IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, vol.E86-A, no.7, pp.1885-1895, Jul. 2003.

Improving Precision of the Subspace Information Criterion

Masashi Sugiyama

Department of Computer Science,
Graduate School of Information Science and Engineering,
Tokyo Institute of Technology.

2-12-1, O-okayama, Meguro-ku, Tokyo, 152-8552, Japan.

sugi@cs.titech.ac.jp
http://sugiyama-www.cs.titech.ac.jp/~sugi/

Abstract

Evaluating the generalization performance of learning machines without using additional test samples is one of the most important issues in the machine learning community. The subspace information criterion (SIC) is one of the methods for this purpose, which is shown to be an unbiased estimator of the generalization error with finite samples. Although the mean of SIC agrees with the true generalization error even in small sample cases, the scatter of SIC can be large under some severe conditions. In this paper, we therefore investigate the causes of degrading the precision of SIC, and discuss how its precision could be improved.

Keywords

machine learning, generalization capability, subspace information criterion, generalized inverse, unbiased estimator, variance.

1 Introduction

Evaluating the generalization performance of learning machines without using additional test samples is one of the central issues in the machine learning community. So far, a number of methods for estimating the generalization error have been proposed.

Akaike's information criterion (AIC) [1, 22, 14, 12] is an estimator of the generalization error defined by the Kullback-Leibler divergence. A notable property of AIC is that it is an *asymptotic* unbiased estimator of the generalization error, i.e., it becomes unbiased as the number of training examples goes to infinity.

Cross-validation (CV) [18, 3] is a method that estimates the expected prediction error by dividing the training set into k disjoint subsets. $(k-1)$ subsets are used for training and the rest is used for validation. This procedure is repeated for all k combinations and the mean validation error is outputted as an estimate of the prediction error. In the extreme case that $k = n$, where n is the number of training examples, it is specially called the leave-one-out CV. It is shown that for finite samples, the leave-one-out CV gives an *almost* unbiased[1] estimate of the expected prediction error [13, 16].

The subspace information criterion (SIC) [20, 23, 19] is an estimator of the generalization error defined by the function space norm. SIC is proved to be an *exact* unbiased estimator of an essential part of the generalization error with finite samples[2].

The methods described above have common excellent properties that their unbiasedness is theoretically guaranteed in the asymptotic, almost, or exact sense. However, they still have weakness, for example, under some severe conditions, their scatter (or variance) can be large [17, 16, 24]. For this reason, further investigation is needed to improve their precision.

In this paper, we therefore investigate the causes of degrading the precision of SIC, and discuss how its precision could be improved. More specifically, we investigate two causes: the large scatter caused by numerical instability of generalized inverse and the large variance caused by the unbiasedness. For the first cause, we show that, under a certain condition, an essential part of SIC can be exactly calculated without actually calculating generalized inverse. This will completely releases us from the first problem. For the second cause, we propose an alternative criterion aimed at reducing the variance of SIC by sacrificing its unbiasedness.

The rest of this article is organized as follows. In Section 2, the problem of supervised learning is mathematically formulated. In Section 3, the derivation of the original SIC is reviewed. In Section 4, we tackle the problem of reducing the scatter of SIC caused by numerical instability of generalized inverse. In Section 5, we tackle the problem of reducing the variance of SIC by sacrificing its unbiasedness. In Section 6, computer simulations are performed to illustrate how the consequences in Sections 4 and 5 contribute to improving the precision of SIC. Finally, Section 7 gives conclusions and future prospects.

[1] The term "almost unbiased" refers to the fact that the leave-one-out CV provides an unbiased estimate for training with $(n-1)$ samples.

[2] Note that we can not simply compare AIC, CV, and SIC because the conditions assumed behind are all different. See [21] for further discussions.

Nomenclature used in this article is summarized in Table 1.

2 Formulation of Supervised Learning

Supervised learning can be viewed as a function approximation problem. For this reason, we discuss the problem of approximating an unknown real-valued function $f(\boldsymbol{x})$ of d real variables from training examples.

Let $\mathcal{D}$ ($\subset \mathbb{R}^d$) be the domain of the learning target function $f(\boldsymbol{x})$. Let $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ be the training examples, where n is the number of training examples, $\boldsymbol{x}_i$ ($\in \mathcal{D}$) is the sample point, and y_i ($\in \mathbb{R}$) is the sample value degraded by unknown zero-mean additive noise ϵ_i, i.e.,

$$y_i = f(\boldsymbol{x}_i) + \epsilon_i. \tag{1}$$

Let us consider the case that the unknown learning target function $f(\boldsymbol{x})$ belongs to a specified reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ (see e.g., [4, 26, 25, 5])[3]. We denote the reproducing kernel of $\mathcal{H}$ by $K(\boldsymbol{x}, \boldsymbol{x}')$. We will obtain the learning result function $\hat{f}(\boldsymbol{x})$ by the following kernel regression model:

$$\hat{f}(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i K(\boldsymbol{x}, \boldsymbol{x}_i), \tag{2}$$

where $\{\alpha_i\}_{i=1}^n$ are parameters to be estimated from training examples. In this article, we focus on the case that the estimated parameters $\{\hat{\alpha}_i\}_{i=1}^n$ are given by linear combinations of sample values $\{y_i\}_{i=1}^n$. More specifically, letting

$$\begin{aligned} \boldsymbol{y} &= (y_1, y_2, \ldots, y_n)^\top, && (3)\\ \hat{\boldsymbol{\alpha}} &= (\hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_n)^\top, && (4) \end{aligned}$$

where $^\top$ denotes the transpose of a vector (or a matrix), we consider the case that the parameter vector $\hat{\boldsymbol{\alpha}}$ is estimated by

$$\hat{\boldsymbol{\alpha}} = \boldsymbol{X}\boldsymbol{y}, \tag{5}$$

where $\boldsymbol{X}$ is an n-dimensional matrix that is irrelevant to the noise $\{\epsilon_i\}_{i=1}^n$. $\boldsymbol{X}$ is called the learning matrix. The form (5) includes, for example, least mean squares estimate, kernel ridge regression [5], and a certain form of Gaussian process regression [27].

The purpose of regression is to obtain the optimal approximation $\hat{f}(\boldsymbol{x})$ to the unknown learning target function $f(\boldsymbol{x})$. We measure the generalization error of $\hat{f}(\boldsymbol{x})$ by the following criterion.

$$J_0 = \mathrm{E}_{\boldsymbol{\epsilon}} \|\hat{f} - f\|_{\mathcal{H}}^2, \tag{6}$$

[3] In our early work [20], only finite dimensional RKHSs could be dealt with. However, this restriction has been completely removed in the reference [19]. The current paper is based on the latter work so we do not impose any restrictions on the choice of the RKHS, e.g., infinite dimensional RKHSs are also allowed.

Table 1: Nomenclature

$f(\boldsymbol{x})$	Learning target function
$\mathcal{D}$	Domain of input $\boldsymbol{x}$
d	Dimension of input space $\mathcal{D}$
$\mathcal{H}$	Reproducing kernel Hilbert space including $f(\boldsymbol{x})$
$K(\cdot,\cdot)$	Reproducing kernel of $\mathcal{H}$
$\boldsymbol{x}_i$	Training sample point
y_i	Training sample value: $y_i = f(\boldsymbol{x}_i) + \epsilon_i$
ϵ_i	Zero-mean noise included in y_i
$\boldsymbol{Q}$	Noise covariance matrix
$(\boldsymbol{x}_i, y_i)$	Training example
n	Number of training examples
$\mathcal{S}$	Subspace of $\mathcal{H}$ spanned by $\{K(\boldsymbol{x}, \boldsymbol{x}_i)\}_{i=1}^{n}$
$f_{\mathcal{S}}(\boldsymbol{x})$	Orthogonal projection of $f(\boldsymbol{x})$ onto $\mathcal{S}$
$\boldsymbol{\epsilon}$	Noise vector: $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)^\top$
$\boldsymbol{z}$	Noiseless sample value vector: $\boldsymbol{z} = (f(\boldsymbol{x}_1), f(\boldsymbol{x}_2), \ldots, f(\boldsymbol{x}_n))^\top$
$\boldsymbol{y}$	Noisy sample value vector: $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^\top$
$^\top$	Transpose of a matrix or vector
$\mathrm{E}_{\boldsymbol{\epsilon}}$	Expectation over noise
$\hat{f}(\boldsymbol{x})$	Kernel regression model: $\hat{f}(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i K(\boldsymbol{x}, \boldsymbol{x}_i)$
α_i	Parameter in kernel regression model $\hat{f}(\boldsymbol{x})$
$\boldsymbol{\alpha}^*$	True parameter vector corresponding to $f_{\mathcal{S}}(\boldsymbol{x})$
$\hat{\boldsymbol{\alpha}}$	Estimated parameter vector
$\boldsymbol{X}$	Learning matrix: $\hat{\boldsymbol{\alpha}} = \boldsymbol{X}\boldsymbol{y}$
$\hat{\boldsymbol{\alpha}}^u$	Unbiased estimate of true $\boldsymbol{\alpha}^*$: $\mathrm{E}_{\boldsymbol{\epsilon}}\hat{\boldsymbol{\alpha}}^u = \boldsymbol{\alpha}^*$
λ	Regularization parameter
$\boldsymbol{T}$	Regularization matrix
$\boldsymbol{K}$	Kernel matrix: $\boldsymbol{K}_{i,j} = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$
$\boldsymbol{I}$	Identity matrix
$\langle\cdot,\cdot\rangle_{\mathcal{H}}$	Inner product in $\mathcal{H}$
$\|\cdot\|_{\mathcal{H}}$	Norm in $\mathcal{H}$
$\langle\cdot,\cdot\rangle$	Euclidean inner product in $\mathbb{R}^n$
$\|\cdot\|$	Euclidean norm in $\mathbb{R}^n$
$\|\cdot\|_{\boldsymbol{K}}$	Weighted norm by $\boldsymbol{K}$: $\|\cdot\|_{\boldsymbol{K}}^2 = \langle \boldsymbol{K}\cdot, \cdot\rangle$
$\mathrm{tr}(\cdot)$	Trace of a matrix
J_0	Generalization error: $J_0 = \mathrm{E}_{\boldsymbol{\epsilon}}\|\hat{f} - f\|_{\mathcal{H}}^2$
J_1	$J_1 = \mathrm{E}_{\boldsymbol{\epsilon}}\|\hat{f} - f_{\mathcal{S}}\|_{\mathcal{H}}^2 = J_0 + \text{constant}$
J_2	$J_2 = \mathrm{E}_{\boldsymbol{\epsilon}}\|\hat{f}\|_{\mathcal{H}}^2 - 2\mathrm{E}_{\boldsymbol{\epsilon}}\langle \hat{f}, f_{\mathcal{S}}\rangle_{\mathcal{H}} = J_1 + \text{constant}$
†	Moore-Penrose generalized inverse
$\mathcal{R}(\cdot)$	Range of a matrix

where $\mathrm{E}_{\boldsymbol{\epsilon}}$ denotes the expectation over the noise and $\|\cdot\|_{\mathcal{H}}$ is the norm in the RKHS $\mathcal{H}$. This error measure is often used in the field of function approximation (e.g., [6, 8, 7]).

In many statistical learning theories (e.g., [1, 22, 14, 12]), the training sample points $\{\boldsymbol{x}_i\}_{i=1}^n$ as well as the noise $\{\epsilon_i\}_{i=1}^n$ are treated as probabilistic and the expectation over both of them is often taken. In contrast, we consider a fixed design $\{\boldsymbol{x}_i\}_{i=1}^n$ and do not take the expectation over the training sample points. Therefore, our approach is more data-dependent.

The generalization error (6) can not be directly calculated since it includes the unknown learning target function $f(\boldsymbol{x})$. This paper is devoted to investigating how the generalization error J_0 is estimated.

3 The Subspace Information Criterion

The subspace information criterion (SIC) [20, 19] is an unbiased estimator of an essential part of the generalization error J_0. In this section, we review the derivation of SIC.

Let $\mathcal{S}$ be the subspace spanned by $\{K(\boldsymbol{x}, \boldsymbol{x}_i)\}_{i=1}^n$, and let $f_{\mathcal{S}}(\boldsymbol{x})$ be the orthogonal projection of $f(\boldsymbol{x})$ onto $\mathcal{S}$. Let ${f_{\mathcal{S}}}^{\perp}$ be

$$ {f_{\mathcal{S}}}^{\perp} = f - f_{\mathcal{S}}. \tag{7}$$

Note that ${f_{\mathcal{S}}}^{\perp}$ is orthogonal to $\hat{f}$ and $f_{\mathcal{S}}$. Let $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ be the inner product in $\mathcal{H}$. Then the generalization error J_0 can be expressed by

$$\begin{aligned} J_0 &= \mathrm{E}_{\boldsymbol{\epsilon}}\|\hat{f} - f_{\mathcal{S}} - {f_{\mathcal{S}}}^{\perp}\|_{\mathcal{H}}^2 \\ &= \mathrm{E}_{\boldsymbol{\epsilon}}\|\hat{f} - f_{\mathcal{S}}\|_{\mathcal{H}}^2 - 2\langle \hat{f} - f_{\mathcal{S}}, {f_{\mathcal{S}}}^{\perp}\rangle_{\mathcal{H}} + \|{f_{\mathcal{S}}}^{\perp}\|_{\mathcal{H}}^2 \\ &= \mathrm{E}_{\boldsymbol{\epsilon}}\|\hat{f} - f_{\mathcal{S}}\|_{\mathcal{H}}^2 + \|{f_{\mathcal{S}}}^{\perp}\|_{\mathcal{H}}^2. \end{aligned} \tag{8}$$

Since the second term $\|{f_{\mathcal{S}}}^{\perp}\|_{\mathcal{H}}^2$ does not depend on $\hat{f}$, we will ignore it and let us denote the first term by J_1 (see Figure 1):

$$ J_1 = \mathrm{E}_{\boldsymbol{\epsilon}}\|\hat{f} - f_{\mathcal{S}}\|_{\mathcal{H}}^2. \tag{9}$$

Since $f_{\mathcal{S}}(\boldsymbol{x})$ belongs to $\mathcal{S}$, it can be expressed by

$$ f_{\mathcal{S}}(\boldsymbol{x}) = \sum_{i=1}^n \alpha_i^* K(\boldsymbol{x}, \boldsymbol{x}_i), \tag{10}$$

where the parameters $\{\alpha_i^*\}_{i=1}^n$ are unknown[4]. Let

$$ \boldsymbol{\alpha}^* = (\alpha_1^*, \alpha_2^*, \ldots, \alpha_n^*)^\top, \tag{11}$$

[4] When $\{K(\boldsymbol{x}, \boldsymbol{x}_i)\}_{i=1}^n$ are linearly dependent, $\{\alpha_i^*\}_{i=1}^n$ are not determined uniquely. In this case, we adopt the minimum norm one.

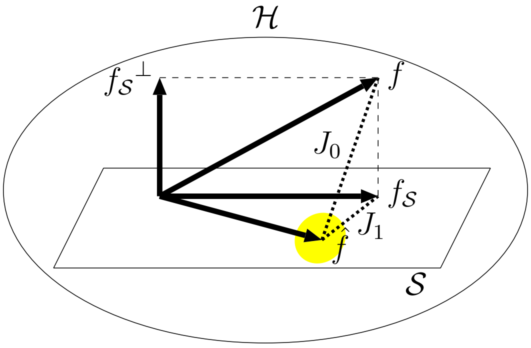


Figure 1: Decomposition of learning target function f.

and $\boldsymbol{K}$ be the so-called kernel matrix, i.e., the (i, j)-th element of $\boldsymbol{K}$ is given by

$$\boldsymbol{K}_{i,j} = K(\boldsymbol{x}_i, \boldsymbol{x}_j). \tag{12}$$

Then J_1 is expressed as

$$\begin{aligned} J_1 &= \mathrm{E}_{\boldsymbol{\epsilon}}\|\hat{f} - f_{\mathcal{S}}\|_{\mathcal{H}}^2 \\ &= \mathrm{E}_{\boldsymbol{\epsilon}}\|\sum_{i=1}^{n}(\hat{\alpha}_i - \alpha_i^*)K(\cdot, \boldsymbol{x}_i)\|_{\mathcal{H}}^2 \\ &= \mathrm{E}_{\boldsymbol{\epsilon}}\sum_{i,j=1}^{n}(\hat{\alpha}_i - \alpha_i^*)(\hat{\alpha}_j - \alpha_j^*)\langle K(\cdot, \boldsymbol{x}_j), K(\cdot, \boldsymbol{x}_i)\rangle_{\mathcal{H}} \\ &= \mathrm{E}_{\boldsymbol{\epsilon}}\sum_{i,j=1}^{n}(\hat{\alpha}_i - \alpha_i^*)(\hat{\alpha}_j - \alpha_j^*)K(\boldsymbol{x}_i, \boldsymbol{x}_j) \\ &= \mathrm{E}_{\boldsymbol{\epsilon}}\langle \boldsymbol{K}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*), \hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\rangle, \end{aligned} \tag{13}$$

where the inner product $\langle\cdot, \cdot\rangle$ in the last equation is the ordinary Euclidean inner product in $\mathbb{R}^n$, i.e., $\langle\boldsymbol{\alpha}_a, \boldsymbol{\alpha}_b\rangle = \boldsymbol{\alpha}_b^{\top}\boldsymbol{\alpha}_a$. For convenience, let us define the weighted norm in $\mathbb{R}^n$:

$$\|\boldsymbol{\alpha}\|_{\boldsymbol{K}}^2 = \langle\boldsymbol{K}\boldsymbol{\alpha}, \boldsymbol{\alpha}\rangle. \tag{14}$$

Then J_1 is expressed as

$$J_1 = \mathrm{E}_{\boldsymbol{\epsilon}}\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_{\boldsymbol{K}}^2. \tag{15}$$

It is known that J_1 can be decomposed into the bias and variance terms [9, 11]:

$$J_1 = \|\mathrm{E}_{\boldsymbol{\epsilon}}\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_{\boldsymbol{K}}^2 + \mathrm{E}_{\boldsymbol{\epsilon}}\|\hat{\boldsymbol{\alpha}} - \mathrm{E}_{\boldsymbol{\epsilon}}\hat{\boldsymbol{\alpha}}\|_{\boldsymbol{K}}^2. \tag{16}$$

Let us define the noiseless sample value vector $\boldsymbol{z}$ and the noise vector $\boldsymbol{\epsilon}$ by

$$\boldsymbol{z} = (f(\boldsymbol{x}_1), f(\boldsymbol{x}_2), \ldots, f(\boldsymbol{x}_n))^{\top}, \tag{17}$$

$$\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)^{\top}. \tag{18}$$

Then the (noisy) sample value vector $\boldsymbol{y}$ defined by Eq.(3) is expressed as

$$\boldsymbol{y} = \boldsymbol{z} + \boldsymbol{\epsilon}. \tag{19}$$

Recalling that the mean noise $\mathrm{E}_{\boldsymbol{\epsilon}}\boldsymbol{\epsilon}$ is zero, we can express the variance term $\mathrm{E}_{\boldsymbol{\epsilon}}\|\hat{\boldsymbol{\alpha}} - \mathrm{E}_{\boldsymbol{\epsilon}}\hat{\boldsymbol{\alpha}}\|_{\boldsymbol{K}}^2$ in Eq.(16) as

$$\begin{aligned}\mathrm{E}_{\boldsymbol{\epsilon}}\|\hat{\boldsymbol{\alpha}} - \mathrm{E}_{\boldsymbol{\epsilon}}\hat{\boldsymbol{\alpha}}\|_{\boldsymbol{K}}^2 &= \mathrm{E}_{\boldsymbol{\epsilon}}\|\boldsymbol{X}\boldsymbol{y} - \mathrm{E}_{\boldsymbol{\epsilon}}\boldsymbol{X}\boldsymbol{y}\|_{\boldsymbol{K}}^2 \\ &= \mathrm{E}_{\boldsymbol{\epsilon}}\|\boldsymbol{X}(\boldsymbol{z}+\boldsymbol{\epsilon}) - \boldsymbol{X}\boldsymbol{z}\|_{\boldsymbol{K}}^2 \\ &= \mathrm{E}_{\boldsymbol{\epsilon}}\|\boldsymbol{X}\boldsymbol{\epsilon}\|_{\boldsymbol{K}}^2 \\ &= \mathrm{tr}\left(\boldsymbol{K}\boldsymbol{X}\boldsymbol{Q}\boldsymbol{X}^\top\right),\end{aligned} \tag{20}$$

where $\boldsymbol{Q}$ is the noise covariance matrix and $\mathrm{tr}(\cdot)$ denotes the trace of a matrix, i.e., the sum of diagonal elements.

Eq.(20) implies that the variance term $\mathrm{E}_{\boldsymbol{\epsilon}}\|\hat{\boldsymbol{\alpha}} - \mathrm{E}_{\boldsymbol{\epsilon}}\hat{\boldsymbol{\alpha}}\|_{\boldsymbol{K}}^2$ in Eq.(16) can be calculated if the noise covariance matrix $\boldsymbol{Q}$ is available. When $\boldsymbol{Q}$ is unknown, one of the practical methods for estimating $\boldsymbol{Q}$ is given as follows:

$$\hat{\boldsymbol{Q}} = \hat{\sigma}^2\boldsymbol{I}, \tag{21}$$

where $\boldsymbol{I}$ is the identity matrix and $\hat{\sigma}^2$ is an estimate of the noise variance given, e.g., by

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n \left(\hat{f}(\boldsymbol{x}_i) - y_i\right)^2}{n - \mathrm{tr}(\boldsymbol{K}\boldsymbol{X})} = \frac{\|\boldsymbol{K}\boldsymbol{X}\boldsymbol{y} - \boldsymbol{y}\|^2}{n - \mathrm{tr}(\boldsymbol{K}\boldsymbol{X})}. \tag{22}$$

Note that $\|\cdot\|$ in the last equation of Eq.(22) denotes the ordinary Euclidean norm in $\mathbb{R}^n$.

On the other hand, the bias term $\|\mathrm{E}_{\boldsymbol{\epsilon}}\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_{\boldsymbol{K}}^2$ in Eq.(16) is totally inaccessible since both $\mathrm{E}_{\boldsymbol{\epsilon}}\hat{\boldsymbol{\alpha}}$ and $\boldsymbol{\alpha}^*$ are unknown. The key idea of SIC is to use an unbiased estimate $\hat{\boldsymbol{\alpha}}^u$ of the unknown true parameter vector $\boldsymbol{\alpha}^*$:

$$\mathrm{E}_{\boldsymbol{\epsilon}}\hat{\boldsymbol{\alpha}}^u = \boldsymbol{\alpha}^*. \tag{23}$$

Indeed, such an unbiased estimator can be obtained by (see [19])

$$\hat{\boldsymbol{\alpha}}^u = \boldsymbol{K}^\dagger\boldsymbol{y}, \tag{24}$$

where † denotes the Moore-Penrose generalized inverse.

Using the unbiased estimate $\hat{\boldsymbol{\alpha}}^u$, we can roughly estimate the bias term $\|\mathrm{E}_{\boldsymbol{\epsilon}}\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_{\boldsymbol{K}}^2$ in Eq.(16) by $\|\hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}^u\|_{\boldsymbol{K}}^2$ (see Figure 2). More specifically, we have

$$\begin{aligned}&\|\mathrm{E}_{\boldsymbol{\epsilon}}\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_{\boldsymbol{K}}^2 \\ &\quad = \|\hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}^u\|_{\boldsymbol{K}}^2 - \|\hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}^u\|_{\boldsymbol{K}}^2 + \|\mathrm{E}_{\boldsymbol{\epsilon}}\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_{\boldsymbol{K}}^2 \\ &\quad = \|\hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}^u\|_{\boldsymbol{K}}^2 \\ &\qquad -\|\mathrm{E}_{\boldsymbol{\epsilon}}(\hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}^u) - \mathrm{E}_{\boldsymbol{\epsilon}}(\hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}^u) + \hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}^u\|_{\boldsymbol{K}}^2\end{aligned}$$

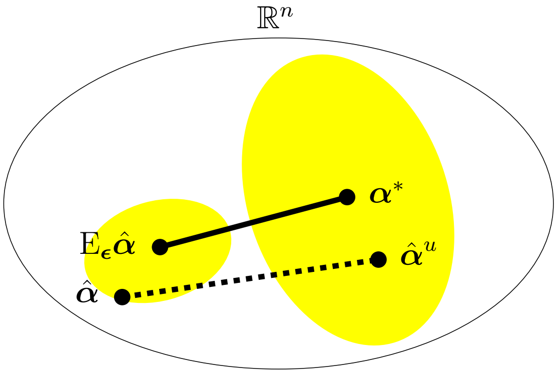

$\mathbb{R}^n$
$\boldsymbol{\alpha}^*$
$\mathrm{E}_{\boldsymbol{\epsilon}}\hat{\boldsymbol{\alpha}}$
$\hat{\boldsymbol{\alpha}}^u$
$\hat{\boldsymbol{\alpha}}$

Figure 2: Basic idea of SIC. The bias term $\|\mathrm{E}_{\boldsymbol{\epsilon}}\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|^2_{\boldsymbol{K}}$ (depicted by the solid line) can be roughly estimated by $\|\hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}^u\|^2_{\boldsymbol{K}}$ (depicted by the dotted line).

$$\begin{aligned}
&+\|\mathrm{E}_{\boldsymbol{\epsilon}}(\hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}^u)\|^2_{\boldsymbol{K}} \\
=\ &\|\hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}^u\|^2_{\boldsymbol{K}} - \|\mathrm{E}_{\boldsymbol{\epsilon}}(\hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}^u)\|^2_{\boldsymbol{K}} \\
&-2\langle \boldsymbol{K}\mathrm{E}_{\boldsymbol{\epsilon}}(\hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}^u), -\mathrm{E}_{\boldsymbol{\epsilon}}(\hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}^u) + \hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}^u\rangle \\
&-\| - \mathrm{E}_{\boldsymbol{\epsilon}}(\hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}^u) + \hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}^u\|^2_{\boldsymbol{K}} \\
&+\|\mathrm{E}_{\boldsymbol{\epsilon}}(\hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}^u)\|^2_{\boldsymbol{K}} \\
=\ &\|\hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}^u\|^2_{\boldsymbol{K}} \\
&+2\langle \boldsymbol{K}\mathrm{E}_{\boldsymbol{\epsilon}}(\hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}^u), \mathrm{E}_{\boldsymbol{\epsilon}}(\hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}^u) - (\hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}^u)\rangle \\
&-\|\mathrm{E}_{\boldsymbol{\epsilon}}(\hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}^u) - (\hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}^u)\|^2_{\boldsymbol{K}}. \qquad (25)
\end{aligned}$$

However, the second and third terms in the last equation of Eq.(25) are still inaccessible since $\mathrm{E}_{\boldsymbol{\epsilon}}(\hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}^u)$ is unknown, so we will average out these terms over the noise. Then the second term vanishes:

$$\mathrm{E}_{\boldsymbol{\epsilon}}\langle \boldsymbol{K}\mathrm{E}_{\boldsymbol{\epsilon}}(\hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}^u), \mathrm{E}_{\boldsymbol{\epsilon}}(\hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}^u) - (\hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}^u)\rangle = 0, \qquad (26)$$

and the third term is reduced to

$$\begin{aligned}
&\mathrm{E}_{\boldsymbol{\epsilon}}\left(\|\mathrm{E}_{\boldsymbol{\epsilon}}(\hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}^u) - (\hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}^u)\|^2_{\boldsymbol{K}}\right) \\
&\quad = \mathrm{E}_{\boldsymbol{\epsilon}}\left(\|\mathrm{E}_{\boldsymbol{\epsilon}}(\boldsymbol{X} - \boldsymbol{K}^\dagger)\boldsymbol{y} - (\boldsymbol{X} - \boldsymbol{K}^\dagger)\boldsymbol{y}\|^2_{\boldsymbol{K}}\right) \\
&\quad = \mathrm{E}_{\boldsymbol{\epsilon}}\left(\|(\boldsymbol{X} - \boldsymbol{K}^\dagger)\boldsymbol{z} - (\boldsymbol{X} - \boldsymbol{K}^\dagger)(\boldsymbol{z} + \boldsymbol{\epsilon})\|^2_{\boldsymbol{K}}\right) \\
&\quad = \mathrm{E}_{\boldsymbol{\epsilon}}\left(\|(\boldsymbol{X} - \boldsymbol{K}^\dagger)\boldsymbol{\epsilon}\|^2_{\boldsymbol{K}}\right) \\
&\quad = \mathrm{tr}\left(\boldsymbol{K}(\boldsymbol{X} - \boldsymbol{K}^\dagger)\boldsymbol{Q}(\boldsymbol{X} - \boldsymbol{K}^\dagger)^\top\right). \qquad (27)
\end{aligned}$$

Consequently we have the subspace information criterion (SIC) [20, 19]:

$$\begin{aligned}
\mathrm{SIC} \ = \ &\|(\boldsymbol{X} - \boldsymbol{K}^\dagger)\boldsymbol{y}\|^2_{\boldsymbol{K}} \\
&-\mathrm{tr}\left(\boldsymbol{K}(\boldsymbol{X} - \boldsymbol{K}^\dagger)\boldsymbol{Q}(\boldsymbol{X} - \boldsymbol{K}^\dagger)^\top\right) \\
&+\mathrm{tr}\left(\boldsymbol{K}\boldsymbol{X}\boldsymbol{Q}\boldsymbol{X}^\top\right). \qquad (28)
\end{aligned}$$

The name subspace information criterion (SIC) came from the fact that it was first introduced for selecting subspace models. The first two terms in SIC are estimates of the bias term in Eq.(16) and the last term corresponds to the variance term in Eq.(16). It was shown that SIC is an unbiased estimator of J_1:

$$\mathrm{E}_{\boldsymbol{\epsilon}}\mathrm{SIC} = J_1. \tag{29}$$

4 Reducing Scatter of SIC Caused by Numerical Instability

As shown in Eq.(29), SIC is an unbiased estimator of J_1 with finite samples. Although this is a useful property, the reference [20] pointed out that the scatter of SIC can be large because the calculation of the generalized inverse matrix $\boldsymbol{K}^\dagger$ is sometimes numerically unstable. In this section, we show how this problem can be avoided.

SIC (28) can be expressed by

$$\begin{aligned}\mathrm{SIC} &= \|\boldsymbol{X}\boldsymbol{y}\|_{\boldsymbol{K}}^2 - 2\langle \boldsymbol{K}\boldsymbol{X}\boldsymbol{y}, \boldsymbol{K}^\dagger\boldsymbol{y}\rangle + \|\boldsymbol{K}^\dagger\boldsymbol{y}\|_{\boldsymbol{K}}^2 \\ &\quad +2\mathrm{tr}\left(\boldsymbol{K}^\dagger\boldsymbol{K}\boldsymbol{X}\boldsymbol{Q}\right) - \mathrm{tr}\left(\boldsymbol{K}^\dagger\boldsymbol{Q}\right). \end{aligned} \tag{30}$$

Since the third and fifth terms are irrelevant to $\boldsymbol{X}$, we will ignore them. Let us denote SIC without the third and fifth terms by SIC_e (SIC essential):

$$\begin{aligned}\mathrm{SIC}_\mathrm{e} &= \boldsymbol{y}^\top\boldsymbol{X}^\top\boldsymbol{K}\boldsymbol{X}\boldsymbol{y} - 2\boldsymbol{y}^\top\boldsymbol{K}^\dagger\boldsymbol{K}\boldsymbol{X}\boldsymbol{y} \\ &\quad +2\mathrm{tr}\left(\boldsymbol{K}^\dagger\boldsymbol{K}\boldsymbol{X}\boldsymbol{Q}\right). \end{aligned} \tag{31}$$

Letting $\mathcal{R}(\cdot)$ be the range of a matrix, we have the following theorem.

Theorem 1 *If the learning matrix $\boldsymbol{X}$ satisfies*

$$\mathcal{R}(\boldsymbol{X}) \subset \mathcal{R}(\boldsymbol{K}), \tag{32}$$

SIC_e can be exactly calculated by

$$SIC_\mathrm{e} = \boldsymbol{y}^\top\boldsymbol{X}^\top\boldsymbol{K}\boldsymbol{X}\boldsymbol{y} - 2\boldsymbol{y}^\top\boldsymbol{X}\boldsymbol{y} + 2\mathrm{tr}\left(\boldsymbol{X}\boldsymbol{Q}\right). \tag{33}$$

A proof of Theorem 1 is given in A. Note that $\mathcal{R}(\boldsymbol{X})$ is the subspace to which the learned parameter vector belongs (see Eq.(5)), while $\mathcal{R}(\boldsymbol{K})$ is the subspace to which the noiseless sample value vector $\boldsymbol{z}$ (Eq.(17)) belongs [20].

As can be seen from the definition, SIC and SIC_e are essentially the same since their difference is only irrelevant terms. However, the above theorem asserts that SIC_e has an excellent property that it can be exactly calculated without $\boldsymbol{K}^\dagger$ when Eq.(32) holds. Therefore, the numerical instability caused by the calculation of generalized inverse can be completely avoided if Eq.(32) holds.

The reference [19] showed that when $\boldsymbol{K}$ is invertible, $\mathrm{SIC_e}$ can be calculated by Eq.(33). However, $\boldsymbol{K}^{\dagger}$ is still needed when $\boldsymbol{K}$ is not invertible. On the other hand, Eq.(32) always holds when $\boldsymbol{K}$ is invertible. Therefore, Theorem 1 can be regarded as a more general condition for exactly calculating $\mathrm{SIC_e}$ by Eq.(33).

Now we show an example of the learning method that satisfies Eq.(32). The regularization learning (with quadratic regularizers) determines the parameter vector $\boldsymbol{\alpha}$ such that the regularized training error is minimized, i.e.,

$$\hat{\boldsymbol{\alpha}} = \operatorname*{argmin}_{\boldsymbol{\alpha}} \left(\sum_{i=1}^{n} \left(\hat{f}(\boldsymbol{x}_i) - y_i \right)^2 + \lambda \boldsymbol{\alpha}^{\top} \boldsymbol{T} \boldsymbol{\alpha} \right), \tag{34}$$

where λ is a positive scalar and $\boldsymbol{T}$ is an n-dimensional symmetric positive semi-definite matrix[5]. λ is called the regularization parameter and $\boldsymbol{T}$ is called the regularization matrix. The parameter vector $\hat{\boldsymbol{\alpha}}$ that satisfies Eq.(34) is given by the following learning matrix $\boldsymbol{X}$:

$$\boldsymbol{X} = (\boldsymbol{K}^2 + \lambda \boldsymbol{T})^{\dagger} \boldsymbol{K}. \tag{35}$$

For the regularization learning, the following theorem holds.

Theorem 2 *Eq.(32) holds if the regularization matrix $\boldsymbol{T}$ satisfies either one of the following conditions.*

1. $\mathcal{R}(\boldsymbol{T}) + \mathcal{R}(\boldsymbol{K}) = \mathbb{R}^n$ *and* $\boldsymbol{TK} = \boldsymbol{KT}$,
2. $\mathcal{R}(\boldsymbol{T}) \perp \mathcal{R}(\boldsymbol{K})$.

A proof of Theorem 2 is given in B.

A practically useful form of Gaussian process regression [27] corresponds to the regularization learning with $\boldsymbol{T} = \boldsymbol{I}$. When $\boldsymbol{T} = \boldsymbol{I}$, the condition (1) in Theorem 2 is fulfilled for any $\boldsymbol{K}$. This means that $\mathrm{SIC_e}$ can always be exactly calculated by Eq.(33) irrespective of the choice of the training sample points $\{\boldsymbol{x}_i\}_{i=1}^{n}$ (see Eq.(12)).

5 Reducing Variance of SIC Caused by Unbiasedness

So far, we discussed how the large scatter of SIC caused by the numerical instability of generalized inverse can be avoided. In this section, we investigate another cause of degrading the precision of SIC based on the fact that SIC is an exact unbiased estimator.

[5] An n-dimensional matrix $\boldsymbol{X}$ is said to be positive semi-definite if $\langle \boldsymbol{X}\boldsymbol{y}, \boldsymbol{y} \rangle \geq 0$ for any element $\boldsymbol{y}$ in $\mathbb{R}^n$ (see e.g., [2]).

5.1 Expected Squared Error of SIC

Let us consider the expected squared error (ESE) between SIC and J_1:

$$\mathrm{ESE} = \mathrm{E}_{\boldsymbol{\epsilon}}(\mathrm{SIC} - J_1)^2. \tag{36}$$

The above ESE can be decomposed into the bias and variance terms [9, 11]:

$$\mathrm{ESE} = (\mathrm{E}_{\boldsymbol{\epsilon}}\mathrm{SIC} - J_1)^2 + \mathrm{E}_{\boldsymbol{\epsilon}}(\mathrm{SIC} - \mathrm{E}_{\boldsymbol{\epsilon}}\mathrm{SIC})^2. \tag{37}$$

Since SIC is proved to be an unbiased estimator of J_1 (see Eq.(29)), the above bias term $(\mathrm{E}_{\boldsymbol{\epsilon}}\mathrm{SIC} - J_1)^2$ is zero. However, the variance term $\mathrm{E}_{\boldsymbol{\epsilon}}(\mathrm{SIC} - \mathrm{E}_{\boldsymbol{\epsilon}}\mathrm{SIC})^2$ is not taken into account in the derivation of SIC so it can be large. As a result, ESE between SIC and J_1 is not necessarily small. This may be another, and essential reason for degrading the precision of SIC.

The reference [24] suggested replacing the unbiased estimate $\hat{\boldsymbol{\alpha}}^u$ introduced in the derivation of SIC by a properly regularized estimate. The simulation results given in that paper showed that the variance of SIC can be drastically reduced in exchange for slight increase in the bias of SIC. Although this idea would be highly effective, we should find a proper degree of regularization, which can be determined only heuristically for the present.

On the other hand, the reference [20] paid attention to the fact that the bias term in Eq.(16) is always non-negative. Based on the fact, that paper proposed modifying SIC such that the terms in SIC which correspond to the bias term in Eq.(16) are kept non-negative. The modified criterion is called the corrected SIC (cSIC):

$$\begin{aligned} \mathrm{cSIC} \;=\; & \max\Big[\, 0, \|(\boldsymbol{X} - \boldsymbol{K}^{\dagger})\boldsymbol{y}\|_{\boldsymbol{K}}^2 \\ & \qquad - \mathrm{tr}\left(\boldsymbol{K}(\boldsymbol{X} - \boldsymbol{K}^{\dagger})\boldsymbol{Q}(\boldsymbol{X} - \boldsymbol{K}^{\dagger})^{\top}\right)\Big] \\ & + \mathrm{tr}\left(\boldsymbol{K}\boldsymbol{X}\boldsymbol{Q}\boldsymbol{X}^{\top}\right). \end{aligned} \tag{38}$$

For cSIC, we have the following theorem.

Theorem 3 *It holds that*

$$|cSIC - J_1| \leq |SIC - J_1|. \tag{39}$$

A proof of Theorem 3 is given in C.

Squaring both sides of Eq.(39) and taking the expectation over the noise, we have

$$\mathrm{E}_{\boldsymbol{\epsilon}}(\mathrm{cSIC} - J_1)^2 \leq \mathrm{E}_{\boldsymbol{\epsilon}}(\mathrm{SIC} - J_1)^2. \tag{40}$$

This implies that cSIC is better than SIC in the ESE sense.

5.2 Toward Further Improving Precision

As can be seen above, several efforts have been made so far for reducing the variance of SIC. Especially, the idea of regularizing $\hat{\boldsymbol{\alpha}}^u$ was shown to be practically useful, and cSIC was theoretically proved to be better than the original SIC in the ESE sense. Here we show another feasible way to find better generalization error estimators.

J_1 defined by Eq.(9) can be decomposed as

$$J_1 = \mathrm{E}_{\boldsymbol{\epsilon}}\|\hat{f}\|_{\mathcal{H}}^2 - 2\mathrm{E}_{\boldsymbol{\epsilon}}\langle\hat{f}, f_{\mathcal{S}}\rangle_{\mathcal{H}} + \|f_{\mathcal{S}}\|_{\mathcal{H}}^2, \tag{41}$$

where the third term is irrelevant to $\hat{f}$. As shown in Eq.(29), SIC is an unbiased estimator of J_1. This implies that SIC also estimates the above irrelevant term $\|f_{\mathcal{S}}\|_{\mathcal{H}}^2$. Indeed, the third and fifth terms in SIC (30), i.e., $\|\boldsymbol{K}^\dagger\boldsymbol{y}\|_{\boldsymbol{K}}^2 - \mathrm{tr}(\boldsymbol{K}^\dagger\boldsymbol{Q})$, correspond to an estimate of $\|f_{\mathcal{S}}\|_{\mathcal{H}}^2$. Now let us denote the first two terms in Eq.(41) by J_2:

$$J_2 = \mathrm{E}_{\boldsymbol{\epsilon}}\|\hat{f}\|_{\mathcal{H}}^2 - 2\mathrm{E}_{\boldsymbol{\epsilon}}\langle\hat{f}, f_{\mathcal{S}}\rangle_{\mathcal{H}}. \tag{42}$$

Then it can be confirmed that $\mathrm{SIC_e}$ is an unbiased estimator of J_2 [19]:

$$\mathrm{E}_{\boldsymbol{\epsilon}}\mathrm{SIC_e} = J_2. \tag{43}$$

From the above fact, we conjecture that a feasible approach is reducing ESE against J_2, rather than reducing ESE against J_1.

5.3 Reducing ESE against J_2

Now we give an alternative criterion that is aimed at improving the precision of $\mathrm{SIC_e}$.

Let us denote the negative half of the second term in J_2 (42) by η:

$$\eta = \mathrm{E}_{\boldsymbol{\epsilon}}\langle\hat{f}, f_{\mathcal{S}}\rangle_{\mathcal{H}}. \tag{44}$$

As shown in the reference [19], η is expressed as

$$\eta = \langle\boldsymbol{K}^\dagger\boldsymbol{K}\boldsymbol{X}\boldsymbol{z}, \boldsymbol{z}\rangle, \tag{45}$$

where $\boldsymbol{z}$ is defined by Eq.(17). The right-hand side of Eq.(45) implies that if $\boldsymbol{K}^\dagger\boldsymbol{K}\boldsymbol{X}$ is positive semi-definite, η is always non-negative. This fact motivates us to modify $\mathrm{SIC_e}$ such that the terms in $\mathrm{SIC_e}$ that correspond to η are kept non-negative. We will call the modified criterion the corrected $\mathrm{SIC_e}$ ($\mathrm{cSIC_e}$) because the idea is similar to cSIC. The negative half of the second and third terms in $\mathrm{SIC_e}$ (31) correspond to an estimate of η, so let us denote them by $\hat{\eta}$:

$$\hat{\eta} = \boldsymbol{y}^\top\boldsymbol{K}^\dagger\boldsymbol{K}\boldsymbol{X}\boldsymbol{y} - \mathrm{tr}(\boldsymbol{K}^\dagger\boldsymbol{K}\boldsymbol{X}\boldsymbol{Q}). \tag{46}$$

Then $\mathrm{cSIC_e}$ is defined as follows.

$$\mathrm{cSIC_e} = \boldsymbol{y}^\top\boldsymbol{X}^\top\boldsymbol{K}\boldsymbol{X}\boldsymbol{y} - 2\max(0, \hat{\eta}). \tag{47}$$

Now we show an example of the learning method such that $\boldsymbol{K}^{\dagger}\boldsymbol{K}\boldsymbol{X}$ is positive semi-definite. Since $\boldsymbol{K}$ is symmetric and positive semi-definite, there exists a unique symmetric square root[6] $\boldsymbol{K}^{\frac{1}{2}}$. Then we have the following theorem.

Theorem 4 *$\boldsymbol{K}^{\dagger}\boldsymbol{K}\boldsymbol{X}$ is positive semi-definite if the regularization matrix $\boldsymbol{T}$ in Eq.(34) satisfies either one of the following conditions.*

1. $\mathcal{R}(\boldsymbol{T}) + \mathcal{R}(\boldsymbol{K}) = \mathbb{R}^n$ *and* $\boldsymbol{T}\boldsymbol{K}^{\frac{1}{2}} = \boldsymbol{K}^{\frac{1}{2}}\boldsymbol{T}$,
2. $\mathcal{R}(\boldsymbol{T}) \perp \mathcal{R}(\boldsymbol{K})$.

A proof of Theorem 4 is given in D.

Fortunately, Theorem 2 also holds when the regularization matrix $\boldsymbol{T}$ satisfies either one of the conditions in Theorem 4 (see D for detail). Therefore, when either one of the conditions in Theorem 4 holds, $\mathrm{cSIC_e}$ can be stably calculated by

$$\begin{aligned}\mathrm{cSIC_e} &= \boldsymbol{y}^{\top}\boldsymbol{X}^{\top}\boldsymbol{K}\boldsymbol{X}\boldsymbol{y} \\ &\quad -2\max\left(0, \boldsymbol{y}^{\top}\boldsymbol{X}\boldsymbol{y} - \mathrm{tr}\left(\boldsymbol{X}\boldsymbol{Q}\right)\right). \end{aligned} \tag{48}$$

Note that a practically useful form of Gaussian process regression [27] corresponds to the regularization learning with $\boldsymbol{T} = \boldsymbol{I}$. When $\boldsymbol{T} = \boldsymbol{I}$, the condition (1) in Theorem 4 is fulfilled. Therefore, for the regularization learning with $\boldsymbol{T} = \boldsymbol{I}$, $\mathrm{cSIC_e}$ can always be stably calculated by Eq.(48).

6 Simulations

In this section, simple computer simulations are performed to illustrate how the precision of SIC is improved.

6.1 Artificial Data Sets

Let the dimension d of the input vector be 1. We use the Gaussian RKHS with width $c = 1$:

$$K(x, x') = \exp\left(-\frac{(x-x')^2}{2c^2}\right). \tag{49}$$

Let the learning target function be

$$f(x) = \mathrm{sinc}\, x. \tag{50}$$

Note that the above sinc function is included in the Gaussian RKHS[7].

The sample points $\{x_i\}_{i=1}^{n}$ are independently drawn from the uniform distribution on $(-\pi, \pi)$. The sample values $\{y_i\}_{i=1}^{n}$ are created as $y_i = f(x_i) + \epsilon_i$, where the noise $\{\epsilon_i\}_{i=1}^{n}$

[6] When $\boldsymbol{K}$ is symmetric and positive semi-definite, it is expressed by $\boldsymbol{K} = \sum_{i=1}^{n}\beta_i\phi_i\phi_i^{\top}$, where ϕ_i is an eigenvector and β_i is an associated eigenvalue that is non-negative. Then a symmetric square root

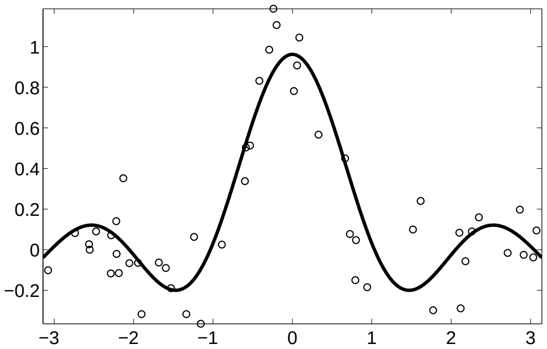


Figure 3: Target function and 50 training examples with noise variance $\sigma^2 = 0.04$.

is independently drawn from the normal distribution with mean zero and variance σ^2. We attempt the following 9 cases as the number n of training examples and the noise variance σ^2:

$$\begin{aligned}(n, \sigma^2) \;=\; & (100, 0.01), (50, 0.01), (25, 0.01),\\ & (100, 0.04), (50, 0.04), (25, 0.04),\\ & (100, 0.09), (50, 0.09), (25, 0.09), \end{aligned} \tag{51}$$

i.e., we investigate the cases with large/medium/small samples and low/medium/high noise levels. The learning target function and an example of the training set are illustrated in Figure 3.

We use the regularization learning (34) with $\boldsymbol{T} = \boldsymbol{I}$ for learning, which can also be regarded as a form of Gaussian process regression [27]. The generalization error prediction performance is investigated as a function of the regularization parameter λ. The following values of λ are attempted:

$$\{10^{-4}, 10^{-3.5}, 10^{-3}, \ldots, 10^{3}\}. \tag{52}$$

The generalization error is measured by an unexpected J_2 (42) over the noise, i.e.,

$$\text{Error} = \|\hat{f}\|_{\mathcal{H}}^2 - 2\langle \hat{f}, f_{\mathcal{S}} \rangle_{\mathcal{H}}. \tag{53}$$

We compare the following 3 criteria:

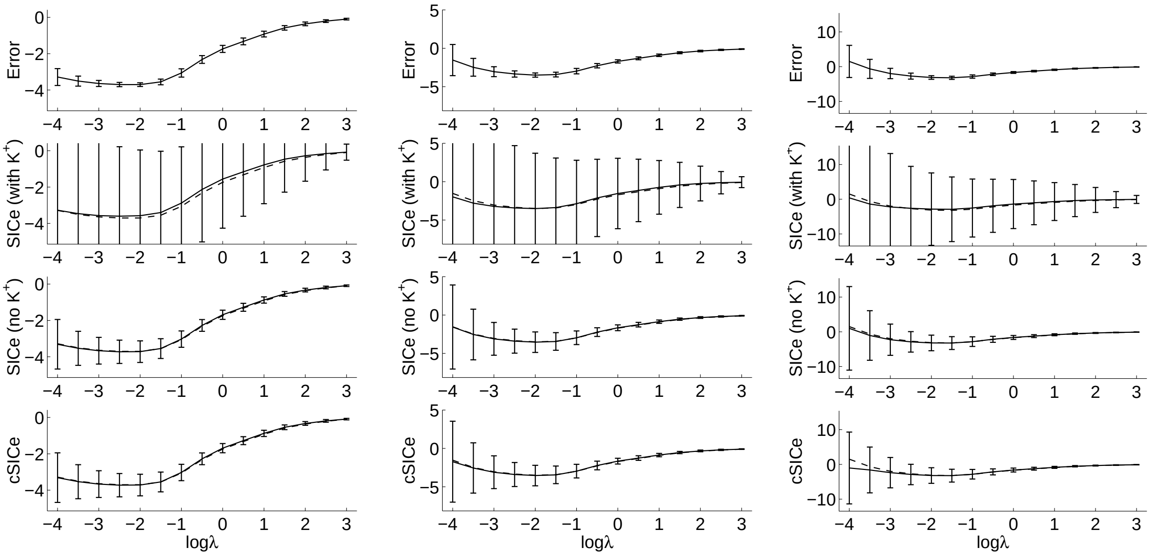
- SIC$_\text{e}$ (31), which includes $\boldsymbol{K}^\dagger$,

matrix is given by $\boldsymbol{K}^{\frac{1}{2}} = \sum_{i=1}^{n} \sqrt{\beta_i}\phi_i\phi_i^\top$.

[7] As described in the reference [10], the Gaussian RKHS is spanned by the function $f(x)$ that belongs to $L_2(\mathbb{R})$ and satisfies

$$\int_{-\infty}^{\infty} \frac{|\tilde{f}(\omega)|^2}{\tilde{k}(\omega)} d\omega < \infty,$$

where $\tilde{f}(\omega)$ is the Fourier transform of the function $f(x)$ and $\tilde{k}(\omega)$ is the Fourier transform of $\exp\left(-\frac{x^2}{2c^2}\right)$. The sinc function belongs to $L_2(\mathbb{R})$, and its Fourier transform is zero for $|\omega| > \pi$. Therefore, the above conditions are fulfilled so the sinc function is included in the Gaussian RKHS.



$(n, \sigma^2) = (50, 0.01)$ $(n, \sigma^2) = (50, 0.04)$ $(n, \sigma^2) = (50, 0.09)$

Figure 4: Means and standard errors of the error (53), $\mathrm{SIC_e}$ with $\boldsymbol{K}^\dagger$, $\mathrm{SIC_e}$ without $\boldsymbol{K}^\dagger$, and $\mathrm{cSIC_e}$ over 1000 trials. The mean values of the error (53) are also plotted in the bottom 3 graphs by the dashed line.

- $\mathrm{SIC_e}$ (33), which does not include $\boldsymbol{K}^\dagger$,
- $\mathrm{cSIC_e}$ (48).

The noise covariance matrix $\boldsymbol{Q}$ included in $\mathrm{SIC_e}$ and $\mathrm{cSIC_e}$ is estimated by Eqs.(21) and (22). The simulations are repeated 1000 times for each (n, σ^2) in Eq.(51), randomly drawing the sample points $\{x_i\}_{i=1}^n$ and noise $\{\epsilon_i\}_{i=1}^n$ from scratch in each trial. This means that we are practically taking the expectation over both the training sample points $\{x_i\}_{i=1}^n$ and the noise $\{\epsilon_i\}_{i=1}^n$. Note that $\mathrm{SIC_e}$ is still an unbiased estimator of J_2 even without taking expectation over the training sample points $\{x_i\}_{i=1}^n$.

The simulation results are depicted in Figure 4. Since the results are almost the same for different n, we only show the results for $n = 50$. In each block of the figure, four graphs are depicted. The horizontal axis of the graphs denotes the values of the regularization parameter λ in log-scale. The vertical axes denote, from top, the error (53), $\mathrm{SIC_e}$ (31) which includes $\boldsymbol{K}^\dagger$, $\mathrm{SIC_e}$ (33) which does not include $\boldsymbol{K}^\dagger$, and $\mathrm{cSIC_e}$ (48). The curves denote the mean values over 1000 trials, while the error bars denote the standard error over 1000 trials. For comparison, the mean values of the error (53) are also plotted in the bottom three graphs by the dashed line.

Since the condition in Theorem 1 is always fulfilled for the regularization learning (34) with $\boldsymbol{T} = \boldsymbol{I}$ (see Theorem 2), $\mathrm{SIC_e}$ (31) can be exactly calculated by $\mathrm{SIC_e}$ (33). Therefore, $\mathrm{SIC_e}$ (31) and $\mathrm{SIC_e}$ (33) should give the same value in principle. First we investigate the effect of avoiding $\boldsymbol{K}^\dagger$, that corresponds to the discussion in Section 4.

Table 2: Root mean squared error (54) of $\mathrm{SIC_e}$ or $\mathrm{cSIC_e}$.

(n, σ^2)	$(100, 0.01)$	$(50, 0.01)$	$(25, 0.01)$
$\mathrm{SIC_e}$ (33)	5.14×10^{-1}	5.68×10^{-1}	6.87×10^{-1}
$\mathrm{cSIC_e}$ (48)	5.14×10^{-1}	5.68×10^{-1}	6.87×10^{-1}
Improvement	0.00%	0.00%	0.00%

(n, σ^2)	$(100, 0.04)$	$(50, 0.04)$	$(25, 0.04)$
$\mathrm{SIC_e}$ (33)	1.58×10^{0}	1.87×10^{0}	1.95×10^{0}
$\mathrm{cSIC_e}$ (48)	1.57×10^{0}	1.83×10^{0}	1.85×10^{0}
Improvement	1.04%	2.30%	5.13%

(n, σ^2)	$(100, 0.09)$	$(50, 0.09)$	$(25, 0.09)$
$\mathrm{SIC_e}$ (33)	3.65×10^{0}	3.97×10^{0}	4.14×10^{0}
$\mathrm{cSIC_e}$ (48)	3.32×10^{0}	3.63×10^{0}	3.66×10^{0}
Improvement	9.07%	8.76%	11.6%

When $\sigma^2 = 0.01$, the mean values of $\mathrm{SIC_e}$ (31) roughly catch the profile of the mean error for all λ (see the left block in Figure 4). The mean values of $\mathrm{SIC_e}$ (33) also well agree with the mean error. This implies that, in practice, the mean values are not sensitive to the instability of generalized inverse, and estimating the noise covariance matrix $\boldsymbol{Q}$ by Eqs.(21) and (22) may be bearable. However, when we turn our attention to the scatter of $\mathrm{SIC_e}$, the error bars of $\mathrm{SIC_e}$ (31) are excessively large. On the other hand, the error bars of $\mathrm{SIC_e}$ (33) are much smaller than those of $\mathrm{SIC_e}$ (31), and are comparable to those of the true error (53). This means that the scatter of $\mathrm{SIC_e}$ is heavily affected by the instability of generalized inverse, and the large error bars can be effectively reduced by $\mathrm{SIC_e}$ (33) if the condition (32) holds.

When $\sigma^2 = 0.04$, the mean values of both $\mathrm{SIC_e}$ (31) and $\mathrm{SIC_e}$ (33) still well agree with the mean error (see the center block in Figure 4). Similar to the case with $\sigma^2 = 0.01$, $\mathrm{SIC_e}$ (31) has very large error bars, and it can be effectively reduced by $\mathrm{SIC_e}$ (33). However, the difference is that even with $\mathrm{SIC_e}$ (33), the error bars are rather large for small λ. The results for $\sigma^2 = 0.09$ are almost equivalent to the case with $\sigma^2 = 0.04$.

The above simulation results show that the mean of $\mathrm{SIC_e}$ is robust against the instability of the generalized inverse for any cases. On the other hand, the scatter of $\mathrm{SIC_e}$ is very sensitive to the instability of the generalized inverse and it can be effectively reduced by $\mathrm{SIC_e}$ if the condition (32) holds. However, even with this improvement, the scatter of $\mathrm{SIC_e}$ can still be large when the noise level is medium/high (i.e., $\sigma^2 = 0.04, 0.09$) and the regularization parameter λ is small.

Now we investigate the performance of $\mathrm{cSIC_e}$, i.e., we would like to investigate whether $\mathrm{cSIC_e}$ contributes to settling the above problem of $\mathrm{SIC_e}$ (33). The results are depicted in the bottom graph of each block in Figure 4. When the noise level is low ($\sigma^2 = 0.01$), cSIC$_{\mathrm{e}}$ seems to perform almost the same as SIC$_{\mathrm{e}}$ (33). In order to investigate the difference between SIC$_{\mathrm{e}}$ and cSIC$_{\mathrm{e}}$ in detail, we examine the root mean squared error of SIC$_{\mathrm{e}}$ or cSIC$_{\mathrm{e}}$ defined by

$$\sqrt{\mathrm{E}_\lambda\left[\mathrm{E}_t(\widehat{\mathrm{Error}} - \mathrm{E}_t\mathrm{Error})^2,\right]} \tag{54}$$

where E_λ denotes the expectation over all regularization parameter λ in Eq.(52), E_t denotes the expectation over 1000 trials, and $\widehat{\mathrm{Error}}$ denotes an estimator of the error, i.e., SIC$_{\mathrm{e}}$ (33) or cSIC$_{\mathrm{e}}$ (48). The values of the root mean squared error are described in Table 2, showing that they are the fairly equivalent. Therefore, cSIC$_{\mathrm{e}}$ works as well as SIC$_{\mathrm{e}}$ when $\sigma^2 = 0.01$.

On the other hand, when it comes to the medium noise cases ($\sigma^2 = 0.04$), the error bars of cSIC$_{\mathrm{e}}$ seem to be slightly smaller than those of SIC$_{\mathrm{e}}$ for $\lambda = 10^{-4}$ (see Figure 4). Indeed, the values of the root mean squared error described in Table 2 show that they are slightly reduced.

Finally, when the noise level is high ($\sigma^2 = 0.09$), the error bars of cSIC$_{\mathrm{e}}$ are again smaller than those of SIC$_{\mathrm{e}}$ for small λ (see Figure 4). Indeed, the values of the root mean squared error described in Table 2 show that they are significantly reduced.

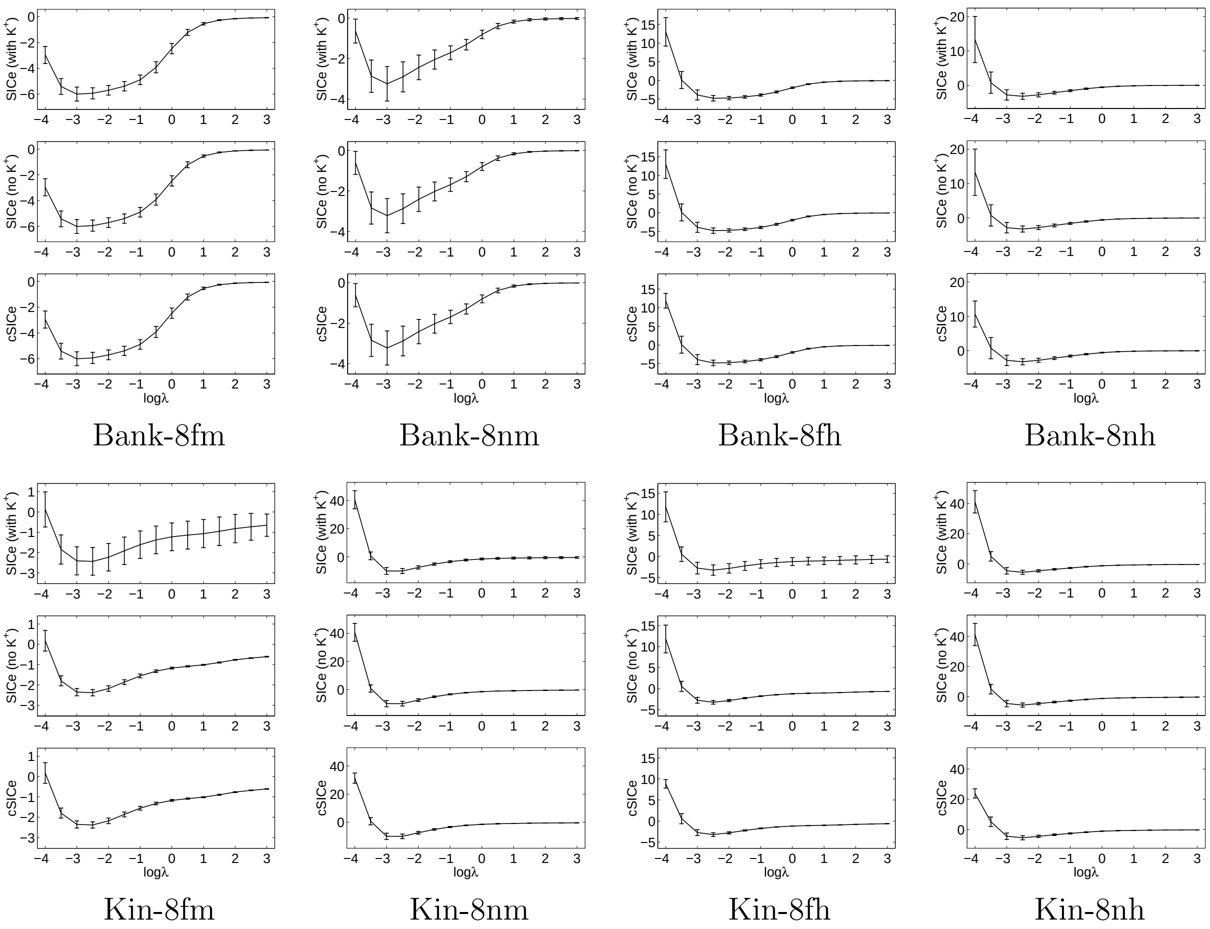
The above simulation results show that cSIC$_{\mathrm{e}}$ inherits the good performance of SIC$_{\mathrm{e}}$ when the noise level is low, and cSIC$_{\mathrm{e}}$ tends to improve the precision of SIC$_{\mathrm{e}}$ when the noise level is medium/high and the regularization parameter λ is small. Consequently, the root mean squared error of cSICe is improved for medium/high noise level cases. This means that the variance of SIC is effectively reduced in exchange for small increase in the bias.

6.2 DELVE Data Sets

Now we apply the proposed methods to practical data sets provided by DELVE [15]: *Bank-8fm, Bank-8nm, Bank-8fh, Bank-8nh, Kin-8fm, Kin-8nm, Kin-8fh*, and *Kin-8nh.* ‘f’ or ‘n’ signifies ‘fairly linear’ or ‘non-linear’, respectively, and ‘m’ or ‘h’ signifies ‘medium unpredictability/noise’ or ‘high unpredictability/noise’, respectively. Each of the 8 data sets includes 8192 samples, consisting of 8-dimensional input and 1-dimensional output data. We used 100 randomly selected samples for training, and the simulation is repeated 100 times. The noise covariance matrix is estimated by Eqs.(21) and (22) with $\lambda = 10^{-3}$. Note that for the above data sets, we can not calculate the true error corresponding to Eq.(53) since the learning target function is unknown.

The simulation results are depicted in Figure 5. The figure shows that SIC$_{\mathrm{e}}$ without $\boldsymbol{K}^\dagger$ gives significantly smaller error bars than SIC$_{\mathrm{e}}$ with $\boldsymbol{K}^\dagger$ for the Kin-8fm and Kin-8fh data sets, and they give almost the same results for other data sets. This implies that, even for the real data sets, SIC$_{\mathrm{e}}$ without $\boldsymbol{K}^\dagger$ sometimes contributes to reducing the scatter of SIC$_{\mathrm{e}}$ with $\boldsymbol{K}^\dagger$.

cSIC$_{\mathrm{e}}$ gives significantly different results from SIC$_{\mathrm{e}}$ for the Bank-8nh, Kin-8nm, Kin-8fh, and Kin-8nh data sets, and the difference appears for small λ. This tendency is similar to the previous simulations with artificial data sets. We can not judge whether cSIC$_{\mathrm{e}}$ is



Bank-8fm Bank-8nm Bank-8fh Bank-8nh

Kin-8fm Kin-8nm Kin-8fh Kin-8nh

Figure 5: Means and standard errors of $\mathrm{SIC_e}$ with $\boldsymbol{K}^{\dagger}$, $\mathrm{SIC_e}$ without $\boldsymbol{K}^{\dagger}$, and $\mathrm{cSIC_e}$ over 100 trials.

better than $\mathrm{SIC_e}$ or not because we can not calculate the true error in the current setting. However, given the fact that similar tendencies to the previous simulations appear, we expect that $\mathrm{cSIC_e}$ improves over $\mathrm{SIC_e}$.

7 Conclusions

The subspace information criterion (SIC) is an unbiased estimator of the generalization error J_1. In this paper, we discussed how the precision of SIC can be improved.

The scatter of SIC sometimes becomes large because the calculation of generalized inverse can be numerically unstable. To overcome this problem, we derived a condition for exactly calculating $\mathrm{SIC_e}$ (an essential part of SIC) without calculating generalized inverse. We showed that this condition is always fulfilled, for example, by a particular form of Gaussian process regression. The simple computer simulations illustrated that our effort surely contributed to reducing the scatter of $\mathrm{SIC_e}$.

However, the precision of $\mathrm{SIC_e}$ can still be degraded when the noise level is very high. This may be caused by the fact that $\mathrm{SIC_e}$ is an exact unbiased estimator of J_2 (an essential part of J_1). To cope with this problem, we gave a basic strategy that we should pursue an estimator that minimizes the expected squared error against J_2. Following this strategy, we proposed $\mathrm{cSIC_e}$, which is aimed at reducing the expected squared error against J_2. The simple computer simulations illustrated that $\mathrm{cSIC_e}$ works as well as $\mathrm{SIC_e}$ when the noise level is low, and it tends to be more precise than $\mathrm{SIC_e}$ as the noise level increases.

Our contribution surely improved the precision of SIC. However, still its precision can be degraded when the noise level is high and the regularization parameter λ is small. In the future, we will theoretically investigate the expected squared error and we would like to further improve the precision of the generalization error estimators.

Acknowledgements

The author would like to thank Prof. Dr. Hidemitsu Ogawa and Mr. Satoru Tanaka for their helpful comments. Special thanks also go to the anonymous reviewers for their comments, which greatly contributed to improving the earlier version of the manuscript. This research was partially supported by MEXT Grant-in-Aid for Scientific Research 14780262 and 14380158.

A Proof of Theorem 1

It is known that (see e.g., Exercises 3.13.1 in the book [2]) Eq.(32) is equivalent to

$$\boldsymbol{K}^{\dagger}\boldsymbol{K}\boldsymbol{X} = \boldsymbol{X}. \tag{55}$$

Therefore, Eq.(31) is reduced to Eq.(33). ■

B Proof of Theorem 2

First, we prove the condition (1). When $\boldsymbol{T}$ satisfies $\mathcal{R}(\boldsymbol{T}) + \mathcal{R}(\boldsymbol{K}) = \mathbb{R}^n$, $(\boldsymbol{K}^2 + \lambda\boldsymbol{T})$ is invertible. Then it holds that

$$(\boldsymbol{K}^2 + \lambda\boldsymbol{T})^{-1}\boldsymbol{K} = \boldsymbol{K}(\boldsymbol{K}^2 + \lambda\boldsymbol{T})^{-1}, \tag{56}$$

which can be confirmed by pre- and post-multiplying $(\boldsymbol{K}^2 + \lambda\boldsymbol{T})$ and using $\boldsymbol{T}\boldsymbol{K} = \boldsymbol{K}\boldsymbol{T}$. Then the learning matrix $\boldsymbol{X}$ given by Eq.(35) is expressed by

$$\boldsymbol{X} = \boldsymbol{K}(\boldsymbol{K}^2 + \lambda\boldsymbol{T})^{-1}, \tag{57}$$

which yields Eq.(32).

Now we prove the condition (2). When $\boldsymbol{T}$ satisfies the condition (2), it holds that

$$(\boldsymbol{K}^2 + \lambda\boldsymbol{T})^{\dagger}\boldsymbol{K} = \boldsymbol{K}^{\dagger}, \tag{58}$$

which can be confirmed by a standard matrix inversion technique (see e.g., Theorem 4.8 in the book [2]). Then the learning matrix $\boldsymbol{X}$ is expressed by

$$\boldsymbol{X} = \boldsymbol{K}^{\dagger}. \tag{59}$$

Recalling that $\mathcal{R}(\boldsymbol{K}^{\dagger}) = \mathcal{R}(\boldsymbol{K})$, we have Eq.(32). ■

C Proof of Theorem 3

Let b be the bias term in Eq.(16), i.e.,

$$b = \|\mathrm{E}_{\boldsymbol{\epsilon}}\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|^2_{\boldsymbol{K}}, \tag{60}$$

and let $\hat{b}$ be the terms in SIC (28) that correspond to b, i.e.,

$$\begin{aligned}\hat{b} \;=\;& \|(\boldsymbol{X} - \boldsymbol{K}^{\dagger})\boldsymbol{y}\|^2_{\boldsymbol{K}} \\ & -\mathrm{tr}\left(\boldsymbol{K}(\boldsymbol{X} - \boldsymbol{K}^{\dagger})\boldsymbol{Q}(\boldsymbol{X} - \boldsymbol{K}^{\dagger})^{\top}\right).\end{aligned} \tag{61}$$

Then it holds that

$$\mathrm{SIC} - J_1 \;=\; \hat{b} - b, \tag{62}$$

$$\mathrm{cSIC} - J_1 \;=\; \max(0, \hat{b}) - b. \tag{63}$$

If $\hat{b} < 0$, then $\max(0, \hat{b}) = 0$. Recalling $b \geq 0$, we have

$$\mathrm{cSIC} - J_1 = \max(0, \hat{b}) - b = -b \leq 0. \tag{64}$$

On the other hand, $\hat{b} < 0 = \max(0, \hat{b})$ so we have

$$\mathrm{SIC} - J_1 = \hat{b} - b < \max(0, \hat{b}) - b. \tag{65}$$

From Eqs.(64) and (65), we have

$$\mathrm{SIC} - J_1 < \mathrm{cSIC} - J_1 \leq 0. \tag{66}$$

Therefore, Eq.(39) holds with strict inequality.

If $\hat{b} \geq 0$, then $\hat{b} = \max(0, \hat{b})$. Therefore, Eqs.(62) and (63) assert that Eq.(39) holds with equality. ■

D Proof of Theorem 4

If $\boldsymbol{T}$ satisfies $\boldsymbol{T}\boldsymbol{K}^{\frac{1}{2}} = \boldsymbol{K}^{\frac{1}{2}}\boldsymbol{T}$, it holds that

$$\begin{aligned}\boldsymbol{T}\boldsymbol{K} \;&=\; \boldsymbol{T}\boldsymbol{K}^{\frac{1}{2}}\boldsymbol{K}^{\frac{1}{2}} = \boldsymbol{K}^{\frac{1}{2}}\boldsymbol{T}\boldsymbol{K}^{\frac{1}{2}} = \boldsymbol{K}^{\frac{1}{2}}\boldsymbol{K}^{\frac{1}{2}}\boldsymbol{T} \\ &=\; \boldsymbol{K}\boldsymbol{T},\end{aligned} \tag{67}$$

where $\boldsymbol{T}\boldsymbol{K}^{\frac{1}{2}} = \boldsymbol{K}^{\frac{1}{2}}\boldsymbol{T}$ is applied twice. This implies that if the condition (1) in Theorem 4 holds, the condition (1) in Theorem 2 also holds. Furthermore, the condition (2) in Theorem 4 and the condition (2) in Theorem 2 are equivalent. Therefore, if either one of the conditions in Theorem 4 holds, it follows from Theorem 2 that

$$\boldsymbol{K}^{\dagger}\boldsymbol{K}\boldsymbol{X} = \boldsymbol{X}. \tag{68}$$

For this reason, we shall prove that $\boldsymbol{X}$ is positive semi-definite if either one of the conditions in Theorem 4 holds.

First we prove the condition (1). If $\boldsymbol{T}$ satisfies $\mathcal{R}(\boldsymbol{T}) + \mathcal{R}(\boldsymbol{K}) = \mathbb{R}^n$, $(\boldsymbol{K}^2 + \lambda\boldsymbol{T})$ is invertible. Then it holds that

$$(\boldsymbol{K}^2 + \lambda\boldsymbol{T})^{-1}\boldsymbol{K}^{\frac{1}{2}} = \boldsymbol{K}^{\frac{1}{2}}(\boldsymbol{K}^2 + \lambda\boldsymbol{T})^{-1}, \tag{69}$$

which can be confirmed by pre- and post-multiplying $(\boldsymbol{K}^2+\lambda\boldsymbol{T})$ and using $\boldsymbol{T}\boldsymbol{K}^{\frac{1}{2}} = \boldsymbol{K}^{\frac{1}{2}}\boldsymbol{T}$. Since $(\boldsymbol{K}^2 + \lambda\boldsymbol{T})^{-1}$ is symmetric and positive semi-definite, $(\boldsymbol{K}^2 + \lambda\boldsymbol{T})^{-\frac{1}{2}}$ exists. Then learning matrix $\boldsymbol{X}$ given by Eq.(35) is expressed as

$$\begin{aligned} \boldsymbol{X} &= (\boldsymbol{K}^2 + \lambda\boldsymbol{T})^{-1}\boldsymbol{K}^{\frac{1}{2}}\boldsymbol{K}^{\frac{1}{2}} \\ &= \boldsymbol{K}^{\frac{1}{2}}(\boldsymbol{K}^2 + \lambda\boldsymbol{T})^{-1}\boldsymbol{K}^{\frac{1}{2}} \\ &= \boldsymbol{K}^{\frac{1}{2}}(\boldsymbol{K}^2 + \lambda\boldsymbol{T})^{-\frac{1}{2}}(\boldsymbol{K}^2 + \lambda\boldsymbol{T})^{-\frac{1}{2}}\boldsymbol{K}^{\frac{1}{2}}. \end{aligned} \tag{70}$$

Therefore, it holds that for any element $\boldsymbol{y}$ in $\mathbb{R}^n$,

$$\langle \boldsymbol{X}\boldsymbol{y}, \boldsymbol{y} \rangle = \|(\boldsymbol{K}^2 + \lambda\boldsymbol{T})^{-\frac{1}{2}}\boldsymbol{K}^{\frac{1}{2}}\boldsymbol{y}\|^2 \geq 0, \tag{71}$$

which shows that $\boldsymbol{X}$ is positive semi-definite.

Now we prove the condition (2). If $\boldsymbol{T}$ satisfies the condition (2), Eq.(35) is expressed by $\boldsymbol{X} = \boldsymbol{K}^{\dagger}$ (see the proof of Theorem 2). Then $\boldsymbol{X}$ is positive semi-definite since $\boldsymbol{K}^{\dagger}$ is positive semi-definite. ■

References

[1] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, AC-19(6):716–723, 1974.

[2] A. Albert. *Regression and the Moore-Penrose Pseudoinverse.* Academic Press, New York and London, 1972.

[3] S. Amari, N. Murata, K.-R. Müller, M. Finke, and H. H. Yang. Asymptotic statistical theory of overtraining and cross-validation. *IEEE Transactions on Neural Networks*, 8(5):985–996, 1997.

[4] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404, 1950.

[5] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods.* Cambridge University Press, Cambridge, 2000.

[6] I. Daubechies. *Ten Lectures on Wavelets.* Society for Industrial and Applied Mathematics, Philadelphia and Pennsylvania, 1992.

[7] D. L. Donoho. De-noising by soft thresholding. *IEEE Transactions on Information Theory*, 41(3):613–627, 1995.

[8] D. L. Donoho and I. M. Johnstone. Ideal spatial adaptation via wavelet shrinkage. *Biometrika*, 81:425–455, 1994.

[9] S. Geman, E. Bienenstock, and R. Doursat. Neural networks and the bias/variance dilemma. *Neural Computation*, 4(1):1–58, 1992.

[10] F. Girosi. An equivalence between sparse approximation and support vector machines. *Neural Computation*, 10(6):1455–1480, 1998.

[11] T. Heskes. Bias/variance decompositions for likelihood-based estimators. *Neural Computation*, 10(6):1425–1433, 1998.

[12] S. Konishi and G. Kitagawa. Generalized information criteria in model selection. *Biometrika*, 83:875–890, 1996.

[13] A. Luntz and V. Brailovsky. On estimation of characters obtained in statistical procedure of recognition. *Technicheskaya Kibernetica*, 3, 1969. in Russian.

[14] N. Murata, S. Yoshizawa, and S. Amari. Network information criterion — Determining the number of hidden units for an artificial neural network model. *IEEE Transactions on Neural Networks*, 5(6):865–872, 1994.

[15] C. E. Rasmussen, R. M. Neal, G. E. Hinton, D. van Camp, M. Revow, Z. Ghahramani, R. Kustra, and R. Tibshirani. The DELVE manual, 1996.

[16] B. Schölkopf and A. J. Smola. *Learning with Kernels.* MIT Press, Cambridge, MA, 2002.

[17] H. Shimodaira. An application of multiple comparison techniques to model selection. *Annals of Institute of Statistical Mathematics*, 50(1):1–13, 1998.

[18] M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, Series B*, 36:111–147, 1974.

[19] M. Sugiyama and K.-R. Müller. The subspace information criterion for infinite dimensional hypothesis spaces. *Journal of Machine Learning Research*, 3(Nov):323–359, 2002.

[20] M. Sugiyama and H. Ogawa. Subspace information criterion for model selection. *Neural Computation*, 13(8):1863–1889, 2001.

[21] M. Sugiyama and H. Ogawa. Theoretical and experimental evaluation of the subspace information criterion. *Machine Learning*, 48(1/2/3):25–50, 2002.

[22] K. Takeuchi. Distribution of information statistics and validity criteria of models. *Mathematical Science*, 153:12–18, 1976. in Japanese.

[23] A. Tanaka, H. Imai, and M. Miyakoshi. Choosing the parameter of image restoration filters by modified subspace information criterion. *IEICE Transactions on Fundamentals*, E85-A(5):1104–1110, 2002.

[24] K. Tsuda, M. Sugiyama, and K.-R. Müller. Subspace information criterion for non-quadratic regularizers — Model selection for sparse regressors. *IEEE Transactions on Neural Networks*, 13(1):70–80, 2002.

[25] V. N. Vapnik. *Statistical Learning Theory.* John Wiley & Sons, Inc., New York, 1998.

[26] H. Wahba. *Spline Model for Observational Data.* Society for Industrial and Applied Mathematics, Philadelphia and Pennsylvania, 1990.

[27] C. K. I. Williams. Prediction with Gaussian processes: From linear regression to linear prediction and beyond. In M. I. Jordan, editor, *Learning in Graphical Models*, pages 599–621. The MIT Press, Cambridge, 1998.