathrm{EB}} \to \eta_B^* \tag{43c}$$

almost surely. Moreover, η_g^*, η_y^*, and η_B^* are roots of the system of equations, respectively, $i = 1, \ldots, p$:

$$\mathrm{Tr}\Big(P(\eta)^{-1}\Sigma^{-2}P(\eta)^{-1}\big(P(\eta)-\theta_0\theta_0^T\big)P(\eta)^{-1}\frac{\partial P(\eta)}{\partial \eta_i}\Big) = 0$$

$$\mathrm{Tr}\Big(P(\eta)^{-1}\Sigma^{-1}P(\eta)^{-1}\big(P(\eta)-\theta_0\theta_0^T\big)P(\eta)^{-1}\frac{\partial P(\eta)}{\partial \eta_i}\Big) = 0$$

$$\mathrm{Tr}\Big(P(\eta)^{-1}\big(P(\eta)-\theta_0\theta_0^T\big)P(\eta)^{-1}\frac{\partial P(\eta)}{\partial \eta_i}\Big) = 0.$$

The Oracle estimators $\widehat{\eta}_{\mathrm{MSEg}}$ and $\widehat{\eta}_{\mathrm{MSEg}}$ are optimal for any data length N in the average sense if we are concerned with the ability to reproduce the true impulse response and predict the future outputs of the system respectively, while the SUREg $\widehat{\eta}_{\mathrm{Sg}}$ and the SUREy $\widehat{\eta}_{\mathrm{Sy}}$ are not optimal in general. Surprisingly, a nice property of $\widehat{\eta}_{\mathrm{Sg}}$ and $\widehat{\eta}_{\mathrm{Sy}}$ is that they converge to the best possible hyperparameter η_g^* and η_y^*, respectively, for any chosen parameterized kernel matrix $P(\eta)$. It is so to speak that the two SURE methods are "asymptotically consistent or asymptotically optimal". This means that when N is sufficiently large, $\widehat{\eta}_{\mathrm{Sg}}$ and $\widehat{\eta}_{\mathrm{Sy}}$ perform as well as $\widehat{\eta}_{\mathrm{MSEg}}$ and $\widehat{\eta}_{\mathrm{MSEy}}$, respectively. It is also worth noting that even with increasing number of data the EB estimator $\widehat{\eta}_{\mathrm{EB}}$ has another preference than to minimize MSEg and MSEy.

Remark 11. In contrast with $W_g(P, \Sigma, \theta_0)$ and $W_y(P, \Sigma, \theta_0)$, a unique property of $W_B(P, \theta_0)$ is that it does not depend on the limit Σ of $\Phi^T\Phi/N$. This can to some extent explain why the EB estimator is more robust than the SUREg and SUREy especially when $\Phi^T\Phi$ is ill-conditioned. Interested readers can find experimental evidence for this in Pillonetto and Chiuso (2015). However, in contrast with the SUREg and SUREy, the EB estimator is not asymptotically optimal in the MSEg/MSEy sense.

Remark 12. The different expressions of the limit functions $W_g(P(\eta), \Sigma, \theta_0)$, $W_y(P(\eta), \Sigma, \theta_0)$, and $W_B(P(\eta), \theta_0)$ imply that the optimal hyperparameters η_g^*, η_y^*, and η_B^* may be different. To check this, we consider a special case: $P = \eta K$, where $\eta > 0$ and K is fixed and positive definite. In this case, (40), (41) and (42) become

$$\begin{aligned}\eta_g^* &= \arg\min_{\eta\geq 0} \frac{\sigma^4}{\eta^2}\theta_0^T K^{-1}\Sigma^{-2}K^{-1}\theta_0 - \frac{2\sigma^4}{\eta}\mathrm{Tr}(\Sigma^{-1}K^{-1}\Sigma^{-1})\\ &= \frac{\theta_0^T K^{-1}\Sigma^{-2}K^{-1}\theta_0}{\mathrm{Tr}(\Sigma^{-1}K^{-1}\Sigma^{-1})}\\ \eta_y^* &= \arg\min_{\eta\geq 0} \frac{\sigma^4}{\eta^2}\theta_0^T K^{-1}\Sigma^{-1}K^{-1}\theta_0 - \frac{2\sigma^4}{\eta}\mathrm{Tr}(\Sigma^{-1}K^{-1})\\ &= \frac{\theta_0^T K^{-1}\Sigma^{-1}K^{-1}\theta_0}{\mathrm{Tr}(\Sigma^{-1}K^{-1})}\\ \eta_B^* &= \arg\min_{\eta\geq 0} \theta_0^T K^{-1}\theta_0/\eta + \log\eta^n + \log\det(K)\\ &= \theta_0^T K^{-1}\theta_0/n\end{aligned}$$

which shows that η_g^*, η_y^* and η_B^* can be different. Clearly, when $K = I_n$ and $\Sigma = dI_n$ with $d > 0$, $\eta_g^* = \eta_y^* = \eta_B^*$. For this case, the optimal value η_B^* has been given in Theorem 7.3 of Pillonetto and Chiuso (2015).

Corollary 2. *Assume that $\Phi^T\Phi/N \to dI_n$ almost surely with $d > 0$ and $P(\eta)$ is any positive definite parameterization of the kernel matrix. Then we have*

$$\eta_g^* = \eta_y^* = \operatorname*{argmin}_{\eta\in\Omega} \theta_0^T P(\eta)^{-2}\theta_0 - 2\mathrm{Tr}(P(\eta)^{-1})$$

$$\eta_B^* = \operatorname*{argmin}_{\eta\in\Omega} \theta_0^T P(\eta)^{-1}\theta_0 + \log\det(P(\eta))$$

and further η_g^* and η_B^* are roots of the following system of equations, respectively:

$$\mathrm{Tr}\Big(P(\eta)^{-2}\big(P(\eta)-\theta_0\theta_0^T\big)P(\eta)^{-1}\frac{\partial P(\eta)}{\partial \eta_i}\Big) = 0,\ i = 1, \ldots, p$$

$$\mathrm{Tr}\Big(P(\eta)^{-1}\big(P(\eta)-\theta_0\theta_0^T\big)P(\eta)^{-1}\frac{\partial P(\eta)}{\partial \eta_i}\Big) = 0,\ i = 1, \ldots, p.$$

In addition, for the diagonal kernel matrix (32), we have

$$\eta_g^* = \eta_y^* = \eta_B^* = \left[(g_1^0)^2, \ldots, (g_n^0)^2\right]^T.$$

In Theorem 1, we have considered the convergence of those hyperparameter estimators. In fact, we can further derive their corresponding convergence rate. To this end, we let $\xi_N = o_p(a_N)$ denote that the sequence $\{\xi_N/a_N\}$ for nonzero sequence $\{a_N\}$ converges in probability to zero, i.e., $\forall\epsilon > 0$, $P(|\xi_N/a_N| > \epsilon) \to 0$ as $N \to \infty$, while $\xi_N = O_p(a_N)$ denote that $\{\xi_N/a_N\}$ is bounded in probability, i.e., $\forall\epsilon > 0$, $\exists L > 0$ such that $P(|\xi_N/a_N| > L) < \epsilon$, $\forall N$. Then we have the following theorem.

Theorem 2. *Assume that $\|\Phi^T\Phi/N - \Sigma\| = O_p(\delta_N)$, where $\|\cdot\|$ denotes the Frobenius norm for a square matrix, $\delta_N \to 0$ as $N \to \infty$ and $P(\eta)$ is any positive definite parameterization of the kernel matrix. Then we have*

$$\|\widehat{\eta}_{\mathrm{MSEg}} - \eta_g^*\| = O_p(\varpi_N),\ \ \|\widehat{\eta}_{\mathrm{Sg}} - \eta_g^*\| = O_p(\mu_N) \tag{44a}$$

$$\|\widehat{\eta}_{\mathrm{MSEy}} - \eta_y^*\| = O_p(\varpi_N),\ \ \|\widehat{\eta}_{\mathrm{Sy}} - \eta_y^*\| = O_p(\mu_N) \tag{44b}$$

$$\|\widehat{\eta}_{\mathrm{EEB}} - \eta_B^*\| = O_p(1/N),\ \ \|\widehat{\eta}_{\mathrm{EB}} - \eta_B^*\| = O_p(1/\sqrt{N}) \tag{44c}$$

where

$$\varpi_N = \max\big(O_p(\delta_N), O_p(1/N)\big) \tag{45a}$$

$$\mu_N = \max\big(O_p(\delta_N), O_p(1/\sqrt{N})\big). \tag{45b}$$

Theorem 2 shows that the convergence rate of $\widehat{\eta}_{\mathrm{EEB}}$ and $\widehat{\eta}_{\mathrm{EB}}$ to η_B^* depends only on the fact $\Phi^T\Phi/N \to \Sigma$ as $N \to \infty$ ($\Phi^T\Phi = O_p(N)$) but not on the rate $\|\Phi^T\Phi/N - \Sigma\| = O_p(\delta_N)$. Moreover, we have

- the convergence rate of $\widehat{\eta}_{\text{EEB}}$ to η_{B}^* is faster than that of $\widehat{\eta}_{\text{MSEg}}$ to η_{g}^* and that of $\widehat{\eta}_{\text{MSEy}}$ to η_{y}^*.
- the convergence rate of $\widehat{\eta}_{\text{EB}}$ to η_{B}^* is faster than that of $\widehat{\eta}_{\text{Sg}}$ to η_{g}^* and that of $\widehat{\eta}_{\text{Sy}}$ to η_{y}^*.
- the convergence rate of $\widehat{\eta}_{\text{MSEg}}$, $\widehat{\eta}_{\text{MSEy}}$ and $\widehat{\eta}_{\text{EEB}}$ to η_{g}^*, η_{y}^* and η_{B}^*, respectively, is faster than that of $\widehat{\eta}_{\text{Sg}}$, $\widehat{\eta}_{\text{Sy}}$ and $\widehat{\eta}_{\text{EB}}$ to η_{g}^*, η_{y}^* and η_{B}^*, respectively.

Theorem 2 has the following corollary.

Corollary 3. *Assume that* $\|\Phi^T\Phi/N - \Sigma\| = O_p(\delta_N)$, *where* $\delta_N \to 0$ *as* $N \to \infty$ *and* $P(\eta)$ *is any positive definite parameterization of the kernel matrix. Then*

$$\|\widehat{\eta}_{\text{MSEg}} - \widehat{\eta}_{\text{Sg}}\| = O_p(\mu_N) \tag{46a}$$

$$\|\widehat{\eta}_{\text{MSEy}} - \widehat{\eta}_{\text{Sy}}\| = O_p(\mu_N) \tag{46b}$$

$$\|\widehat{\eta}_{\text{EEB}} - \widehat{\eta}_{\text{EB}}\| = O_p(1/\sqrt{N}) \tag{46c}$$

where μ_N *is defined in* (45b).

This corollary shows that the convergence rate of $\|\widehat{\eta}_{\text{EEB}} - \widehat{\eta}_{\text{EB}}\|$ to zero is faster than that of $\|\widehat{\eta}_{\text{MSEg}} - \widehat{\eta}_{\text{Sg}}\|$ and $\|\widehat{\eta}_{\text{MSEy}} - \widehat{\eta}_{\text{Sy}}\|$ to zero.

6. Numerical simulation

In this section, we illustrate the theoretical results with numerical simulation.

6.1. Test data-bank

The method in Chen et al. (2012) and Pillonetto and Chiuso (2015) is used to generate 1000 30th order test systems. Then for each test system, we consider four different test inputs:

- The first two test inputs are implemented by the MATLAB command `idinput` choosing the bandlimited white Gaussian noise with normalized bands [0, 0.6] and [0, 1], respectively, and denoted by IT1 and IT2, respectively.
- The third and fourth test inputs are the white Gaussian noise of unit variance filtered by a second order rational transfer function $1/(1 - aq^{-1})^2$ with a chosen to be 0.95 and 0.05, respectively, and denoted by IT3 and IT4, respectively.

To generate the data set, we simulate each system with one of the four test inputs to get the output, which is then corrupted by an additive white Gaussian noise. 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 kept the same for the four test inputs.

Finally, in order to test the finite sample and asymptotic behaviour of the hyperparameter estimators, we consider data sets with different data lengths $N = 500$ and 8000, respectively.

6.2. Simulation setup

The performance of the RLS estimator (6b) is evaluated by the measure of fit (Ljung, 2012) defined as follows :

$$\text{Fit} = 100 \times \left(1 - \frac{\|\widehat{\theta} - \theta_0\|}{\|\theta_0 - \bar{\theta}_0\|}\right), \quad \bar{\theta}_0 = \frac{1}{n}\sum_{k=1}^{n} g_k^0$$

where n is set to 200. This fit is actually to evaluate the RLS estimator in the MSEg sense.

Here the unknown inputs $u_{-1}, \ldots, u_{-n+1}$ are not used (nonwindowed). The TC kernel (14c) is considered and its hyperparameter $\eta = [c, \alpha]^T$ is estimated by using the estimators SUREg (18a), SUREy (18b), and EB (15a), respectively. For reference, we also consider their corresponding Oracle counterparts, i.e., the estimators MSEg (19a), MSEy (19b), and EEB (19c), respectively. The notations Sg, Sy, EB, MSEg, MSEy, and EEB are used to denote the corresponding simulation results, respectively.

Table 1
Average fits for 1000 test systems and data sets.

	MSEg	Sg	MSEy	Sy	EEB	EB
IT1						
$N = 500$	80.34	−2.4E9	78.07	53.83	77.98	77.26
$N = 8000$	90.63	−8.6E8	88.08	78.39	88.39	88.36
IT2						
$N = 500$	87.11	84.46	87.02	86.03	86.60	86.16
$N = 8000$	96.67	96.60	96.67	96.60	96.47	96.44
IT3						
$N = 500$	46.95	−2220	41.61	−146.4	39.47	39.03
$N = 8000$	57.67	−176.8	53.63	38.86	51.05	50.86
IT4						
$N = 500$	86.78	83.89	86.69	85.66	86.24	85.84
$N = 8000$	96.57	96.49	96.56	96.49	96.38	96.35

6.3. Simulation results

The average fits are given in Table 1. The boxplots of the 1000 fits for IT1 and IT2 are displayed in Figs. 1–2, respectively. The boxplots for IT3 and IT4 are skipped because of their similarity with IT1 and IT2.

6.4. Findings

Firstly, for all tested cases and in terms of average accuracy and robustness, the Oracle estimators MSEg and MSEy (not implementable in practice) are better than Sg and Sy, respectively, while EB is just a little bit worse than but very close to its Oracle estimator EEB.

Secondly, we consider the cases with input IT1, where $\Phi^T\Phi$ is very ill-conditioned for both $N = 500$ and $N = 8000$. In this case and in terms of average accuracy and robustness, Sg performs badly because it depends on $(\Phi^T\Phi)^{-1}$. Moreover, Sy is better than Sg, but worse than EB.

Thirdly, we consider the case with input IT2 and $N = 500$, where $\Phi^T\Phi$ is much better conditioned than the cases with input IT1. In this case and in terms of average accuracy and robustness, Sg behaves much better in contrast with the cases with input IT1. Moreover, EB and Sy are quite close though EB is a little bit better, and they are all better than Sg.

Lastly, we consider the case with input IT2 and $N = 8000$, where $\Phi^T\Phi$ is very well-conditioned and in terms of average accuracy and robustness, Sg behaves much better in contrast with all the other cases, and performs as well as Sy and better than EB. Moreover, Sg and Sy are very close to the corresponding Oracle estimators MSEg and MSEy. These observations coincide with the results found in Theorem 1 and Corollary 2. Namely, Sg and Sy are asymptotically optimal but EB is not in the MSEg/MSEy senses and moreover, Sg and Sy give the same optimal hyperparameter estimate as their Oracle counterparts MSEg and MSEy, because the limit $\Sigma = I_n$ of $\Phi^T\Phi/N$ as $N \to \infty$. It can also be seen from Figs. 1 and 2 that the boxplots of EEB and EB are closer than that of MSEg and Sg and that of MSEy and Sy. This observation coincides with the result found in Corollary 3, that is, the convergence rate of $\|\widehat{\eta}_{\text{EEB}} - \widehat{\eta}_{\text{EB}}\|$ to zero is faster than that of $\|\widehat{\eta}_{\text{MSEg}} - \widehat{\eta}_{\text{Sg}}\|$ and $\|\widehat{\eta}_{\text{MSEy}} - \widehat{\eta}_{\text{Sy}}\|$ to zero.

Based on the theoretical and simulation results, we have the following suggestions for choosing the hyperparameter estimators:

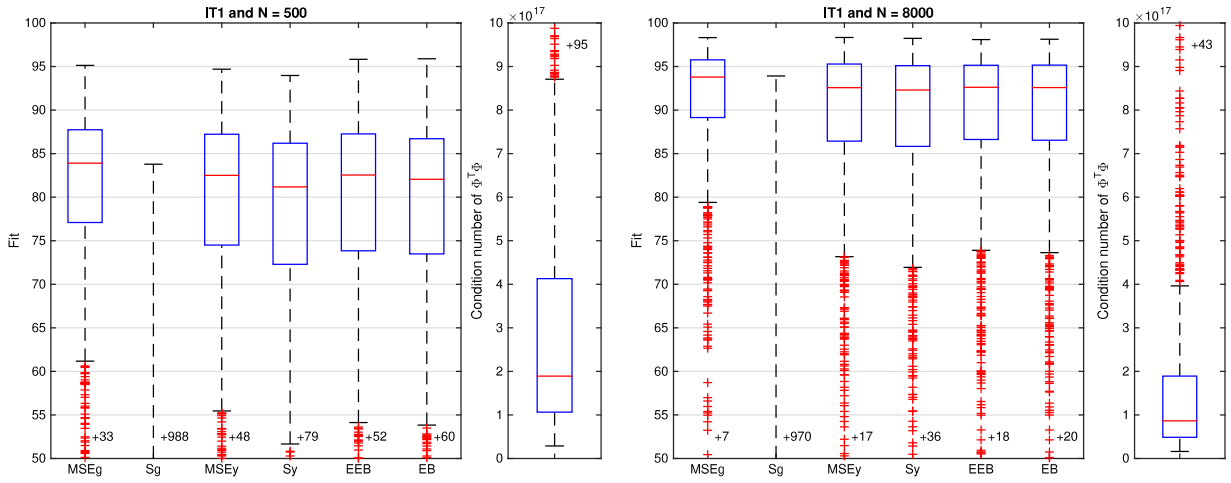


Fig. 1. Boxplot of the 1000 fits for the bandlimited white Gaussian noise input with the normalized band [0, 0.6] and boxplot of the condition numbers of the matrix $\Phi^T\Phi$: data lengths $N = 500$ (left) and $N = 8000$ (right).

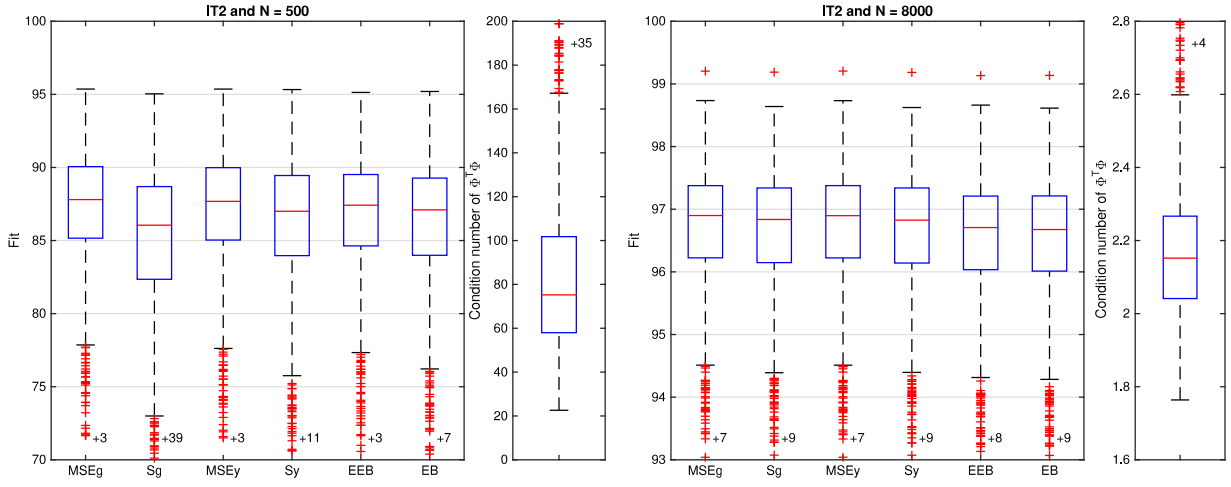


Fig. 2. Boxplot of the 1000 fits for the bandlimited white Gaussian noise input with the normalized band [0, 1] and boxplot of the condition numbers of the matrix $\Phi^T\Phi$: data lengths $N = 500$ (left) and $N = 8000$ (right).

(i) When the regression matrix is well-conditioned and the data is sufficiently long, the two SUREs should be used since they are asymptotically optimal;
(ii) When the regression matrix is ill-conditioned or the data is short, the EB estimator should be used.

7. Conclusions

Kernel matrix design and hyperparameter estimation are two core issues for the kernel based regularization methods. In contrast with the former issue, there are few results reported for the latter issue. In this paper, we focused on the latter issue and studied the properties of several hyperparameter estimators including the empirical Bayes (EB) estimator, two Stein's unbiased risk estimators (SURE) and their corresponding Oracle counterparts, with an emphasis on the asymptotic properties of these hyperparameter estimators. Our major results are the following:

- The first order optimality conditions of these hyperparameter estimators are put in similar forms that better expose their relation and lead to several insights on these hyperparameter estimators.
- As the number of data goes to infinity, the two SUREs converge to the best hyperparameter minimizing the corresponding mean square error, respectively, while the more widely used EB estimator converges to another best hyperparameter minimizing the expectation of the EB estimation criterion. This indicates that the two SUREs are asymptotically optimal in the corresponding MSE sense but the EB estimator is not.
- The convergence rate of two SUREs is slower than that of the EB estimator, and moreover, unlike the two SUREs, the EB estimator is independent of the convergence rate of $\Phi^T\Phi/N$ to its limit.

The results enhance our understanding about these hyperparameter estimators and are one step forward towards the goal of building a theory of the hyperparameter estimation for the kernel-based regularization methods.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China under contract Nos. 61773329 and 61603379,

the Thousand Youth Talents Plan funded by the central government of China, the Shenzhen Research Projects Ji-20170189 and Ji-20160207 funded by the Shenzhen Science and Technology Innovation Council, the Presidential Fund PF. 01.000249 and the start-up grant 2014.0003.23 funded by the Chinese University of Hong Kong, Shenzhen, a research grant for junior researchers under contract No. 2014-5894 funded by Swedish Research Council, the National Key Basic Research Program of China (973 Program) under contract No. 2014CB845301, and the Presidential Fund of the Academy of Mathematics and Systems Science, CAS under contract No. 2015-hwyxqnrc-mbq.

Appendix A

Appendix A contains the proof of the results in the paper, for which the technical lemmas are placed in Appendix B. The proofs of Propositions 1, 5, 7, 8 and Corollaries 1, 2, and 3 are straightforward and thus omitted.

A.1. Proof of Proposition 2

Under the setting $P^{-1} = \beta A/\sigma^2$, the MSEg (11a) of the RLS estimator (6b) is a function of β for a given A:

$$\text{MSEg}(\beta) = \text{Bias}(\beta) + \text{Var}(\beta) \text{ where} \tag{A.1}$$
$$\text{Bias}(\beta) = \beta^2\theta_0^T A(\Phi^T\Phi + \beta A)^{-1}(\Phi^T\Phi + \beta A)^{-1}A\theta_0$$
$$\text{Var}(\beta) = \sigma^2\text{Tr}\big((\Phi^T\Phi + \beta A)^{-1}\Phi^T\Phi(\Phi^T\Phi + \beta A)^{-1}\big).$$

Note that $\text{MSEg}(0) = \sigma^2\text{Tr}\big((\Phi^T\Phi)^{-1}\big)$ corresponds to the MSEg of the LS estimator (5b). The derivatives of $\text{Bias}(\beta)$ and $\text{Var}(\beta)$ with respect to β are as follows:

$$\begin{aligned}\frac{\text{dBias}(\beta)}{\text{d}\beta} &= 2\beta\theta_0^T A(\Phi^T\Phi + \beta A)^{-1}(\Phi^T\Phi + \beta A)^{-1}A\theta_0 \\ &\quad - 2\beta^2\theta_0^T A(\Phi^T\Phi + \beta A)^{-1}A(\Phi^T\Phi + \beta A)^{-1} \\ &\quad \times (\Phi^T\Phi + \beta A)^{-1}A\theta_0 \end{aligned}\tag{A.2}$$

$$\begin{aligned}\frac{\text{dVar}(\beta)}{\text{d}\beta} &= -2\sigma^2\text{Tr}\big((\Phi^T\Phi + \beta A)^{-1}A(\Phi^T\Phi + \beta A)^{-1} \\ &\quad \times \Phi^T\Phi(\Phi^T\Phi + \beta A)^{-1}\big)\end{aligned}\tag{A.3}$$

where the formula $\frac{\text{dC}^{-1}(\beta)}{\text{d}\beta} = -C^{-1}(\beta)\frac{\text{dC}(\beta)}{\text{d}\beta}C^{-1}(\beta)$ for an invertible matrix $C(\beta)$ is used. Then we have

$$\frac{\text{dBias}(\beta)}{\text{d}\beta}\Big|_{\beta\to 0^+} = 0$$
$$\frac{\text{dVar}(\beta)}{\text{d}\beta}\Big|_{\beta\to 0^+} = -2\sigma^2\text{Tr}\big((\Phi^T\Phi)^{-1}A(\Phi^T\Phi)^{-1}\big) < 0.$$

Therefore, we have $\frac{\text{dMSEg}(\beta)}{\text{d}\beta}\Big|_{\beta\to 0^+} < 0$. This means that $\text{MSEg}(\beta) < \text{MSEg}(0)$ in some small right neighbourhood of the origin $\beta = 0$.

Under the assumption that A is positive definite, denote

$$M(\beta) \triangleq E(\widehat{\theta}^{\text{R}} - \theta_0)(\widehat{\theta}^{\text{R}} - \theta_0)^T.$$

We first prove $M(0) - M(\beta) > 0$ for $0 < \beta < 2\sigma^2/(\theta_0^T A\theta_0)$. A straightforward calculation gives

$$\begin{aligned}&M(0) - M(\beta) \\ &= \sigma^2(\Phi^T\Phi)^{-1} - \sigma^2(\Phi^T\Phi + \beta A)^{-1}\Phi^T\Phi(\Phi^T\Phi + \beta A)^{-1} \\ &\quad - \beta^2(\Phi^T\Phi + \beta A)^{-1}A\theta_0\theta_0^T A(\Phi^T\Phi + \beta A)^{-1} \\ &= \beta(\Phi^T\Phi + \beta A)^{-1}\big(\sigma^2[2A + \beta A(\Phi^T\Phi)^{-1}A] - \beta A\theta_0\theta_0^T A\big) \\ &\quad \times (\Phi^T\Phi + \beta A)^{-1}.\end{aligned}$$

As a result, to prove $M(0) - M(\beta) > 0$, it suffices to show

$$\sigma^2[2A + \beta A(\Phi^T\Phi)^{-1}A] - \beta A\theta_0\theta_0^T A > 0 \tag{A.4}$$

which is true if $2\sigma^2 I_n - \beta A^{1/2}\theta_0\theta_0^T A^{1/2} > 0$ due to

$$\begin{aligned}&\sigma^2[2A + \beta A(\Phi^T\Phi)^{-1}A] - \beta A\theta_0\theta_0^T A \\ &\quad > 2\sigma^2 A - \beta A\theta_0\theta_0^T A \\ &\quad = A^{1/2}(2\sigma^2 I_n - \beta A^{1/2}\theta_0\theta_0^T A^{1/2})A^{1/2} > 0\end{aligned}$$

where $A^{1/2}$ is the symmetric and positive definite square root of A when A is positive definite. In addition, the eigenvalues of $A^{1/2}\theta_0\theta_0^T A^{1/2}$ are $\theta_0^T A\theta_0$ and zero (with multiplicity $n-1$). This shows $2\sigma^2 I_n - \beta A^{1/2}\theta_0\theta_0^T A^{1/2} > 0$ for $0 < \beta < 2\sigma^2/(\theta_0^T A\theta_0)$.

Note that $\text{MSEg}(\beta) = \text{Tr}(M(\beta))$. One has proved that $M(0) - M(\beta)$ is positive definite if $0 < \beta < 2\sigma^2/(\theta_0^T A\theta_0)$, so we have $\text{MSEg}(0) - \text{MSEg}(\beta) = \text{Tr}(M(0) - M(\beta)) > 0$.

The proof for the MSEy (11b) is similar to that for the MSEg (11a) by using the connection (10).

Remark A1. When $\beta \to \infty$, from the MSEg (A.1) we have

(1) $\text{Bias}(\beta) \to \theta_0^T\theta_0$ and $\frac{\text{dBias}(\beta)}{\text{d}\beta} \to 0$,
(2) $\text{Var}(\beta) \to 0$ and $\frac{\text{dVar}(\beta)}{\text{d}\beta} \to 0$,
(3) $\text{MSEg}(\beta) \to \theta_0^T\theta_0$ and $\frac{\text{dMSEg}(\beta)}{\text{d}\beta} \to 0$.

A.2. Proof of Proposition 3

To prove (24a), let us set

$$\mathscr{F}_{\text{Sg}_1}(P) = \sigma^4 Y^T Q^{-T}\Phi(\Phi^T\Phi)^{-2}\Phi^T Q^{-1}Y$$
$$\mathscr{F}_{\text{Sg}_2}(P) = \sigma^2\text{Tr}\big(2R^{-1} - (\Phi^T\Phi)^{-1}\big).$$

By (B.1) and (B.4), the derivative of $\mathscr{F}_{\text{Sg}_1}(P)$ is

$$\begin{aligned}\frac{\partial\mathscr{F}_{\text{Sg}_1}(P)}{\partial P} &= \sigma^4\sum_{i,j}\big(2\Phi(\Phi^T\Phi)^{-2}\Phi^T Q^{-1}YY^T\big)_{ij}\frac{\partial(Q^{-1})_{ij}}{\partial P} \\ &= -2\sigma^4\sum_{i,j}(\Phi(\Phi^T\Phi)^{-2}\Phi^T Q^{-1}YY^T)_{ij}\Phi^T Q^{-T}J_{ij}Q^{-T}\Phi \\ &= -2\sigma^4\Phi^T Q^{-T}\Phi(\Phi^T\Phi)^{-2}\Phi^T Q^{-1}YY^T Q^{-T}\Phi\end{aligned}\tag{A.5}$$

and using (B.16) implies the derivative of $\mathscr{F}_{\text{Sg}_2}(P)$

$$\begin{aligned}\frac{\partial\mathscr{F}_{\text{Sg}_2}(P)}{\partial P} &= 2\sigma^2\sum_{i=1}^{n}\frac{\partial(R^{-1})_{ii}}{\partial P} \\ &= 2\sigma^4 P^{-T}R^{-T}R^{-T}P^{-T} = 2\sigma^4 H^{-T}H^{-T}.\end{aligned}\tag{A.6}$$

Combining (A.5) with (A.6) derives (24a).

The proof of (24b) and (24c) is similar by using Lemma B1 and so it is omitted.

A.3. Proof of Proposition 4

Recall that $R = \Phi^T\Phi + \sigma^2 P^{-1}$ and V is the noise vector. It follows from (6b) that

$$\begin{aligned}\widehat{\theta}^{\text{R}} - \theta_0 &= R^{-1}\Phi^T Y - \theta_0 \\ &= -\sigma^2 R^{-1}P^{-1}\theta_0 + R^{-1}\Phi^T V \\ &= -\sigma^2 H^{-1}\theta_0 + R^{-1}\Phi^T V\end{aligned}$$

which derives

$$\begin{aligned}\text{MSEg}(P) &= \sigma^4\theta_0^T H^{-T}H^{-1}\theta_0 + \sigma^2\text{Tr}(R^{-1}\Phi^T\Phi R^{-T}) \\ &= \text{MSEg1}(P) + \text{MSEg2}(P).\end{aligned}$$

For the term MSEg1(P), using the formulae (B.1) and (B.4) gives

$$\begin{aligned}\frac{\partial \mathrm{MSEg1}(P)}{\partial P} &= \sigma^4 \sum_{i,j} \left(2H^{-1}\theta_0\theta_0^T\right)_{ij} \frac{\partial \left(H^{-1}\right)_{ij}}{\partial P} \\ &= \sigma^4 \sum_{i,j} \left(2H^{-1}\theta_0\theta_0^T\right)_{ij} \left(-H^{-T} J_{ij} H^{-T} \Phi^T \Phi\right) \\ &= -2\sigma^4 H^{-T} H^{-1} \theta_0 \theta_0^T H^{-T} \Phi^T \Phi \\ &= -2\sigma^4 H^{-T} H^{-1} \theta_0 \theta_0^T \Phi^T Q^{-T} \Phi. \end{aligned} \tag{A.7}$$

By using the formulae (B.6) and (B.16), one derives

$$\begin{aligned}\frac{\mathrm{MSEg2}(P)}{\partial P} &= \sigma^2 \sum_{i,j} \left(2R^{-1}\Phi^T\Phi\right)_{ij} \frac{\partial \left(R^{-1}\right)_{ij}}{\partial P} \\ &= \sigma^2 \sum_{i,j} \left(2R^{-1}\Phi^T\Phi\right)_{ij} (\sigma^2 P^{-T} R^{-T} J_{ij} R^{-T} P^{-T}) \\ &= 2\sigma^4 P^{-T} R^{-T} R^{-1} \Phi^T \Phi R^{-T} P^{-T} \\ &= 2\sigma^4 H^{-T} H^{-1} P \Phi^T Q^{-T} \Phi. \end{aligned} \tag{A.8}$$

Combining (A.7) with (A.8) implies the conclusion (26a).

The proof of (26b) and (26c) is similar by using Lemma B1 and so it is omitted.

A.4. Proof of Proposition 6

Let us first consider the EB estimator. Under the assumptions that $P(\eta) = \eta I_n$ and $\Phi^T\Phi = NI_n$, we have $S^{-1} = 1/(\eta + \sigma^2/N) I_n$. Further, by using (28f) and (23), we obtain

$$\frac{\mathrm{d}\mathscr{F}_{\mathrm{EB}}(P(\eta))}{\mathrm{d}\eta} = \frac{nN^2}{(\eta N + \sigma^2)^2}\left(\eta + \frac{\sigma^2}{N} - \frac{(\widehat{\theta}^{\mathrm{LS}})^T \widehat{\theta}^{\mathrm{LS}}}{n}\right).$$

By using the Lagrange multiplier, the resulting Lagrangian is

$$\mathscr{L}_{\mathrm{EB}}(\eta, \lambda) = \mathscr{F}_{\mathrm{EB}}(P(\eta)) - \lambda\eta$$

where λ is the Lagrange multiplier. Thus the corresponding Karush–Kuhn–Tucker (KKT) conditions (Boyd & Vandenberghe, 2004) are

$$\begin{aligned}&\frac{nN^2}{(\eta N + \sigma^2)^2}\left(\eta + \frac{\sigma^2}{N} - \frac{(\widehat{\theta}^{\mathrm{LS}})^T \widehat{\theta}^{\mathrm{LS}}}{n}\right) - \lambda = 0 \\ &\lambda\eta = 0, \quad \lambda \geq 0, \quad \eta \geq 0\end{aligned}$$

and its solution is

$$\begin{cases} \widehat{\eta} = \dfrac{(\widehat{\theta}^{\mathrm{LS}})^T \widehat{\theta}^{\mathrm{LS}}}{n} - \dfrac{\sigma^2}{N}, \widehat{\lambda} = 0 & \text{if } \dfrac{(\widehat{\theta}^{\mathrm{LS}})^T \widehat{\theta}^{\mathrm{LS}}}{n} \geq \dfrac{\sigma^2}{N} \\ \widehat{\eta} = 0, \widehat{\lambda} = \dfrac{nN^2}{\sigma^4}\left(\dfrac{\sigma^2}{N} - \dfrac{(\widehat{\theta}^{\mathrm{LS}})^T \widehat{\theta}^{\mathrm{LS}}}{n}\right) & \text{if } \dfrac{(\widehat{\theta}^{\mathrm{LS}})^T \widehat{\theta}^{\mathrm{LS}}}{n} < \dfrac{\sigma^2}{N}. \end{cases}$$

Therefore, we obtain the hyperparameter η estimated by the EB estimator is

$$\max\left(0, \frac{(\widehat{\theta}^{\mathrm{LS}})^T \widehat{\theta}^{\mathrm{LS}}}{n} - \frac{\sigma^2}{N}\right). \tag{A.9}$$

Under the same setting $\Phi^T\Phi = NI_n$, the KKT conditions corresponding to the SUREg and SUREy estimators derive that the arguments minimizing $\mathscr{F}_{\mathrm{Sg}}(P(\eta))$ and $\mathscr{F}_{\mathrm{Sy}}(P(\eta))$ under the constraint $\eta \geq 0$ are also (A.9), respectively.

Likewise, by using (28a), (28c), and (28e) we have all the hyperparameters minimizing the MSEg($P(\eta)$), MSEy($P(\eta)$), and EEB($P(\eta)$) satisfy the equation

$$\mathrm{Tr}\left(\eta I_n - \theta_0\theta_0^T\right) = 0$$

and its solution is $\theta_0^T\theta_0/n$.

A.5. Proof of Proposition 9

Under the assumptions that $\Phi^T\Phi/N \to \Sigma > 0$ and the white noise $v(t)$, we have $(\Phi^T\Phi)^{-1} = O_p(1/N) \to 0$, $S^{-1} \to P^{-1}$, $NR^{-1} \to \Sigma^{-1}$, $R^{-1}\Phi^T\Phi \to I_n$, and $\widehat{\theta}^{\mathrm{LS}} \to \theta_0$ almost surely as $N \to \infty$.

Let us first prove (36a). Using (27), we rewrite MSEg(P) in (11a) as follows:

$$\begin{aligned}\mathrm{MSEg}(P) = {} & \sigma^4 \theta_0^T S^{-1} (\Phi^T\Phi)^{-2} S^{-1} \theta_0 \\ & + \sigma^2 \mathrm{Tr}(R^{-1}\Phi^T\Phi R^{-1}).\end{aligned}$$

The limit

$$\begin{aligned}&N^2\left(R^{-1}\Phi^T\Phi R^{-1} - (\Phi^T\Phi)^{-1}\right) \\ =& -\sigma^2 N^2 R^{-1}\left(2P^{-1} + \sigma^2 P^{-1}(\Phi^T\Phi)^{-1}P^{-1}\right)R^{-1} \\ \to& -2\sigma^2 \Sigma^{-1} P^{-1} \Sigma^{-1}\end{aligned} \tag{A.10}$$

yields that

$$\begin{aligned}&N^2(\mathrm{MSEg}(P) - \sigma^2 \mathrm{Tr}((\Phi^T\Phi)^{-1})) \\ =& \sigma^4 \theta_0^T S^{-1}(N^2(\Phi^T\Phi)^{-2}) S^{-1}\theta_0 \\ &+ \sigma^2 N^2 \mathrm{Tr}\left(R^{-1}\Phi^T\Phi R^{-1} - (\Phi^T\Phi)^{-1}\right) \\ \to& \sigma^4 \theta_0^T P^{-1} \Sigma^{-2} P^{-1} \theta_0 - 2\sigma^4 \mathrm{Tr}\left(\Sigma^{-1} P^{-1} \Sigma^{-1}\right) \\ =& W_g(P, \Sigma, \theta_0).\end{aligned} \tag{A.11}$$

To prove (36b), note that the first term of $\mathscr{F}_{\mathrm{Sg}}(P)$ can be rewritten as $\sigma^4 (\widehat{\theta}^{\mathrm{LS}})^T S^{-1} (\Phi^T\Phi)^{-2} S^{-1} \widehat{\theta}^{\mathrm{LS}}$. Thus one derives

$$\begin{aligned}&N^2(\mathscr{F}_{\mathrm{Sg}}(P) - \sigma^2 \mathrm{Tr}((\Phi^T\Phi)^{-1})) \\ &\quad = \sigma^4 (\widehat{\theta}^{\mathrm{LS}})^T S^{-1} N^2 (\Phi^T\Phi)^{-2} S^{-1} \widehat{\theta}^{\mathrm{LS}} \\ &\qquad + 2\sigma^2 N^2 \mathrm{Tr}\left(R^{-1} - (\Phi^T\Phi)^{-1}\right) \\ &\quad \to W_g(P, \Sigma, \theta_0)\end{aligned} \tag{A.12}$$

where we use the limit

$$\begin{aligned}N^2(R^{-1} - (\Phi^T\Phi)^{-1}) &= -\sigma^2 NR^{-1} P^{-1} N (\Phi^T\Phi)^{-1} \\ &\to -\sigma^2 \Sigma^{-1} P^{-1} \Sigma^{-1}.\end{aligned}$$

Similarly, we can rewrite MSEy(P) as

$$\begin{aligned}\mathrm{MSEy}(P) = {} & \sigma^4 \theta_0^T S^{-1} (\Phi^T\Phi)^{-1} S^{-1} \theta_0 + N\sigma^2 \\ & + \mathrm{Tr}\left(R^{-1}\Phi^T\Phi R^{-1}\Phi^T\Phi\right)\end{aligned} \tag{A.13}$$

and hence the assertion (36c) is proved by

$$\begin{aligned}&N(\mathrm{MSEy}(P) - (n+N)\sigma^2) \\ &\quad = \sigma^4 \theta_0^T S^{-1} N (\Phi^T\Phi)^{-1} S^{-1} \theta_0 \\ &\qquad + \sigma^2 N \mathrm{Tr}\left(R^{-1}\Phi^T\Phi R^{-1}\Phi^T\Phi - I_n\right) \\ &\quad \to W_y(P, \Sigma, \theta_0)\end{aligned} \tag{A.14}$$

where we use the formulae

$$\begin{aligned}&N(R^{-1}\Phi^T\Phi R^{-1}\Phi^T\Phi - I_n) \\ =& -\sigma^2 NR^{-1}\left[2P^{-1} + \sigma^2 P^{-1}(\Phi^T\Phi)^{-1}P^{-1}\right]R^{-1}\Phi^T\Phi \\ \to& -2\sigma^2 \Sigma^{-1} P^{-1}.\end{aligned}$$

To prove (36d), we need some identities. A straightforward calculation shows that

$$Q(I_N - \Phi(\Phi^T\Phi)^{-1}\Phi^T)Q = \sigma^4(I_N - \Phi(\Phi^T\Phi)^{-1}\Phi^T).$$

This means that

$$\sigma^4 Q^{-1}(I_N - \Phi(\Phi^T\Phi)^{-1}\Phi^T)Q^{-1} = I_N - \Phi(\Phi^T\Phi)^{-1}\Phi^T$$

and hence we derive

$$\sigma^4 Y^T Q^{-2} Y + Y^T \Phi(\Phi^T\Phi)^{-1}\Phi^T Y - Y^T Y = \sigma^4 Y^T Q^{-1}\Phi(\Phi^T\Phi)^{-1}\Phi^T Q^{-1} Y.$$

It follows from (B.11) and (B.14) that

$$\begin{aligned} &N\big[\mathscr{F}_{\mathrm{Sy}}(P) + Y^T\Phi(\Phi^T\Phi)^{-1}\Phi^T Y - Y^T Y - 2n\sigma^2\big] \\ &= N\big[\sigma^4 Y^T Q^{-1}\Phi(\Phi^T\Phi)^{-1}\Phi^T Q^{-1} Y + 2\sigma^2 \mathrm{Tr}(R^{-1}\Phi^T\Phi - I_n)\big] \\ &= N\big[\sigma^4(\widehat{\theta}^{\mathrm{LS}})^T S^{-1}(\Phi^T\Phi)^{-1} S^{-1}\widehat{\theta}^{\mathrm{LS}} + 2\sigma^2\mathrm{Tr}(R^{-1}\Phi^T\Phi - I_n)\big] \\ &\to W_{\mathrm{y}}(P, \Sigma, \theta_0) \end{aligned} \tag{A.15}$$

where we use the limit

$$N(R^{-1}\Phi^T\Phi - I_n) = -\sigma^2 N R^{-1}P^{-1} \to -\sigma^2\Sigma^{-1}P^{-1}.$$

Similarly, we need two identities to prove (36e). Using the Sylvester's determinant identity $\det(I_n + AB) = \det(I_N + BA)$ derives

$$\det(Q) = \sigma^{2(N-n)}\det(\Phi^T\Phi)\det(P + \sigma^2(\Phi^T\Phi)^{-1})$$

which implies

$$\begin{aligned} &\log\det(Q) - (N-n)\log\sigma^2 - \log\det(\Phi^T\Phi) \\ &\qquad = \log\det(S) \to \log\det(P). \end{aligned} \tag{A.16}$$

Starting with the identity $I_N = \sigma^2 Q^{-1} + \Phi P\Phi^T Q^{-1}$ gives

$$\begin{aligned} \sigma^2\mathrm{Tr}(Q^{-1}) &= N - \mathrm{Tr}(\Phi P\Phi^T Q^{-1}) \\ &= N - \mathrm{Tr}(R^{-1}\Phi^T\Phi) \to N - n. \end{aligned}$$

Therefore, the limit (36e) is proved by

$$\begin{aligned} &\mathrm{EEB}(P) - (N-n) - (N-n)\log\sigma^2 - \log\det(\Phi^T\Phi) \\ &= \theta_0^T S^{-1}\theta_0 + \big(\sigma^2\mathrm{Tr}(Q^{-1}) - (N-n)\big) \\ &\quad + \log\det(Q) - (N-n)\log\sigma^2 - \log\det(\Phi^T\Phi) \\ &\to \theta_0^T P^{-1}\theta_0 + \log\det(P) = W_{\mathrm{B}}(P, \theta_0). \end{aligned} \tag{A.17}$$

At last, we finish the proof by checking (36f). The identity

$$Q(I_N - \Phi(\Phi^T\Phi)^{-1}\Phi^T)/\sigma^2 = I_N - \Phi(\Phi^T\Phi)^{-1}\Phi^T$$

implies that

$$\begin{aligned} &Y^T Q^{-1} Y + Y^T\Phi(\Phi^T\Phi)^{-1}\Phi^T Y/\sigma^2 - Y^T Y/\sigma^2 \\ &\quad = Y^T Q^{-1}\Phi(\Phi^T\Phi)^{-1}\Phi^T Y. \end{aligned} \tag{A.18}$$

It follows from (A.16), (A.18), and (B.11) that

$$\begin{aligned} &\mathscr{F}_{\mathrm{EB}}(P) + Y^T\Phi(\Phi^T\Phi)^{-1}\Phi^T Y/\sigma^2 - Y^T Y/\sigma^2 \\ &\quad - (N-n)\log\sigma^2 - \log\det(\Phi^T\Phi) \\ &= Y^T Q^{-1} Y + Y^T\Phi(\Phi^T\Phi)^{-1}\Phi^T Y/\sigma^2 - Y^T Y/\sigma^2 \\ &\quad + \log\det(Q) - (N-n)\log\sigma^2 - \log\det(\Phi^T\Phi) \\ &= Y^T Q^{-1}\Phi(\Phi^T\Phi)^{-1}\Phi^T Y + \log\det(S) \\ &\to W_{\mathrm{B}}(P, \theta_0). \end{aligned} \tag{A.19}$$

A.6. Proof of Theorem 1

Firstly, we prove $\widehat{\eta}_{\mathrm{MSEg}} \to \eta_g^*$ as $N \to \infty$. Define

$$\overline{\mathrm{MSE}}\mathrm{g}(P) \triangleq N^2\big(\mathrm{MSEg}(P) - \sigma^2\mathrm{Tr}((\Phi^T\Phi)^{-1})\big). \tag{A.20}$$

Clearly, we have $\widehat{\eta}_{\mathrm{MSEg}}$ also minimizes $\overline{\mathrm{MSE}}\mathrm{g}(P(\eta))$, i.e.,

$$\widehat{\eta}_{\mathrm{MSEg}} = \underset{\eta\in\Omega}{\mathrm{argmin}}\ \overline{\mathrm{MSE}}\mathrm{g}(P(\eta)).$$

Under Assumption 2, there exists a compact set

$$\overline{\Omega} \subset \Omega \tag{A.21}$$

containing η_g^* such that $0 < d_1 \le \|P(\eta)\| \le d_2 < \infty$ for all $\eta \in \overline{\Omega}$. Then by Lemma B3 in Appendix B, to prove $\widehat{\eta}_{\mathrm{MSEg}} \to \eta_g^*$ as $N \to \infty$, it suffices to show that $\overline{\mathrm{MSE}}\mathrm{g}(P(\eta))$ converges to $W_g(P(\eta), \Sigma, \theta_0)$ almost surely and uniformly in $\overline{\Omega}$, as $N \to \infty$.

It follows from (A.11) and (A.10) that

$$\begin{aligned} &\overline{\mathrm{MSE}}\mathrm{g}(P(\eta)) - W_g(P, \Sigma, \theta_0) \\ &\quad = \sigma^4 Z_1 + 2\sigma^4\mathrm{Tr}(Z_2) - \sigma^6\mathrm{Tr}(Z_3) \end{aligned} \tag{A.22}$$

$$\begin{aligned} Z_1 &= \theta_0^T S^{-1}(N^2(\Phi^T\Phi)^{-2})S^{-1}\theta_0 - \theta_0^T P^{-1}\Sigma^{-2}P^{-1}\theta_0 \\ Z_2 &= \Sigma^{-1}P^{-1}\Sigma^{-1} - N^2 R^{-1}P^{-1}R^{-1} \\ Z_3 &= -N^2 R^{-1}P^{-1}(\Phi^T\Phi)^{-1}P^{-1}R^{-1}. \end{aligned} \tag{A.23}$$

For the term Z_1, we have

$$\begin{aligned} Z_1 = {}& \theta_0^T(S^{-1} - P^{-1})(N^2(\Phi^T\Phi)^{-2})S^{-1}\theta_0 \\ &+ \theta_0^T P^{-1}(N^2(\Phi^T\Phi)^{-2} - \Sigma^{-2})S^{-1}\theta_0 \\ &+ \theta_0^T P^{-1}\Sigma^{-2}(S^{-1} - P^{-1})\theta_0 \end{aligned} \tag{A.24}$$

where

$$S^{-1} - P^{-1} = -\sigma^2 S^{-1}(\Phi^T\Phi)^{-1}P^{-1}. \tag{A.25}$$

Note that $\Phi^T\Phi/N \to \Sigma$ implies that $\|N(\Phi^T\Phi)^{-1}\| = O_p(1)$. Then further noting that $d_1 \le \|P(\eta)\| \le d_2$ and $\|S(\eta)^{-1}\| < \|(P(\eta))^{-1}\| \le 1/d_1$ for $\eta \in \overline{\Omega}$, we have Z_1 converges to zero almost surely and uniformly in $\overline{\Omega}$.

For the term Z_2, we have

$$\begin{aligned} &\Sigma^{-1}P^{-1}\Sigma^{-1} - N^2 R^{-1}P^{-1}R^{-1} \\ &= (\Sigma^{-1} - NR^{-1})P^{-1}\Sigma^{-1} + NR^{-1}P^{-1}(\Sigma^{-1} - NR^{-1}). \end{aligned}$$

The assertions $NR^{-1} \to \Sigma^{-1}$ and $\|NR^{-1} - \Sigma^{-1}\| = O_p(1)$ yield that Z_2 converges to zero almost surely and uniformly in $\overline{\Omega}$. Finally, by noting $(\Phi^T\Phi)^{-1} \to 0$ as $N \to \infty$ it is easy to see that Z_3 also converges to zero almost surely and uniformly. Making use of these facts shows that $\overline{\mathrm{MSE}}\mathrm{g}(P(\eta))$ converges to $W_g(P(\eta), \Sigma, \theta_0)$ almost surely and uniformly in $\overline{\Omega}$ and hence, by Lemma B3, $\widehat{\eta}_{\mathrm{MSEg}} \to \eta_g^*$ as $N \to \infty$ almost surely.

Secondly, we prove that $\widehat{\eta}_{\mathrm{Sg}} \to \eta_g^*$ as $N \to \infty$ and the proof is similar to that of $\widehat{\eta}_{\mathrm{MSEg}} \to \eta_g^*$ as $N \to \infty$. Define

$$\overline{\mathscr{F}}_{\mathrm{Sg}}(P(\eta)) \triangleq N^2\big(\mathscr{F}_{\mathrm{Sg}}(P(\eta)) - \sigma^2\mathrm{Tr}((\Phi^T\Phi)^{-1})\big).$$

Then, we have

$$\widehat{\eta}_{\mathrm{Sg}} = \underset{\eta\in\Omega}{\mathrm{argmin}}\ \overline{\mathscr{F}}_{\mathrm{Sg}}(P(\eta)). \tag{A.26}$$

It follows from (A.12) that

$$\begin{aligned} &\overline{\mathscr{F}}_{\mathrm{Sg}}(P(\eta)) - W_g(P, \Sigma, \theta_0) = \sigma^4 Z_1' + 2\sigma^4\mathrm{Tr}(Z_2') \\ &Z_1' = (\widehat{\theta}^{\mathrm{LS}})^T S^{-1} N^2(\Phi^T\Phi)^{-2} S^{-1}\widehat{\theta}^{\mathrm{LS}} - \theta_0^T P^{-1}\Sigma^{-2}P^{-1}\theta_0 \\ &Z_2' = \Sigma^{-1}P^{-1}\Sigma^{-1} - NR^{-1}P^{-1}N(\Phi^T\Phi)^{-1}. \end{aligned}$$

For the terms Z_1' and Z_2', we have

$$\begin{aligned} Z_1' = {}& \big(\widehat{\theta}^{\mathrm{LS}} - \theta_0\big)^T S^{-1} N^2(\Phi^T\Phi)^{-2} S^{-1}\widehat{\theta}^{\mathrm{LS}} \\ &+ \theta_0^T\big(S^{-1} - P^{-1}\big)N^2(\Phi^T\Phi)^{-2}S^{-1}\widehat{\theta}^{\mathrm{LS}} \\ &+ \theta_0^T P^{-1}\big(N^2(\Phi^T\Phi)^{-2} - \Sigma^{-2}\big)S^{-1}\widehat{\theta}^{\mathrm{LS}} \\ &+ \theta_0^T P^{-1}\Sigma^{-2}\big(S^{-1} - P^{-1}\big)\widehat{\theta}^{\mathrm{LS}} \\ &+ \theta_0^T P^{-1}\Sigma^{-2}P^{-1}\big(\widehat{\theta}^{\mathrm{LS}} - \theta_0\big) \end{aligned} \tag{A.27}$$

$$\begin{aligned} Z_2' = {}& \big(\Sigma^{-1} - NR^{-1}\big)P^{-1}\Sigma^{-1} \\ &+ NR^{-1}P^{-1}\big(\Sigma^{-1} - N(\Phi^T\Phi)^{-1}\big). \end{aligned} \tag{A.28}$$

Then, noting that $\widehat{\theta}^{\mathrm{LS}} \to \theta_0$, $S^{-1} \to P^{-1}$, $N(\Phi^T\Phi)^{-1} \to \Sigma^{-1}$, $NR^{-1} \to \Sigma^{-1}$ almost surely as $N \to \infty$, and $\|NR^{-1}\| = O_p(1)$, $\|\widehat{\theta}^{\mathrm{LS}}\| = O_p(1)$, and $d_1 \le \|P(\eta)\| \le d_2$, $\|S(\eta)^{-1}\| < \|(P(\eta))^{-1}\| \le 1/d_1$, for $\eta \in \overline{\Omega}$, one can show that each term of (A.27) and (A.28), and thus both Z_1' and Z_2' converge to zero almost surely and uniformly in $\overline{\Omega}$. Therefore, $\overline{\mathscr{F}}_{\mathrm{Sg}}(P(\eta))$ converges to $W_g(P, \Sigma, \theta_0)$ almost surely and uniformly in $\overline{\Omega}$. It then follows from Lemma B3 that $\widehat{\eta}_{\mathrm{Sg}} \to \eta_{\mathrm{g}}^*$ almost surely as $N \to \infty$.

The proof of (43b) and (43c) can be done similarly and thus is omitted. The first order optimality conditions of η_{g}^*, η_{y}^*, and η_{B}^* can be derived in a similar way as Proposition 4 and thus is omitted. This completes the proof.

A.7. Proof of Theorem 2

We first prove that $\|\widehat{\eta}_{\mathrm{MSEg}} - \eta_{\mathrm{g}}^*\| = O_p(\varpi_N)$.

Noting (A.11), the ith elements of the gradient vectors of $\overline{\mathrm{MSE}}\mathrm{g}(P(\eta))$ and $W_g(P(\eta), \Sigma, \theta_0)$ with respect to η are, respectively, for $1 \le i \le p$,

$$\begin{aligned}\frac{\partial \overline{\mathrm{MSE}}\mathrm{g}(P(\eta))}{\partial \eta_i} &= 2\sigma^4 N^2 \theta_0^T S^{-1}(\Phi^T\Phi)^{-2}\frac{\partial S^{-1}}{\partial \eta_i}\theta_0 \\ &\quad + 2\sigma^2 N^2 \mathrm{Tr}\Big(\frac{\partial R^{-1}}{\partial \eta_i}\Phi^T\Phi R^{-1}\Big) \\ \frac{\partial W_g(P(\eta), \Sigma, \theta_0)}{\partial \eta_i} &= 2\sigma^4 \theta_0^T P^{-1}\Sigma^{-2}\frac{\partial P^{-1}}{\partial \eta_i}\theta_0 \\ &\quad - 2\sigma^4 \mathrm{Tr}\Big(\Sigma^{-1}\frac{\partial P^{-1}}{\partial \eta_i}\Sigma^{-1}\Big). \end{aligned} \tag{A.29}$$

Using the identity $\frac{\partial R^{-1}}{\partial \eta_i} = -R^{-1}\frac{\partial R}{\partial \eta_i}R^{-1} = -\sigma^2 R^{-1}\frac{\partial P^{-1}}{\partial \eta_i}R^{-1}$, we see their difference is

$$\frac{\partial \overline{\mathrm{MSE}}\mathrm{g}(P(\eta))}{\partial \eta_i} - \frac{\partial W_g(P(\eta), \Sigma, \theta_0)}{\partial \eta_i} = 2\sigma^4\big(\Upsilon_1 + \mathrm{Tr}(\Upsilon_2)\big)$$

where $\Upsilon_1 = \theta_0^T S^{-1}\big(N^2(\Phi^T\Phi)^{-2}\big)\frac{\partial S^{-1}}{\partial \eta_i}\theta_0 - \theta_0^T P^{-1}\Sigma^{-2}\frac{\partial P^{-1}}{\partial \eta_i}\theta_0$,

$$\Upsilon_2 = \Sigma^{-1}\frac{\partial P^{-1}}{\partial \eta_i}\Sigma^{-1} - NR^{-1}\frac{\partial P^{-1}}{\partial \eta_i}R^{-1}\Phi^T\Phi NR^{-1}.$$

Noting $\|N(\Phi^T\Phi)^{-1} - \Sigma^{-1}\| = O_p(\delta_N)$, $\|S^{-1} - P^{-1}\| = O_p(1/N)$, $\big\|\frac{\partial S^{-1}}{\partial \eta_i} - \frac{\partial P^{-1}}{\partial \eta_i}\big\| = O_p(1/N)$, $\big\|R^{-1}\Phi^T\Phi - I_n\big\| = O_p(1/N)$, $\big\|NR^{-1} - \Sigma^{-1}\big\| = O_p(\delta_N)$, and $d_1 \le \|P(\eta)\| \le d_2$ and $\|S(\eta)^{-1}\| < \|(P(\eta))^{-1}\| \le 1/d_1$ for $\eta \in \overline{\Omega}$ yields

$$|\Upsilon_1| = O_p(\varpi_N), \quad |\mathrm{Tr}(\Upsilon_2)| = O_p(\varpi_N) \tag{A.30}$$

uniformly in $\overline{\Omega}$, where $\overline{\Omega}$ is defined in (A.21). Therefore,

$$\Big\|\frac{\partial \overline{\mathrm{MSE}}\mathrm{g}(P(\eta))}{\partial \eta} - \frac{\partial W_g(P(\eta), \Sigma, \theta_0)}{\partial \eta}\Big\| = O_p(\varpi_N)$$

uniformly for any $\eta \in \overline{\Omega}$. Since $\widehat{\eta}_{\mathrm{MSEg}}$ and η_{g}^* minimize $\overline{\mathrm{MSE}}\mathrm{g}(P)$ and $W_g(P, \Sigma, \theta_0)$, respectively, we have

$$\frac{\partial \overline{\mathrm{MSE}}\mathrm{g}(P(\eta))}{\partial \eta}\Big|_{\eta=\widehat{\eta}_{\mathrm{MSEg}}} = 0 \text{ and } \frac{\partial W_g(P(\eta), \Sigma, \theta_0)}{\partial \eta}\Big|_{\eta=\eta_{\mathrm{g}}^*} = 0.$$

It follows that

$$\frac{\partial \overline{\mathrm{MSE}}\mathrm{g}(P(\eta))}{\partial \eta}\Big|_{\eta=\eta_{\mathrm{g}}^*} = O_p(\varpi_N).$$

In addition, by using (A.29), the (i, j)-element of the Hessian matrix of $W_g(P(\eta), \Sigma, \theta_0)$ is

$$\begin{aligned}&\frac{\partial^2 W_g(P(\eta), \Sigma, \theta_0)}{\partial \eta_i \partial \eta_j} \\ &= 2\sigma^4\theta_0^T P^{-1}\Sigma^{-2}\frac{\partial^2 P^{-1}}{\partial \eta_i \partial \eta_j}\theta_0 + 2\sigma^4\theta_0^T\frac{\partial P^{-1}}{\partial \eta_j}\Sigma^{-2}\frac{\partial P^{-1}}{\partial \eta_i}\theta_0 \\ &\quad - 2\sigma^4\mathrm{Tr}\Big(\Sigma^{-1}\frac{\partial^2 P^{-1}}{\partial \eta_i \partial \eta_j}\Sigma^{-1}\Big). \end{aligned} \tag{A.31}$$

The Hessian matrix $\frac{\partial^2 \overline{\mathrm{MSE}}\mathrm{g}(P(\eta))}{\partial \eta \partial \eta^T}$ of $\overline{\mathrm{MSE}}\mathrm{g}(P(\eta))$ is omitted here for simplicity. Then, it can be shown that

$$\Big\|\frac{\partial^2 \overline{\mathrm{MSE}}\mathrm{g}(P(\eta))}{\partial \eta \partial \eta^T} - \frac{\partial^2 W_g(P(\eta), \Sigma, \theta_0)}{\partial \eta \partial \eta^T}\Big\| = o_p(1)$$

uniformly for any $\eta \in \overline{\Omega}$. Applying the Taylor expansion to $\frac{\partial \overline{\mathrm{MSE}}\mathrm{g}(P(\eta))}{\partial \eta}$ yields

$$\begin{aligned}0 &= \frac{\partial \overline{\mathrm{MSE}}\mathrm{g}(P(\eta))}{\partial \eta}\Big|_{\eta=\widehat{\eta}_{\mathrm{MSEg}}} = \frac{\partial \overline{\mathrm{MSE}}\mathrm{g}(P(\eta))}{\partial \eta}\Big|_{\eta=\eta_{\mathrm{g}}^*} \\ &\quad + \frac{\partial^2 \overline{\mathrm{MSE}}\mathrm{g}(P(\eta))}{\partial \eta \partial \eta^T}\Big|_{\eta=\bar{\eta}}(\widehat{\eta}_{\mathrm{MSEg}} - \eta_{\mathrm{g}}^*)\end{aligned}$$

where $\bar{\eta}$ lies between $\widehat{\eta}_{\mathrm{MSEg}}$ and η_{g}^*.

Clearly,

$$\frac{\partial^2 W_g(P(\eta), \Sigma, \theta_0)}{\partial \eta \partial \eta^T}\Big|_{\eta=\eta_{\mathrm{g}}^*} = O_p(1).$$

Then under Assumption 2, we have $\frac{\partial^2 W_g(P(\eta), \Sigma, \theta_0)}{\partial \eta \partial \eta^T}\Big|_{\eta=\eta_{\mathrm{g}}^*}$ is positive definite. For sufficiently large N, $\bar{\eta}$ would be close to η_{g}^*. In this case, we also have $\frac{\partial^2 W_g(P(\eta), \Sigma, \theta_0)}{\partial \eta \partial \eta^T}\Big|_{\eta=\bar{\eta}}$ is positive definite. Then it follows that

$$\begin{aligned}&\widehat{\eta}_{\mathrm{MSEg}} - \eta_{\mathrm{g}}^* \\ &= -\Big(\frac{\partial^2 \overline{\mathrm{MSE}}\mathrm{g}(P(\eta))}{\partial \eta \partial \eta^T}\Big|_{\eta=\bar{\eta}}\Big)^{-1}\frac{\partial \overline{\mathrm{MSE}}\mathrm{g}(P(\eta))}{\partial \eta}\Big|_{\eta=\eta_{\mathrm{g}}^*} \\ &= O_p(1)O_p(\varpi_N) = O_p(\varpi_N).\end{aligned}$$

Now, we prove $\|\widehat{\eta}_{\mathrm{Sg}} - \eta_{\mathrm{g}}^*\| = O_p(\mu_N)$ and the proof is similar to that of $\|\widehat{\eta}_{\mathrm{MSEg}} - \eta_{\mathrm{g}}^*\| = O_p(\varpi_N)$. By (A.12), the ith element of gradient vector of $\overline{\mathscr{F}}_{\mathrm{Sg}}(P(\eta))$ is

$$\begin{aligned}\frac{\partial \overline{\mathscr{F}}_{\mathrm{Sg}}(P(\eta))}{\partial \eta_i} &= 2\sigma^4(\widehat{\theta}^{\mathrm{LS}})^T S^{-1}N^2(\Phi^T\Phi)^{-2}\frac{\partial S^{-1}}{\partial \eta_i}\widehat{\theta}^{\mathrm{LS}} \\ &\quad + 2\sigma^2 N^2\mathrm{Tr}\Big(\frac{\partial R^{-1}}{\partial \eta_i}\Big).\end{aligned} \tag{A.32}$$

Using the identity $\frac{\partial R^{-1}}{\partial \eta_i} = -\sigma^2 R^{-1}\frac{\partial P^{-1}}{\partial \eta_i}R^{-1}$, we see

$$\frac{\partial \overline{\mathscr{F}}_{\mathrm{Sg}}(P(\eta))}{\partial \eta_i} - \frac{\partial W_g(P(\eta), \Sigma, \theta_0)}{\partial \eta_i} = 2\sigma^4\Upsilon_1' + 2\sigma^4\mathrm{Tr}\big(\Upsilon_2'\big)$$

where $\Upsilon_1' = (\widehat{\theta}^{\mathrm{LS}})^T S^{-1}N^2(\Phi^T\Phi)^{-2}\frac{\partial S^{-1}}{\partial \eta_i}\widehat{\theta}^{\mathrm{LS}} - \theta_0^T P^{-1}\Sigma^{-2}\frac{\partial P^{-1}}{\partial \eta_i}\theta_0$

$$\Upsilon_2' = \Sigma^{-1}\frac{\partial P^{-1}}{\partial \eta_i}\Sigma^{-1} - NR^{-1}\frac{\partial P^{-1}}{\partial \eta_i}NR^{-1}. \tag{A.33}$$

Since $\Phi^T\Phi/N \to \Sigma$ and $v(t)$ is a white noise, we have $\|\widehat{\theta}^{\rm LS}-\theta_0\| = O_p(1/\sqrt{N})$. Then noting that $\|N(\Phi^T\Phi)^{-1}-\Sigma^{-1}\| = O_p(\delta_N)$, $\|S^{-1}-P^{-1}\| = O_p(1/N)$, $\left\|\frac{\partial S^{-1}}{\partial \eta_i}-\frac{\partial P^{-1}}{\partial \eta_i}\right\| = O_p(1/N)$, $\left\|NR^{-1}-\Sigma^{-1}\right\| = O_p(\delta_N)$, and $\|NR^{-1}\| = O_p(1)$, $\|\widehat{\theta}^{\rm LS}\| = O_p(1)$, and $d_1 \le \|P(\eta)\| \le d_2$ and $\|S(\eta)^{-1}\| < \|(P(\eta))^{-1}\| \le 1/d_1$ for $\eta \in \overline{\Omega}$, yields

$$|\Upsilon_1'| = \max\big(O_p(1/\sqrt{N}), O_p(1/N), O_p(\delta_N)\big) = O_p(\mu_N)$$
$$|\mathrm{Tr}\big(\Upsilon_2'\big)| = O_p(\delta_N)$$

uniformly in $\overline{\Omega}$. It follows that

$$\left\|\frac{\partial \overline{\mathscr{F}}_{\rm Sg}(P(\eta))}{\partial \eta} - \frac{\partial W_g(P(\eta),\,\Sigma,\,\theta_0)}{\partial \eta}\right\| = O_p(\mu_N)$$

uniformly for any $\eta \in \overline{\Omega}$. This implies

$$\frac{\partial \overline{\mathscr{F}}_{\rm Sg}(P(\eta))}{\partial \eta}\Big|_{\eta=\eta_{\rm g}^*} = O_p(\mu_N). \tag{A.34}$$

Similarly, one can obtain the Hessian matrix of $\overline{\mathscr{F}}_{\rm Sg}(P(\eta))$ and can show that

$$\left\|\frac{\partial^2 \overline{\mathscr{F}}_{\rm Sg}(P(\eta))}{\partial \eta \partial \eta^T} - \frac{\partial^2 W_g(P(\eta),\,\Sigma,\,\theta_0)}{\partial \eta \partial \eta^T}\right\| = o_p(1) \tag{A.35}$$

uniformly for any $\eta \in \overline{\Omega}$. Applying the Taylor expansion of $\frac{\partial \overline{\mathscr{F}}_{\rm Sg}(P(\eta))}{\partial \eta}$ shows

$$\begin{aligned} 0 &= \frac{\partial \overline{\mathscr{F}}_{\rm Sg}(P(\eta))}{\partial \eta}\Big|_{\eta=\widehat{\eta}_{\rm Sg}} = \frac{\partial \overline{\mathscr{F}}_{\rm Sg}(P(\eta))}{\partial \eta}\Big|_{\eta=\eta_{\rm g}^*} \\ &\quad + \frac{\partial^2 \overline{\rm MSEg}(P(\eta))}{\partial \eta \partial \eta^T}\Big|_{\eta=\widetilde{\eta}}(\widehat{\eta}_{\rm MSEg} - \eta_{\rm g}^*) \end{aligned}$$

where $\widetilde{\eta}$ lies between $\widehat{\eta}_{\rm Sg}$ and $\eta_{\rm g}^*$. For sufficiently large N, we have

$$\begin{aligned} &\widehat{\eta}_{\rm Sg} - \eta_{\rm g}^* \\ &= -\Big(\frac{\partial^2 \overline{\mathscr{F}}_{\rm Sg}(P(\eta))}{\partial \eta \partial \eta^T}\Big|_{\eta=\widetilde{\eta}}\Big)^{-1} \frac{\partial \overline{\mathscr{F}}_{\rm Sg}(P(\eta))}{\partial \eta}\Big|_{\eta=\eta_{\rm g}^*} \\ &= O_p(1)O_p(\mu_N) = O_p(\mu_N). \end{aligned}$$

The proof of (44b) and (44c) can be done in a similar way and thus is omitted. This completes the proof.

Appendix B

This appendix contains the technical lemmas used in the proof in Appendix A.

B.1. Matrix differentials and related identities

This section introduces the differentiation of a function $f(X)$ where X is a matrix. It is assumed that X has no special structure, i.e., that the elements of X are independent. For convenience and readability, the formulae used in the paper are stated in the following lemmas.

Lemma B1 (*Petersen & Pedersen, 2012*)**.** *Assume that b is a column vector, and A, B and X are matrices with compatible dimensions. Then we have*

$$\frac{\partial b^T X^T A X b}{\partial X} = (A + A^T)Xbb^T \tag{B.1}$$
$$\frac{\partial b^T X^{-1} b}{\partial X} = -X^{-T}bb^TX^{-T} \tag{B.2}$$
$$\frac{\partial \log|\det(X)|}{\partial X} = X^{-T} \tag{B.3}$$
$$\frac{\partial (X^{-1})_{kl}}{\partial X_{ij}} = -(X^{-1})_{ki}(X^{-1})_{jl} \tag{B.4}$$
$$\frac{\partial \mathrm{Tr}(AX^{-1}B)}{\partial X} = -(X^{-1}BAX^{-1})^T \tag{B.5}$$
$$\frac{\partial \mathrm{Tr}(AXBX^TA^T)}{\partial X} = A^TAX(B+B^T) \tag{B.6}$$

where $(\cdot)_{ij}$ denotes the (i,j)th element of a matrix.

Lemma B2. *We have the following identities:*

$$\sum_{ij}(A)_{ij}J_{ij} = A \tag{B.7}$$
$$Y - \Phi\widehat{\theta}^{\rm R} = \sigma^2 Q^{-1}Y \tag{B.8}$$
$$\widehat{\theta}^{\rm LS} - \widehat{\theta}^{\rm R} = \sigma^2(\Phi^T\Phi)^{-1}\Phi^TQ^{-1}Y \tag{B.9}$$
$$A(I_N + BA)^{-1} = (I_n + AB)^{-1}A \tag{B.10}$$
$$\Phi^TQ^{-1}\Phi = S^{-1}\ \ \Phi^TQ^{-1}Y = S^{-1}\widehat{\theta}^{\rm LS} \tag{B.11}$$
$$\Phi^TQ^{-T}Q^{-T}\Phi = S^{-T}(\Phi^T\Phi)^{-1}S^{-T} \tag{B.12}$$
$$\Phi^TQ^{-T}Q^{-1}Y = S^{-T}(\Phi^T\Phi)^{-1}S^{-1}\widehat{\theta}^{\rm LS} \tag{B.13}$$
$$I_N - \sigma^2Q^{-1} = \Phi P\Phi^TQ^{-1} = Q^{-1}\Phi P\Phi^T = \Phi R^{-1}\Phi^T \tag{B.14}$$
$$\frac{\partial (Q^{-1})_{ij}}{\partial P} = -\Phi^TQ^{-T}J_{ij}Q^{-T}\Phi \tag{B.15}$$
$$\frac{\partial (R^{-1})_{ij}}{\partial P} = \sigma^2P^{-T}R^{-T}J_{ij}R^{-T}P^{-T} \tag{B.16}$$

where J_{ij} is the matrix whose (i,j)-element is one and zero for all other elements.

Proof. The identities (B.7)–(B.14) can be verified by a straightforward calculation. Using (B.4) gives

$$\begin{aligned} \frac{\partial (Q^{-1})_{ij}}{\partial P_{st}} &= \sum_{a,b}\frac{\partial (Q^{-1})_{ij}}{\partial Q_{ab}}\frac{\partial Q_{ab}}{\partial P_{st}} \\ &= -\sum_{a,b}(Q^{-1})_{ia}(Q^{-1})_{bj}\Phi_{as}(\Phi^T)_{tb} \\ &= -\sum_{a,b}(\Phi^T)_{sa}(Q^{-T})_{ai}(Q^{-T})_{jb}\Phi_{bt} \\ &= -(\Phi^TQ^{-T})_{si}(Q^{-T}\Phi)_{jt} \end{aligned}$$

which implies (B.15). While (B.16) can be proved in a similar way. □

B.2. Convergence result for extremum estimators

Lemma B3 (*Ljung, 1999, Theorem 8.2*)**.** *Assume that*

(1) *$M(\eta)$ is a deterministic function that is continuous in $\eta \in \Omega$ and minimized at the set*

$$\mathscr{D} = \operatorname*{argmin}_{\eta\in\Omega} M(\eta) = \big\{\eta | \eta \in \Omega, M(\eta) = \min_{\eta'\in\Omega} M(\eta')\big\}$$

where Ω is a compact subset of $\mathbb{R}^p$.

(2) *A sequence of functions $\{M_N(\eta)\}$ converges to $M(\eta)$ almost surely and uniformly in Ω as N goes to ∞.*

Then $\widehat{\eta}_N = \arg\min_{\eta\in\Omega} M_N(\eta)$ converges to $\mathscr{D}$ almost surely, namely,

$$\inf_{\eta^*\in\mathscr{D}} \|\widehat{\eta}_N - \eta^*\| \to 0, \quad \text{as } N \to \infty.$$

References

Aravkin, A., Burke, J. V., Chiuso, A., & Pillonetto, G. (2012a). On the estimation of hyperparameters for empirical bayes estimators: Maximum marginal likelihood vs minimum mse. In *Proceeding of the IFAC symposium on system identification* (pp. 125–130), Brussels, Belgium.

Aravkin, A., Burke, J. V., Chiuso, A., & Pillonetto, G. (2012b). On the MSE properties of empirical Bayes methods for sparse estimation. In *Proceeding of the IFAC symposium on system identification* (pp. 965–970), Brussels, Belgium.

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

Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge University Press.

Carli, F. P., Chen, T., & Ljung, L. (2017). Maximum entropy kernels for system identification. *IEEE Transactions on Automatic Control*, *62*, 1471–1477.

Chen, T. Continuous-time DC kernel-a stable generalized first-order spline kernel. IEEE Transactions on Automatic Control, to appear in 2019. https://doi.org/10.1109/TAC.2018.2825365.

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

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

Chen, T., Ardeshiri, T., Carli, F. P., Chiuso, A., Ljung, L., & Pillonetto, G. (2016). Maximum entropy properties of discrete-time first-order stable spline kernel. *Automatica*, *66*, 34–38.

Chen, T., Ohlsson, H., & Ljung, L. (2012). On the estimation of transfer functions, regularizations and Gaussian processes–Revisited. *Automatica*, *48*, 1525–1535.

Chen, T., & Pillonetto, G. On the stability of reproducing kernel Hilbert spaces of discrete-time impulse responses. Automatica, accepted on April 18, 2018, to appear.

Dinuzzo, F. (2015). Kernels for linear time invariant system identification. *SIAM Journal on Control and Optimization*, *53*, 3299–3317.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*, 55–67.

Ljung, L. (1999). *System identification: Theory for the user*. Upper Saddle River, NJ: Prentice-Hall.

Ljung, L. (2012). *System identification toolbox for use with MATLAB* (8th ed.). Natick, MA: The MathWorks, Inc..

Ljung, L., Singh, R., & Chen, T. (2015). Regularization features in the system identification toolbox. In *Proceedings of the IFAC symposium on system identification* (pp. 745–750), Beijing, China.

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

Petersen, K. B., & Pedersen, M. S. (2012). The Matrix Cookbook. http://matrixcookbook.com.

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

Pillonetto, G., & Chiuso, A. (2015). Tuning complexity in regularized kernel-based regression and linear system identification: The robustness of the marginal likelihood estimator. *Automatica*, *58*, 106–117.

Pillonetto, G., Chiuso, A., & De Nicolao, G. (2011). Prediction error identification of linear systems: A nonparametric gaussian regression approach. *Automatica*, *47*, 291–305.

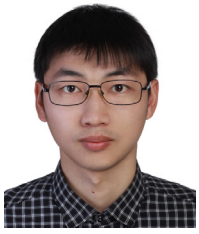
Pillonetto, G., & De Nicolao, G. (2010). A new kernel-based approach for linear system identification. *Automatica*, *46*, 81–93.

Pillonetto, G., Dinuzzo, F., Chen, T., De Nicolao, G., & Ljung, L. (2014). Kernel methods in system identification, machine learning and function estimation: A survey. *Automatica*, *50*, 657–682.

Theobald, C. M. (1974). Generalizations of mean square error applied to ridge regression. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *36*, 103–106.

Zorzi, M. (2017). On the robustness of the Bayes and Wiener estimators under model uncertainty. *Automatica*, *83*, 133–140.

Zorzi, M., & Chiuso, A. (2017). The harmonic analysis of kernel functions. arXiv preprint arXiv:1703.05216.



Biqiang Mu received the Bachelor of Engineering degree in Material Formation and Control Engineering from Sichuan University in 2008 and the Ph.D. degree in Operations Research and Cybernetics from the Academy of Mathematics and Systems Science, Chinese Academy of Sciences in 2013. He was a postdoc at the Wayne State University from 2013 to 2014 and also at the Western Sydney University from 2015 to 2016. He is a postdoc in the Division of Automatic Control, Department of Electrical Engineering, Linköping University since 2016 and he is also an assistant professor at the Academy of Mathematics and Systems Science, Chinese Academy of Sciences. His research interests include system identification (data-driven modelling and analysis), machine learning, and their applications.

Tianshi Chen received his Bachelor's degree and Master's degree both from Harbin Institute of Technology in 2001 and 2005, respectively. He received his Ph.D. degree in Automation and Computer-Aided Engineering from The Chinese University of Hong Kong in December 2008. From April 2009 to December 2015, he was working in the Division of Automatic Control, Department of Electrical Engineering, Linköping University, Linköping, Sweden, first as a Postdoc (April 2009–March 2011) and then as an Assistant Professor (April 2011–December 2015). In May 2015, he received the Youth Talents Award of the Thousand Talents Plan of China, and in December 2015, he returned to China and joined the Chinese University of Hong Kong, Shenzhen, as an Associate Professor. He has also been holding a visiting associate professor position at Linköping University since 2016.

He has been mainly working in the area of system identification (data-driven modelling and analysis), statistical signal processing, machine learning, data science, nonlinear control, and their applications. He has participated in several projects in Sweden, Europe and China. He is an Associate Editor for Automatica (2017–present), System & Control Letters (2017–present), and IEEE Control System Society Conference Editorial Board (2016–present).

Lennart Ljung received his Ph.D. in Automatic Control from Lund Institute of Technology in 1974. Since 1976 he is Professor of the chair of Automatic Control In Linkoping, Sweden. He has held visiting positions at Stanford and MIT and has written several books on System Identification and Estimation. He is an IEEE Fellow, an IFAC Fellow and an IFAC Advisor as well as a member of the Royal Swedish Academy of Sciences (KVA), a member of the Royal Swedish Academy of Engineering Sciences (IVA), an Honorary Member of the Hungarian Academy of Engineering and a Foreign Associate of the US National Academy of Engineering (NAE).

He has received honorary doctorates from the Baltic State Technical University in St Petersburg, from Uppsala University, Sweden, from the Technical University of Troyes, France, from the Catholic University of Leuven, Belgium and from Helsinki University of Technology, Finland. In 2002 he received the Quazza Medal from IFAC, in 2003 he received the Hendrik W. Bode Lecture Prize from the IEEE Control Systems Society, and he was the recipient of the IEEE Control Systems Award for 2007.