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