Available online at www.sciencedirect.com

ScienceDirect

IFAC PapersOnline CONFERENCE PAPER ARCHIVE

IFAC PapersOnLine 51-15 (2018) 203–208

Asymptotic Properties of Generalized Cross Validation Estimators for Regularized System Identification ⋆

Biqiang Mu [*,**] **Tianshi Chen** [***] **Lennart Ljung** [*]

[*] *Division of Automatic Control, Linköping University, Linköping, Sweden (e-mail: Biqiang.Mu@liu.se, Lennart.Ljung@liu.se).*
[**] *Key Laboratory of Systems and Control, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China*
[***] *School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, China (e-mail: tschen@cuhk.edu.cn)*

Abstract: In this paper, we study the asymptotic properties of the generalized cross validation (GCV) hyperparameter estimator and establish its connection with the Stein's unbiased risk estimators (SURE) as well as the mean squared error (MSE). It is shown that as the number of data goes to infinity, the GCV has the same asymptotic property as the SURE does and both of them converge to the best hyperparameter in the MSE sense. We illustrate the efficacy of the result by Monte Carlo simulations.

Keywords: Regularized system identification, Generalized cross-validation, Stein's unbiased risk estimators, Asymptotic analysis

1. INTRODUCTION

During the past few years, kernel-based regularization methods (KRM) for linear system identification, first introduced to the system identification community in Pillonetto and De Nicolao (2010) and then further developed in Pillonetto et al. (2011); Chen et al. (2012, 2014), have attracted intense interest in the community and have become a complement to the classical maximum likelihood/prediction error methods (ML/PEM) (Pillonetto and Chiuso, 2015; Chen et al., 2012; Ljung et al., 2015). The advantage of the KRM has been verified by a number of experimental evidences in Chen et al. (2012); Pillonetto et al. (2014) and also by the theoretic result given in Mu et al. (2018) that the KRM can reach a smaller mean squared error (MSE) than the ML/PEM if the kernel matrix is carefully chosen. Recent works for linear system identification by using this method include, e.g., the kernel design (Prando et al., 2017; Zorzi and Chiuso, 2017; Chen, 2018b, 2019; Chen and Pillonetto, 2018; Chen et al., 2018), hyperparameter estimators (Pillonetto and Chiuso, 2015; Mu et al., 2017b, 2018; Hong et al., 2018), input design (Fujimoto and Sugie, 2018; Mu et al., 2017a; Mu and Chen, 2018) and frequency domain counterpart (Lataire and Chen, 2016).

The implementation of the KRM involves two successive steps: kernel design and hyperparameter estimation, which aim at finding a good kernel matrix based on the data. The former is regarding how to embed the prior knowledge of the underlying system to be identified into the kernel matrix parameterized by a parameter vector, called hyperparameter and the latter is regarding how to estimate the hyperparameter based on the data, or equivalently, to tune model complexity of the estimated model in a continuous manner such that a good balance between the adherence to the data and model complexity is achieved.

The kernel design is to determine the underlying model structure of the kernel matrix for the KRM, which is analogous to the model structure selection for the ML/PEM. So far, many works have been done on this aspect and several kernels embedding various types of prior knowledge have been proposed, e.g., Pillonetto and De Nicolao (2010); Pillonetto et al. (2011); Chen et al. (2012, 2014); Dinuzzo (2015); Chen et al. (2016); Carli et al. (2017); Marconato et al. (2016); Pillonetto et al. (2016); Zorzi and Chiuso (2017); Chen (2018b, 2019); Chen and Pillonetto (2018).

The hyperparameter estimation plays a similar role as the model order selection for the ML/PEM. The survey of the KRM in Pillonetto et al. (2014) and the paper Pillonetto and Chiuso (2015) introduced many popular methods for hyperparameter estimation, such as the empirical Bayes (EB), C_p statistics, Stein's unbiased risk estimator (SURE), cross-validation (CV), and etc. There have been some results on the properties of the hyperparameter estimators reported in Aravkin et al. (2012a,b, 2014); Chen et al. (2014); Pillonetto and Chiuso (2015).

⋆ This work was supported in part by the National Natural Science Foundation of China under contract Nos. 61603379 and 61773329, the National Key Basic Research Program of China (973 Program) under contract No. 2014CB845301, the President Fund of Academy of Mathematics and Systems Science, CAS under contract No. 2015-hwyxqnrc-mbq, 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 funded by the Chinese University of Hong Kong, Shenzhen, and a research grant for junior researchers under contract No. 2014-5894 funded by Swedish Research Council.

2405-8963 © 2018, IFAC (International Federation of Automatic Control) Hosting by Elsevier Ltd. All rights reserved.
Peer review under responsibility of International Federation of Automatic Control.
10.1016/j.ifacol.2018.09.130

Recent works on this aspect are Mu et al. (2017b, 2018), where it is shown that the SURE method converges to the best hyperparameter minimizing the MSE as the number of data goes to infinity, while the more widely used EB estimator converges to the hyperparameter minimizing another different criterion.

In addition to the EB and SURE methods, the CV method is another major tool for hyperparameter estimation. The leave-one-out cross validation (LOOCV), also known as predicted residual sums of squares (PRESS) (Allen, 1974), is an important one of the CV family. The calculation of the PRESS is time-consuming and so the generalized cross validation (GCV) (Golub et al., 1979) can be thought of as a simplification of the PRESS. The general asymptotic properties of the CV method for discrete index (e.g., model order selection) have been extensively studied: e.g., Li (1987); Shao (1997). The application of the CV method to the KRM where the tuning parameter (hyperparameter) is continuous is less studied, except for some special cases, e.g. ridge regression and smoothing splines, Li (1986).

In this paper, we explore the asymptotic properties of the GCV for the KRM, where the ridge regression can be treated as a special case. Regardless of the parameterization of the kernel matrix, we show that the GCV is also asymptotically to minimize the MSE as the SURE does. This means that both GCV and SURE methods are asymptotically optimal and are asymptotically consistent estimates of the MSE. The computational complexity of the GCV and SURE methods is almost the same. Moreover, a merit of GCV is that it does not require to estimate the variance of the noise in comparison with the SURE method. This implies that the GCV may perform better than the SURE method for short data or ill-conditioned inputs. The simulation result given in Section 4 also indicates that the PRESS may be also asymptotically optimal for the cases considered in the simulation.

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 kernel design. In Section 3, we introduce the PRESS and GCV hyperparameter estimators and prove that the GCV is asymptotically optimal. In Section 4, we illustrate our theoretical results with Monte Carlo simulations. Finally, we conclude this paper in Section 4.4.

2. KERNEL-BASED REGULARIZATION METHODS FOR FIR MODEL ESTIMATION

2.1 Problem Statement

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, q is the forward shift operator: $qu(t) = u(t+1)$, $y(t), u(t)$ are the output and input, respectively, the noise $v(t)$ is a zero mean white noise with finite variance $0 < \sigma^2 < \infty$ and is independent of the input $u(t)$. Assume that the input $u(t)$ is known (deterministic) and the input-output data is collected at time instants $t = 1, \cdots, N$. The target is to estimate the rational transfer function

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

determined by the impulse response coefficients $\{g_k^0, k = 1, \cdots, \infty\}$, as well as possible based on the the available data $\{u(t-1), y(t)\}_{t=1}^N$.

The stability of $G_0(q)$ impliess that it is possible to truncate the infinite impulse response at a sufficiently high order, leading to the finite impulse response (FIR) model:

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

Accordingly, system (1) becomes a linear regression form

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

where $\phi(t) = [u(t-1), \cdots, u(t-n)]^T \in \mathbb{R}^n$, 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

$$\widehat{\theta}^{\mathrm{LS}} = \arg\min_{\theta\in\mathbb{R}^n} \|Y - \Phi\theta\|^2 \tag{5a}$$
$$= (\Phi^T\Phi)^{-1}\Phi^T Y, \tag{5b}$$

where $\|\cdot\|$ is the Euclidean norm, is unbiased but may have large variance and mean square error (MSE) (e.g., when the input is low-pass filtered white noise). The large variance problem can be mitigated if some bias is allowed.

2.2 Regularized Least Squares Methods

One feasible way to reduce the variance 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) estimate:

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

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. The mean squared error (MSE) of the RLS estimate relating to the prediction performance is given by, see e.g., Pillonetto and Chiuso (2015); Mu et al. (2018),

$$\mathrm{MSEy}(P) = E\left[\sum_{t=1}^{N}\left(\phi^T(t)\theta_0 + v^*(t) - \widehat{y}(t)\right)^2\right] \tag{7}$$
$$= \|\Phi P\Phi^T Q^{-1}\Phi\theta_0 - \Phi\theta_0\|^2 + N\sigma^2$$
$$+\sigma^2\mathrm{Tr}(\Phi P\Phi^T Q^{-2}\Phi P^T\Phi^T)$$
$$Q = \Phi P\Phi^T + \sigma^2 I_N,$$

where $E(\cdot)$ is the mathematical expectation, $\mathrm{Tr}(\cdot)$ is the trace of a square matrix, $\theta_0 = [g_1^0, \cdots, g_n^0]^T$ with g_i^0, $i = 1, \dots, n$, defined in (2), $\widehat{y}(t)$ is the i-th element of the predicted output

$$\widehat{Y} = \Phi\widehat{\theta}^{\mathrm{R}} = HY \tag{8}$$

$$H \triangleq \Phi P\Phi^T(\Phi P\Phi^T + \sigma^2 I_N)^{-1} \tag{9}$$

and $v^*(t)$ is an independent copy of the noise $v(t)$. It has been shown in Mu et al. (2018, Prop. 2) that for a suitably

chosen kernel matrix P, the RLS estimator (6b) has a smaller MSE than the LS estimator (5b).

2.3 Kernel Design

Kernel design is the first step of the kernel-based regularization method, which is regarding how to parameterize the kernel to embed the prior knowledge of the system to be identified

$$P(\eta), \quad \eta \in \Omega \subset \mathbb{R}^p. \tag{10}$$

Kernel design plays an analogous role in the model structure selection for the ML/PEM, and also determines the underlying model structure for the regularized FIR model (6b). So far, several kernels have been proposed, such as the diagonal correlated (DC) kernel and the tuned-correlated (TC) kernel (Chen et al., 2012), which are defined as follows:

$$\begin{aligned} \text{DC}: \quad & P_{kj}(\eta) = c\alpha^{(k+j)/2}\rho^{|j-k|}, \qquad (11)\\ & \eta = [c, \alpha, \rho] \in \Omega = \{c \geq 0, 0 \leq \alpha \leq 1, |\rho| \leq 1\};\\ \text{TC}: \quad & P_{kj}(\eta) = c\alpha^{\max(k,j)}, \qquad (12)\\ & \eta = [c, \alpha] \in \Omega = \{c \geq 0, 0 \leq \alpha \leq 1\}. \end{aligned}$$

where the TC kernel (12) is a special case of the DC kernel with $\rho = \sqrt{\lambda}$ (Chen et al., 2012).

3. HYPERPARAMETER ESTIMATION

When a parameterized family of the kernel matrix $P(\eta)$ has been chosen, the next task is to estimate, or "tune", a good hyperparameter η based on the data. Hyperparameter estimation plays a similar role as choosing the model order in the traditional parameter framework, which has a great impact on the regularization performance. Some effective tuning methods have been suggested in the literature, see e.g., Section 14 of Pillonetto et al. (2014), including the empirical Bayes (EB) method, the SURE methods, and the cross-validation. The papers Mu et al. (2017b, 2018) report the asymptotic properties of the EB and SURE method. It is shown that the SURE method is asymptotically optimal, while the EB is biased in general.

Lemma 1. (Mu et al., 2018, Theorem 1) Consider the hyperparameter estimators:

$$\text{SUREy}: \widehat{\eta}_{\text{Sy}} = \arg\min_{\eta\in\Omega} \mathcal{F}_{\text{Sy}}(P(\eta)) \tag{13}$$

$$\text{MSEy}: \widehat{\eta}_{\text{MSEy}} = \arg\min_{\eta\in\Omega} \text{MSEy}(P(\eta)) \tag{14}$$

$$\text{EB}: \widehat{\eta}_{\text{EB}} = \arg\min_{\eta\in\Omega} \mathcal{F}_{\text{EB}}(P(\eta)) \tag{15}$$

$$\begin{aligned} \mathcal{F}_{\text{Sy}}(P) &= \|Y - \Phi\widehat{\theta}^{\text{R}}\|^2 + 2\sigma^2\text{Tr}\big(H\big)\\ \mathcal{F}_{\text{EB}}(P) &= Y^TQ^{-1}Y + \log\det(Q). \end{aligned}$$

where $\text{MSEy}(P)$ is defined in (7). The asymptotically best hyperparameter in the MSEy sense is defined by

$$\begin{aligned} \eta_{\text{y}}^* &= \arg\min_{\eta\in\Omega} W_y(P(\eta), \Sigma, \theta_0)\\ W_y(P, \Sigma, \theta_0) &= \sigma^4\theta_0^TP^{-1}\Sigma^{-1}P^{-1}\theta_0 - 2\sigma^4\text{Tr}\big(\Sigma^{-1}P^{-1}\big) \end{aligned}$$

where the positive definite matrix Σ is the limit of $\Phi^T\Phi/N$. Suppose that $P(\eta)$ is a symmetric and positive definite parameterization. Thus we have as $N \to \infty$

$$\widehat{\eta}_{\text{Sy}} \to \eta_{\text{y}}^*, \quad \widehat{\eta}_{\text{MSEy}} \to \eta_{\text{y}}^*$$

almost surely, while

$$\widehat{\eta}_{\text{EB}} \to \eta_{\text{B}}^* = \arg\min_{\eta\in\Omega} \theta_0^TP(\eta)^{-1}\theta_0 + \log\det(P(\eta))$$

almost surely. In general, $\eta_{\text{B}}^* \neq \eta_{\text{y}}^*$.

Cross-validation (CV) is another widely used technique to estimate the hyperparameters besides the EB and SURE methods. The main idea of CV is to split data into two disjoint parts called estimation data and validation data, respectively. The hyperparameter value is estimated from the training data and the quality of the estimate is evaluated on the validation data. The hyperparameter value that gives the best performance on validation data are then selected.

The LOOCV, also known as PRESS, is a popular one of the CV family, where the validation set has only one data at each time. For the linear regression problem (4), the hyperparameter η is estimated by

$$\text{PRESS}: \quad \widehat{\eta} = \arg\min_{\eta\in\Omega} \sum_{t=1}^{N} \left(\frac{y(t) - \widehat{y}(t)}{1 - h_{tt}}\right)^2 \tag{16}$$

where $\widehat{y}(t)$ is the t-th element of the predicted output $\widehat{Y}$ defined in (8) and h_{tt} is the (t,t)-element of H defined in (9). In general, the computation of PRESS is time-consuming and hence the weights h_{tt} in the PRESS are replaced by their average for reducing the computational complexity. This leads to the generalized cross validation (GCV), which estimates η by

$$\text{GCV}: \quad \widehat{\eta}_{\text{GCV}} = \arg\min_{\eta\in\Omega} \mathcal{F}_{\text{GCV}}(P(\eta)) \tag{17a}$$

$$\mathcal{F}_{\text{GCV}}(P) = \frac{\sum_{t=1}^{N}\big(y(t) - \widehat{y}(t)\big)^2}{\big(1 - \text{Tr}(H)/N\big)^2}. \tag{17b}$$

In this paper, we will explore the asymptotic property of the GCV (17).

Theorem 1. Consider the hyperparameter estimation criterion GCV (17). Suppose that P is nonsingular and

$$\Phi^T\Phi/N \to \Sigma \tag{18}$$

almost surely as $N \to \infty$, where Σ is positive definite. Then we have as $N \to \infty$

$$\begin{aligned} &N\big(\mathcal{F}_{\text{GCV}}(P) - (Y^TY - Y^T\Phi(\Phi^T\Phi)^{-1}\Phi^TY)(1 + 2n/N)\big)\\ &\qquad \to W_y(P, \Sigma, \theta_0) + 3n^2\sigma^2 \end{aligned}$$

almost surely. In addition, suppose $P(\eta)$ is a symmetric and positive definite parameterization. Then we have as $N \to \infty$

$$\widehat{\eta}_{\text{GCV}} \to \eta_{\text{y}}^*$$

almost surely.

Remark 1. Assumption (18) is a relatively mild condition on the regressor sequence $\phi(t)$.

Proof. We have the expansion

$$\begin{aligned} \mathcal{F}_{\text{GCV}}(P) &= \frac{\|Y - \Phi\widehat{\theta}^{\text{R}}\|^2}{\big(1 - \text{Tr}(H)/N\big)^2}\\ &= \|Y - \Phi\widehat{\theta}^{\text{R}}\|^2\Big(1 + \frac{2\text{Tr}(H)}{N} + \frac{3(\text{Tr}(H))^2}{N^2} + O\Big(\frac{1}{N^3}\Big)\Big). \end{aligned}$$

by the Taylor formula

$$\frac{1}{(1-x)^2} = 1 + 2x + 3x^2 + O(x^3)$$

around $x = 0$ and $\mathrm{Tr}(H)/N = O(1/N)$. Let us define an estimate for the noise variance σ^2:

$$\begin{aligned}\widehat{\sigma}^2 &\triangleq \frac{1}{N}\|(I_N - \Phi(\Phi^T\Phi)^{-1}\Phi^T)Y\|^2 \\ &= \frac{1}{N}\left(Y^TY - Y^T\Phi(\Phi^T\Phi)^{-1}\Phi^TY\right) \to \sigma^2\end{aligned}$$

which is independent of P. Firstly, we have

$$\begin{aligned}&N\left(\|Y - \Phi\widehat{\theta}^{\mathrm{R}}\|^2 + Y^T\Phi(\Phi^T\Phi)^{-1}\Phi^TY - Y^TY\right) \\ &= \sigma^4 Y^TQ^{-1}\Phi\left(N(\Phi^T\Phi)^{-1}\right)\Phi^TQ^{-1}Y \\ &\to \sigma^4\theta_0P^{-1}\Sigma^{-1}P^{-1}\theta_0. \qquad (19)\end{aligned}$$

Further, we have

$$\begin{aligned}&N\left(\mathrm{Tr}(H)\frac{\|Y - \Phi\widehat{\theta}^{\mathrm{R}}\|^2}{N} - n\widehat{\sigma}^2\right) \\ &= \widehat{\sigma}^2N(\mathrm{Tr}(H) - n) + \mathrm{Tr}(H)\sigma^4Y^TQ^{-1}\Phi(\Phi^T\Phi)^{-1}\Phi^TQ^{-1}Y \\ &\to -\sigma^4\mathrm{Tr}(\Sigma^{-1}P^{-1}) \qquad (20)\end{aligned}$$

based on the limits

$$\begin{aligned}&N(\mathrm{Tr}(H) - n) \to -\sigma^2\mathrm{Tr}\left(\Sigma^{-1}P^{-1}\right), \quad \widehat{\sigma}^2 \to \sigma^2 \\ &\sigma^4Y^TQ^{-1}\Phi(\Phi^T\Phi)^{-1}\Phi^TQ^{-1}Y \to 0, \quad \mathrm{Tr}(H) \to n.\end{aligned}$$

At last, we have

$$\begin{aligned}&N(\mathrm{Tr}(H))^2\frac{\|Y - \Phi\widehat{\theta}^{\mathrm{R}}\|^2}{N^2} \\ &\qquad = (\mathrm{Tr}(H))^2\frac{\|Y - \Phi\widehat{\theta}^{\mathrm{R}}\|^2}{N} \to n^2\sigma^2. \qquad (21)\end{aligned}$$

Combining (19), (20), and (21), one yields

$$\begin{aligned}&\overline{\mathcal{F}}_{\mathrm{GCV}}(P) \triangleq N\left(\mathcal{F}_{\mathrm{GCV}}(P) - (1 + 2n/N)\widehat{\sigma}^2\right) \\ &\to \sigma^4\theta_0P^{-1}\Sigma^{-1}P^{-1}\theta_0 - 2\sigma^4\mathrm{Tr}(\Sigma^{-1}P^{-1}) + 3n^2\sigma^2.\end{aligned}$$

Since $(1 + 2n/N)\widehat{\sigma}^2$ is independent of P, we see that

$$\widehat{\eta}_{\mathrm{GCV}} = \arg\min_{\eta\in\Omega}\overline{\mathcal{F}}_{\mathrm{GCV}}(P(\eta)).$$

Thus we derive

$$\widehat{\eta}_{\mathrm{GCV}} \to \eta_{\mathrm{y}}^*$$

as $N \to \infty$ by applying the convergence result for extremum estimators in Ljung (1999, Theorem 8.2). ■

Remark 2. Comparing Theorem 1 and Lemma 1, we see that the GCV hyperparameter estimator (17) is also asymptotically optimal if we are concerned with the predictive performance of the estimated model.

4. SIMULATION RESULTS

In this section, we test the hyperparameter estimators PRESS and GCV given in (16) and (17), respectively, by the data used in Mu et al. (2018).

4.1 Test data-bank

The true system of order 30 is randomly generated by the method in Chen et al. (2012). Then for each test system, we consider four different test inputs: The first two test inputs are 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. The noise-free output is corrupted by an additive white Gaussian noise such that the signal-to-noise ratio (SNR), i.e., the ratio between the variance of the noise-free output and the noise, is uniformly distributed over $[1, 10]$, and is kept unchanged for the four test inputs. We consider data sets with the length $N = 500$ and 8000, respectively, for showing the small sample and large sample behavior of the hyperparameter estimators.

4.2 Simulation Setup

The measure of fit (Ljung, 2012) defined as follows :

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

where n is set to 200, is used to evaluate the quality of the RLS estimator (6b).

The TC kernel (12) is adopted and its hyperparameter $\eta = [c, \alpha]^T$ is estimated by using the estimators MSEy (14), SUREy (13), PRESS (16), and GCV (17).

4.3 Simulation results

We tested 1000 systems for each case. The average fits are given in Table 1. The boxplots of the 1000 fits for IT1, IT2, IT3, and IT4 are illustrated in Figs. 1–4, respectively.

Inputs	Sizes	MSEy	Sy	PRESS	GCV
IT1	$N=500$	78.07	53.83	56.61	55.74
	$N=8000$	88.08	78.39	78.26	78.41
IT2	$N=500$	87.02	86.03	86.24	86.24
	$N=8000$	96.67	96.60	96.60	96.60
IT3	$N=500$	41.61	-146.4	-85.95	-84.84
	$N=8000$	53.63	38.86	38.79	38.89
IT4	$N=500$	86.69	85.66	85.96	85.95
	$N=8000$	96.56	96.49	96.49	96.49

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

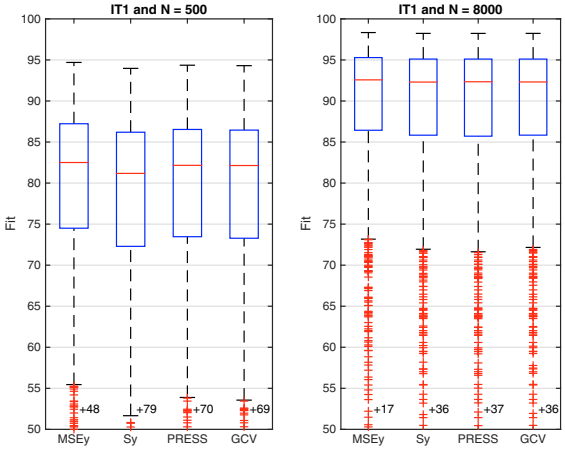

Fig. 1. Boxplot of the 1000 fits for the bandlimited white Gaussian noise input with the normalized band $[0, 0.6]$: $N = 500$ (left) and $N = 8000$ (right).

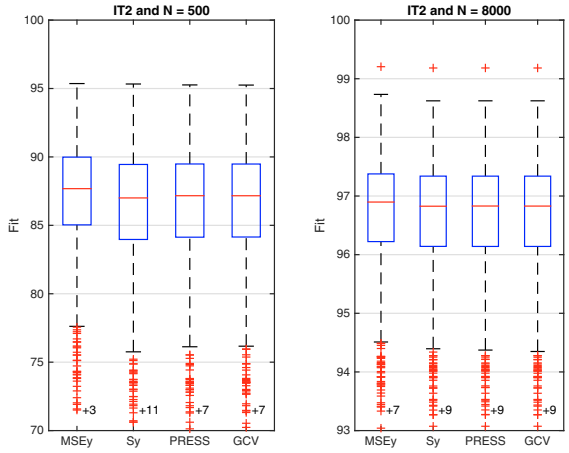

Fig. 2. Boxplot of the 1000 fits for the bandlimited white Gaussian noise input with the normalized band $[0, 1]$: data length $N = 500$ (left) and $N = 8000$ (right).

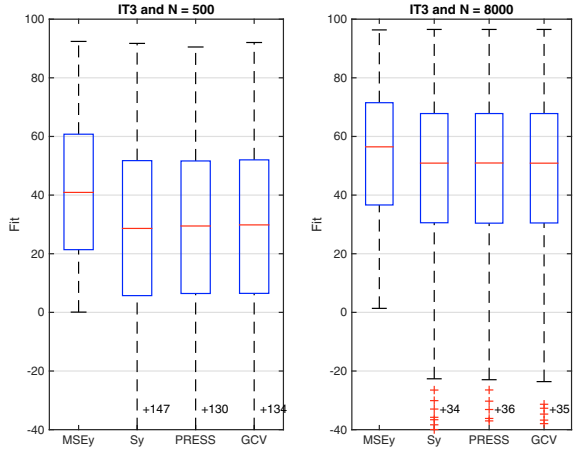

Fig. 3. Boxplot of the 1000 fits for the input (the white Gaussian noise of unit variance filtered by a second order rational transfer function $1/(1 - 0.95q^{-1})^2$): data length $N = 500$ (left) and $N = 8000$ (right).

4.4 Findings

Firstly, for all the tested cases, the fits given by PRESS and GCV are quite close and they are a little better than that of the SURE method, especially for the ill-conditioned inputs IT1 and IT3. This may be because they do not require to estimate the variance σ^2.

Secondly, the estimators including PRESS, GCV, and SURE perform indistinguishably from each other when $N = 8000$. In addition, they are very close to the oracle estimator MSEy for the well-conditioned inputs IT2 and IT4. This indicates the convergence stated in the Theorem 1 as we move from 500 to 8000 data.

Lastly, the simulation result indicates that the PRESS may be also asymptotically optimal in the cases considered here even though we have not proved this in this paper.

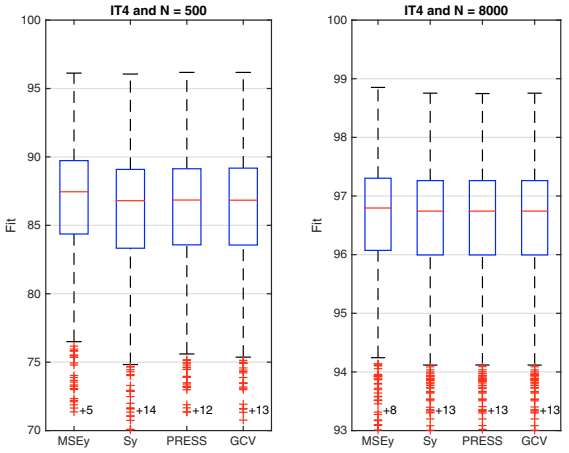

Fig. 4. Boxplot of the 1000 fits for the input (the white Gaussian noise of unit variance filtered by a second order rational transfer function $1/(1 - 0.05q^{-1})^2$): data length $N = 500$ (left) and $N = 8000$ (right).

5. CONCLUSION

This paper investigated the asymptotic behavior of the GCV as the number of data goes to infinity. We found that the GCV and SURE method have the same asymptotic properties and both of them are asymptotically optimal in the MSE sense. This provides us a theoretical support to adopt the GCV method to tune the hyperparameter of the KRM .

REFERENCES

Allen, D.M. (1974). The relationship between variable selection and data agumentation and a method for prediction. *Technometrics*, 16(1), 125–127.

Aravkin, A., Burke, J.V., Chiuso, A., and 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*, 125–130. Brussels, Belgium.

Aravkin, A., Burke, J.V., Chiuso, A., and Pillonetto, G. (2012b). On the mse properties of empirical bayes methods for sparse estimation. In *Proceeding of the IFAC Symposium on System Identification*, 965–970. Brussels, Belgium.

Aravkin, A., Burke, J.V., Chiuso, A., and 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*, 15(1), 217–252.

Carli, F.P., Chen, T., and Ljung, L. (2017). Maximum entropy kernels for system identification. *IEEE Transactions on Automatic Control*, 62(3), 1471–1477.

Chen, T., Andersen, M.S., Mu, B., Yin, F., Ljung, L., and Qin, S.J. (2018). Regularized lti system identification with multiple regularization matrix. In *The 18th IFAC Symposium on System Identification (SYSID)*.

Chen, T. (2019). Continuous-time DC kernel – a stable generalized first-order spline kernel. *IEEE Transactions on Automatic Control*.

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

Chen, T., Andersen, M.S., Ljung, L., Chiuso, A., and Pillonetto, G. (2014). System identification via sparse multiple kernel-based regularization using sequential convex optimization techniques. *IEEE Transactions on Automatic Control*, 59(11), 2933–2945.

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

Chen, T., Ohlsson, H., and Ljung, L. (2012). On the estimation of transfer functions, regularizations and gaussian processes–revisited. *Automatica*, 48(8), 1525–1535.

Chen, T. and Pillonetto, G. (2018). On the stability of reproducing kernel hilbert spaces of discrete-time impulse responses. *Automatica*.

Dinuzzo, F. (2015). Kernels for linear time invariant system identification. *SIAM Journal on Control and Optimization*, 53(5), 3299–3317.

Fujimoto, Y. and Sugie, T. (2018). Informative input design for kernel-based system identification. *Automatica*, 89, 37–43.

Golub, G.H., Heath, M., and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2), 215–223.

Hong, S., Mu, B., Yin, F., Andersen, M.S., and Chen, T. (2018). Multiple kernel based regularized system identification with SURE hyper-parameter estimator. In *The 18th IFAC Symposium on System Identification (SYSID)*.

Lataire, J. and Chen, T. (2016). Transfer function and transient estimation by gaussian process regression in the frequency domain. *Automatica*, 72, 217–229.

Li, K.C. (1986). Asymptotic optimality of cl and generalized cross-validation in ridge regression with application to spline smoothing. *The Annals of Statistics*, 1101–1112.

Li, K.C. (1987). Asymptotic optimality for c_p, c_l, cross-validation and generalized cross-validation: discrete index set. *The Annals of Statistics*, 958–975.

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

Ljung, L. (2012). *System Identification Toolbox for Use with MATLAB*. The MathWorks, Inc., Natick, MA, 8th ed. edition.

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

Marconato, A., Schoukens, M., and Schoukens, J. (2016). Filter-based regularisation for impulse response modelling. *IET Control Theory & Applications*, 11(2), 194–204.

Mu, B. and Chen, T. (2018). On input design for regularized lti system identification: Power-constrained input. *arXiv:1708.05539*.

Mu, B., Chen, T., and Ljung, L. (2017a). On the input design for kernel-based regularized lti system identification: Power-constrained inputs. In *Proceeding of the 56th IEEE Conference on Decision and Control*, 5262–5267. Melbourne, Australia.

Mu, B., Chen, T., and Ljung, L. (2017b). Tuning of hyperparameters for fir models–an asymptotic theory. In *Proceedings of the 20th IFAC World Congress*, 2818–2823. Toulouse, France.

Mu, B., Chen, T., and Ljung, L. (2018). On asymptotic properties of hyperparameter estimators for kernel-based regularization methods. *Automatica*.

Pillonetto, G., Chen, T., Chiuso, A., De Nicolao, G., and 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. and 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., and De Nicolao, G. (2011). Prediction error identification of linear systems: A nonparametric gaussian regression approach. *Automatica*, 47(2), 291–305.

Pillonetto, G. and De Nicolao, G. (2010). A new kernel-based approach for linear system identification. *Automatica*, 46(1), 81–93.

Pillonetto, G., Dinuzzo, F., Chen, T., De Nicolao, G., and Ljung, L. (2014). Kernel methods in system identification, machine learning and function estimation: A survey. *Automatica*, 50(3), 657–682.

Prando, G., Chiuso, A., and Pillonetto, G. (2017). Maximum entropy vector kernels for mimo system identification. *Automatica*, 79, 326–339.

Shao, J. (1997). An asymptotic theory for linear model selection. *Statistica Sinica*, 221–242.

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