On the Influence of Input Noise on a Generalization Error Estimator

Masashi Sugiyama[1,2], Yuta Okabe[2], Hidemitsu Ogawa[2]

[1]Fraunhofer FIRST, Intelligent Data Analysis Group
Kekuléstr. 7, 12489, Berlin, Germany

[2]Department of Computer Science, Tokyo Institute of Technology
2-12-1, O-okayama, Meguro-ku, Tokyo, 152-8552, Japan

E-mail: sugi@cs.titech.ac.jp

ABSTRACT
Estimating the generalization capability is one of the most important problems in supervised learning. Therefore, various generalization error estimators have been proposed so far, in the presence of noise in output values. On the other hand, noise often exists in input values as well as output values. In this paper, we therefore investigate the influence of input noise on a generalization error estimator. We focus on a particular generalization error estimator called the subspace information criterion (SIC), which is shown to be unbiased in the absence of input noise. Intuitively, small input noise does not seem to affect the unbiasedness of SIC severely because small input noise varies the output values only slightly if the learning target function is continuous. On the contrary to this intuition, we show that even small input noise can totally corrupt the unbiasedness of SIC. This fact casts doubt on the use of SIC in the presence of input noise. To cope with this problem, we provide a sufficient condition to guarantee that SIC is unbiased in the limit of small input noise. We finally show that this condition is always fulfilled when the standard ridge estimation is used for learning, which allows us to use SIC without concern even in the presence of small input noise.

KEY WORDS
Machine Learning, Generalization Error, Input Noise, Measurement Noise, Perturbation, Model Selection

1 Introduction

Estimating an underlying function from training examples is the goal of supervised learning. The training examples consist of input points and corresponding output values, and they are often degraded by noise. Therefore, effectively suppressing the influence of noise in training examples is one of the keys to success in learning. To this end, several sophisticated theories of learning in the presence of noise in the output values (or labels) have been developed so far [18, 13, 3, 7]. On the other hand, there are cases where the noise is also included in the input values. For example, robot motor control, bioinformatics data analysis, and speech or image recognition, where input values as well as output values are measured. Time series prediction of multiple-step ahead can also be regarded as a case with input noise because estimated uncertain output values are recursively used as input values. In the statistics community, noise in the input values is called the measurement error and various methods for handling the measurement error have been explored [10, 4]. Also, in the field of neural information processing, a method for efficiently propagating the influence of uncertainty in time series prediction of multiple-step ahead has been proposed within the framework of Gaussian processes [11].

Estimating the generalization capability is one of the most important ingredients for successful learning because an accurate estimator of the generalization error can be used for model selection. Therefore, various generalization error estimators have been proposed so far, in the presence of output noise. However, it seems that generalization error estimation in the presence of input noise has not been well studied previously. In this paper, we therefore investigate how the accuracy of generalization error estimators can be influenced when input noise exists. More specifically, we focus on a particular generalization error estimator called the subspace information criterion (SIC) [17, 16], which is an unbiased estimator of a particular generalization error in the absence of input noise. In this paper, we investigate how the input noise influences the unbiasedness of SIC.

When the learning target function is continuous, small input noise varies the output values only slightly. Therefore, it intuitively seems that small input noise does not severely affect the unbiasedness of SIC. However, our interesting finding in this paper shows that this intuition is not always true. That is, the difference between the mean SIC and true generalization error does not always converge to zero in the limit of small input noise. Even worse, the difference between the mean SIC and true generalization error can go to infinity. This negative fact implies that simply using SIC in the presence of input noise is rather questionable. To cope with this problem, we investigate why such an extremely small input noise can totally corrupt the unbiasedness of SIC, and show how this problem can be overcome. More specifically, we show that under a mild condition on the learning method, the difference between the mean SIC and true generalization error always converges to zero as the size of input noise goes to zero,

which guarantees the robustness of SIC against small input noise. We finally show that a standard learning method such as the ridge estimation [12] satisfies this mild condition, which allows us to use SIC without concern even in the presence of small input noise.

2 Regression and Generalization Error Estimation

In this section, we formulate the regression problem of approximating a target function from training samples, and introduce an estimator of the generalization error called the subspace information criterion.

Let us denote the learning target function by $f(\boldsymbol{x})$, which is a real-valued function of d variables defined on the domain $\mathcal{D}$ (=$\mathbb{R}^d$). We are given a set of n samples called the *training examples*. A training example consists of a *sample point* $\boldsymbol{x}_i$ in $\mathcal{D}$ and a *sample value* y_i in $\mathbb{R}$. Sampling is actually carried out at $\boldsymbol{v}_i$ but we can not access to the true sample point $\boldsymbol{v}_i$. Instead, we have a noisy sample point $\boldsymbol{x}_i$ which is degraded by unknown additive noise $\boldsymbol{\xi}_i$. The sample value y_i also includes unknown additive noise ϵ_i. That is, the training examples are expressed as follows (see also Figure 1):

$$\{(\boldsymbol{x}_i, y_i) \mid \boldsymbol{x}_i = \boldsymbol{v}_i + \boldsymbol{\xi}_i, \;\; y_i = f(\boldsymbol{v}_i) + \epsilon_i\}_{i=1}^n. \quad (1)$$

In this paper, we treat $\{\epsilon_i\}_{i=1}^n$ as random variables, while we regard $\{\boldsymbol{\xi}_i\}_{i=1}^n$ as deterministic variables because we are interested in directly investigating the influence of the input noise $\{\boldsymbol{\xi}_i\}_{i=1}^n$. We assume that $\{\epsilon_i\}_{i=1}^n$ are drawn independently from a distribution with mean zero and variance σ^2.

Let us consider the cases where the unknown learning target function $f(\boldsymbol{x})$ belongs to a specified *reproducing kernel Hilbert space* (RKHS) $\mathcal{H}$. The *reproducing kernel* of a functional Hilbert space $\mathcal{H}$, denoted by $K(\boldsymbol{x}, \boldsymbol{x}')$, is a bivariate function defined on $\mathcal{D} \times \mathcal{D}$ that satisfies the following conditions [2, 19, 18, 5]:

- For any fixed $\boldsymbol{x}'$ in $\mathcal{D}$, $K(\boldsymbol{x}, \boldsymbol{x}')$ is a function of $\boldsymbol{x}$ in $\mathcal{H}$.
- For any function f in $\mathcal{H}$ and for any $\boldsymbol{x}'$ in $\mathcal{D}$, it holds that
$$\langle f(\cdot), K(\cdot, \boldsymbol{x}')\rangle_{\mathcal{H}} = f(\boldsymbol{x}'), \quad (2)$$
where $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ stands for the inner product in $\mathcal{H}$.

We will employ the following kernel regression model $\hat{f}(\boldsymbol{x})$ for learning:

$$\hat{f}(\boldsymbol{x}) = \sum_{i=1}^n \alpha_i K(\boldsymbol{x}, \boldsymbol{x}_i), \quad (3)$$

where $\{\alpha_i\}_{i=1}^n$ are parameters. We estimate the parameters by a linear estimation. More specifically, letting

$$\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^\top, \quad (4)$$
$$\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_n)^\top, \quad (5)$$

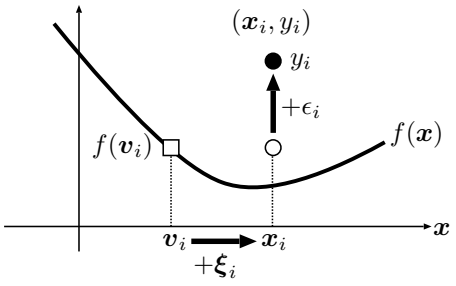

Figure 1. A training example is expressed by $(\boldsymbol{x}_i, y_i)$, where $\boldsymbol{x}_i = \boldsymbol{v}_i + \boldsymbol{\xi}_i$ and $y_i = f(\boldsymbol{v}_i) + \epsilon_i$. Sampling is carried out at $\boldsymbol{v}_i$ but we can not access to the true sample point $\boldsymbol{v}_i$. Instead, we have a noisy sample point $\boldsymbol{x}_i$ which is degraded by unknown additive noise $\boldsymbol{\xi}_i$. The sample value y_i also includes unknown additive noise ϵ_i. We will later consider the limit of small input noise $\boldsymbol{\xi}_i$.

where $^\top$ denotes the transpose of a vector (or a matrix) and $\{\hat{\alpha}_i\}_{i=1}^n$ are the estimated parameters, we estimate the parameters by

$$\hat{\boldsymbol{\alpha}} = \boldsymbol{X}\boldsymbol{y}, \quad (6)$$

where $\boldsymbol{X}$ is an n-dimensional matrix that does not depend on the output noise $\{\epsilon_i\}_{i=1}^n$. The matrix $\boldsymbol{X}$, which we call the *learning matrix*, can be any matrix but it is usually determined based on $\{\boldsymbol{x}_i\}_{i=1}^n$. A popular choice of $\boldsymbol{X}$ is the ridge estimation [12].

The purpose of regression is to obtain a good approximation $\hat{f}(\boldsymbol{x})$ to the unknown learning target function $f(\boldsymbol{x})$. For this purpose, we need a criterion that measures the *closeness* between two functions (i.e., the generalization measure). In this paper, we measure the generalization error by the expected squared norm in the RKHS $\mathcal{H}$.

$$\mathrm{E}_{\boldsymbol{\epsilon}}\|\hat{f} - f\|_{\mathcal{H}}^2, \quad (7)$$

where $\mathrm{E}_{\boldsymbol{\epsilon}}$ denotes the expectation over the output noise $\{\epsilon_i\}_{i=1}^n$, and $\|\cdot\|_{\mathcal{H}}$ denotes the norm in the RKHS $\mathcal{H}$. Using the function space norm as the error measure is rather common in the field of function approximation [6, 9, 8]. For further discussions on this generalization measure, readers may refer to [16]. For simplicity, we shall subtract a constant $\|f\|_{\mathcal{H}}^2$ from Eq.(7), and use the following J as the generalization measure.

$$\begin{aligned} J[\boldsymbol{X}] &= \mathrm{E}_{\boldsymbol{\epsilon}}\|\hat{f} - f\|_{\mathcal{H}}^2 - \|f\|_{\mathcal{H}}^2 \\ &= \mathrm{E}_{\boldsymbol{\epsilon}}\|\hat{f}\|_{\mathcal{H}}^2 - 2\mathrm{E}_{\boldsymbol{\epsilon}}\langle\hat{f}, f\rangle_{\mathcal{H}}, \end{aligned} \quad (8)$$

where $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ denotes the inner product in $\mathcal{H}$.

As can be seen from Eq.(8), J includes the unknown learning target function $f(\boldsymbol{x})$ so it can not be directly calculated. The subspace information criterion (SIC) [17, 16] is an estimator of the above generalization error J:

$$\begin{aligned} \mathrm{SIC}[\boldsymbol{X}] = \langle\boldsymbol{K}\boldsymbol{X}\boldsymbol{y}, \boldsymbol{X}\boldsymbol{y}\rangle - 2\langle\boldsymbol{K}\boldsymbol{X}\boldsymbol{y}, \boldsymbol{K}^\dagger\boldsymbol{y}\rangle \\ +2\sigma^2\mathrm{tr}(\boldsymbol{K}^\dagger\boldsymbol{K}\boldsymbol{X}), \end{aligned} \quad (9)$$

where † denotes the Moore-Penrose generalized inverse [1], $\mathrm{tr}(\cdot)$ denotes the trace of a matrix, and $\boldsymbol{K}$ is the so-called kernel matrix, i.e., the (i,j)-th element of $\boldsymbol{K}$ is given by

$$\boldsymbol{K}_{i,j} = K(\boldsymbol{x}_i, \boldsymbol{x}_j). \tag{10}$$

In the absence of input noise, SIC is shown to be an unbiased estimator of J for any learning matrix $\boldsymbol{X}$:

$$\mathrm{E}_{\boldsymbol{\epsilon}}\mathrm{SIC}[\boldsymbol{X}] = J[\boldsymbol{X}], \quad \text{when } \|\boldsymbol{\xi}_i\| = 0 \text{ for } i = 1, 2, \ldots, n. \tag{11}$$

The purpose of this paper is to investigate how this unbiasedness is influenced when input noise exists.

3 Influence of Small Input Noise on Unbiasedness of SIC

If the learning target function is continuous, it intuitively seems that small input noise does not affect the unbiasedness of SIC severely because small input noise varies the output values only slightly. In this section, we show that this intuition is not always true, and discuss how this problem can be overcome.

We first show the relation between the mean SIC and true generalization error J in the presence of input noise. Let $\boldsymbol{z}$ be a vector of sample values at the true sample points $\{\boldsymbol{v}_i\}_{i=1}^n$ and $\boldsymbol{z}_{\boldsymbol{x}}$ be a vector of sample values at the noisy sample points $\{\boldsymbol{x}_i\}_{i=1}^n$:

$$\boldsymbol{z} = (f(\boldsymbol{v}_1), f(\boldsymbol{v}_2), \ldots, f(\boldsymbol{v}_n))^\top, \tag{12}$$

$$\boldsymbol{z}_{\boldsymbol{x}} = (f(\boldsymbol{x}_1), f(\boldsymbol{x}_2), \ldots, f(\boldsymbol{x}_n))^\top. \tag{13}$$

Note that both $\boldsymbol{z}$ and $\boldsymbol{z}_{\boldsymbol{x}}$ are inaccessible because $f(\boldsymbol{x})$ and $\{\boldsymbol{v}_i\}_{i=1}^n$ are unknown. Then we have the following lemma.

Lemma 1 *In the presence of input noise $\{\boldsymbol{\xi}_i\}_{i=1}^n$, it holds that*

$$\mathrm{E}_{\boldsymbol{\epsilon}}\mathrm{SIC} = J + \Delta J, \tag{14}$$

where

$$\Delta J = 2\langle \boldsymbol{K}^\dagger \boldsymbol{K}\boldsymbol{X}\boldsymbol{z}, \boldsymbol{z}_{\boldsymbol{x}} - \boldsymbol{z}\rangle. \tag{15}$$

(Proof) Since $\{\epsilon_i\}_{i=1}^n$ are drawn independently from a distribution with mean zero and variance σ^2, we have

$$\begin{aligned}
\mathrm{E}_{\boldsymbol{\epsilon}}\mathrm{SIC} &= \mathrm{E}_{\boldsymbol{\epsilon}}\langle \boldsymbol{K}\boldsymbol{X}\boldsymbol{y}, \boldsymbol{X}\boldsymbol{y}\rangle - 2\mathrm{E}_{\boldsymbol{\epsilon}}\langle \boldsymbol{K}\boldsymbol{X}\boldsymbol{y}, \boldsymbol{K}^\dagger \boldsymbol{y}\rangle \\
&\quad + 2\sigma^2 \mathrm{tr}\left(\boldsymbol{K}^\dagger \boldsymbol{K}\boldsymbol{X}\right) \\
&= \langle \boldsymbol{K}\boldsymbol{X}\boldsymbol{z}, \boldsymbol{X}\boldsymbol{z}\rangle + \sigma^2 \mathrm{tr}\left(\boldsymbol{X}^\top \boldsymbol{K}\boldsymbol{X}\right) \\
&\quad - 2\langle \boldsymbol{K}\boldsymbol{X}\boldsymbol{z}, \boldsymbol{K}^\dagger \boldsymbol{z}\rangle - 2\sigma^2 \mathrm{tr}\left(\boldsymbol{K}^\dagger \boldsymbol{K}\boldsymbol{X}\right) \\
&\quad + 2\sigma^2 \mathrm{tr}\left(\boldsymbol{K}^\dagger \boldsymbol{K}\boldsymbol{X}\right) \\
&= \langle \boldsymbol{K}\boldsymbol{X}\boldsymbol{z}, \boldsymbol{X}\boldsymbol{z}\rangle + \sigma^2 \mathrm{tr}\left(\boldsymbol{X}^\top \boldsymbol{K}\boldsymbol{X}\right) \\
&\quad - 2\langle \boldsymbol{K}\boldsymbol{X}\boldsymbol{z}, \boldsymbol{K}^\dagger \boldsymbol{z}\rangle.
\end{aligned} \tag{16}$$

On the other hand, as shown in [16], J is expressed by

$$\begin{aligned}
J &= \langle \boldsymbol{K}\boldsymbol{X}\boldsymbol{z}, \boldsymbol{X}\boldsymbol{z}\rangle + \sigma^2 \mathrm{tr}(\boldsymbol{X}^\top \boldsymbol{K}\boldsymbol{X}) \\
&\quad - 2\langle \boldsymbol{K}\boldsymbol{X}\boldsymbol{z}, \boldsymbol{K}^\dagger \boldsymbol{z}_{\boldsymbol{x}}\rangle,
\end{aligned} \tag{17}$$

where only the third term is different from Eq.(16). Subtracting Eq.(17) from Eq.(16), we immediately have Eqs.(14) and (15). ■

Lemma 1 shows that, in the presence of input noise, SIC is generally no longer an unbiased estimator of J, but it is biased by ΔJ.

We are interested in investigating whether $|\Delta J|$ is small when the size of input noise is small. This may not be true for discontinuous learning target functions because $f(\boldsymbol{v}_i + \boldsymbol{\xi}_i)$ and $f(\boldsymbol{v}_i)$ can be totally different values even when $\|\boldsymbol{\xi}_i\|$ is small. So we focus on the cases where, roughly, the difference in the output values monotonically decreases as the input noise decreases. More specifically, for

$$\delta = \|\boldsymbol{z}_{\boldsymbol{x}} - \boldsymbol{z}\|, \tag{18}$$

we consider the cases where δ goes to zero as $\|\boldsymbol{\xi}_i\|$ goes to 0 for all $i = 1, 2, \ldots, n$. Under the above condition, we shall investigate the following question.

> Does $|\Delta J|$ converge to 0 as $\|\boldsymbol{\xi}_i\|$ goes to 0 for all $i = 1, 2, \ldots, n$?

If the answer is yes, then the unbiasedness of SIC is almost maintained even when small input noise exists. Therefore, we may use SIC without concern even in the presence of small input noise. Unfortunately, however, the following counterexample shows that this is not always true.

Example 2 *Let the input dimension d be 1, and let $\mathcal{H}$ be a Gaussian RKHS with reproducing kernel*

$$K(x, x') = \exp\left(-(x - x')^2/(2c^2)\right), \tag{19}$$

where we let $c = 1/\sqrt{2}$. Let the learning target function be

$$f(x) = \mathrm{sinc}\, x = \begin{cases} \sin \pi x/(\pi x) & \text{if } x \neq 0, \\ 1 & \text{if } x = 0, \end{cases} \tag{20}$$

which is included in the above Gaussian RKHS $\mathcal{H}$. Let $v_1 = v_2 = 0$, and let the learning matrix be

$$\boldsymbol{X} = \begin{pmatrix} (\mathrm{sinc}\, x_1 - 1)^2 & 0 \\ 0 & (\mathrm{sinc}\, x_2 - 1)^2 \end{pmatrix}^\dagger. \tag{21}$$

Then we have

$$\Delta J = 2(\mathrm{sinc}\, \xi_1 - 1)^{-1} + 2(\mathrm{sinc}\, \xi_2 - 1)^{-1} \quad \text{for } \xi_1 \neq 0 \text{ and } \xi_2 \neq 0. \tag{22}$$

This implies that $|\Delta J| \to \infty$ as $|\xi_1| \to 0$ and $|\xi_2| \to 0$.

Although the above example is fairly artificial, at least it clearly shows that there exists a case where $|\Delta J|$ does not converge to zero as the size of input noise goes to zero.

Even worse, $|\Delta J|$ goes to infinity in the above example. This fact casts doubt on the use of SIC in the presence of input noise.

On the contrary to this negative fact, the following theorem shows that this critical problem can be resolved by imposing a mild condition on the learning matrix $\boldsymbol{X}$.

Theorem 3 *Let $\|\boldsymbol{X}\|$ be the matrix norm defined by*

$$\|\boldsymbol{X}\| = \sup_{\boldsymbol{z}\neq 0} \frac{\|\boldsymbol{X}\boldsymbol{z}\|}{\|\boldsymbol{z}\|}. \tag{23}$$

If the learning matrix $\boldsymbol{X}$ satisfies

$$\|\boldsymbol{X}\| = o(1/\delta), \tag{24}$$

then $|\Delta J|$ converges to zero as $\|\boldsymbol{\xi}_i\|$ goes to 0 for all $i = 1, 2, \ldots, n$.

(Proof) From the Cauchy-Schwarz inequality, we have

$$|\Delta J| \leq 2\|\boldsymbol{K}^\dagger \boldsymbol{K}\boldsymbol{X}\boldsymbol{z}\| \cdot \|\boldsymbol{z}_{\boldsymbol{x}} - \boldsymbol{z}\| = 2\delta \|\boldsymbol{K}^\dagger \boldsymbol{K}\boldsymbol{X}\boldsymbol{z}\|. \tag{25}$$

On the other hand, it follows from Eq.(23) that for a bounded matrix $\boldsymbol{B}$

$$\|\boldsymbol{B}\boldsymbol{z}\| \leq \|\boldsymbol{B}\| \cdot \|\boldsymbol{z}\|. \tag{26}$$

Then we have

$$|\Delta J| \leq 2\delta \|\boldsymbol{K}^\dagger \boldsymbol{K}\| \cdot \|\boldsymbol{X}\| \cdot \|\boldsymbol{z}\|. \tag{27}$$

Since $\boldsymbol{K}^\dagger \boldsymbol{K}$ is an orthogonal projection matrix, $\|\boldsymbol{K}^\dagger \boldsymbol{K}\|$ is either 0 or 1. When $\|\boldsymbol{K}^\dagger \boldsymbol{K}\| = 0$, we have $|\Delta J| = 0$. When $\|\boldsymbol{K}^\dagger \boldsymbol{K}\| = 1$, we have

$$|\Delta J| \leq 2\delta \|\boldsymbol{X}\| \cdot \|\boldsymbol{z}\|. \tag{28}$$

Since $\|\boldsymbol{z}\|$ does not depend on $\{\boldsymbol{\xi}_i\}_{i=1}^n$, the upper bound $2\delta\|\boldsymbol{X}\| \cdot \|\boldsymbol{z}\|$ converges to zero as δ goes to zero if $\|\boldsymbol{X}\| = o(1/\delta)$. ■

Now we are interested in finding a learning matrix $\boldsymbol{X}$ that satisfies the above sufficient condition. Let us consider the ridge estimation [12], which determines $\boldsymbol{X}$ so that the regularized training error is minimized.

$$\min \left(\sum_{i=1}^n \left(\hat{f}(\boldsymbol{x}_i) - y_i \right)^2 + \lambda \sum_{j=1}^n \alpha_j^2 \right), \tag{29}$$

where λ is a positive scalar called the *ridge parameter*. A minimizer of Eq.(29) is given by the following learning matrix:

$$\boldsymbol{X} = (\boldsymbol{K}^2 + \lambda \boldsymbol{I})^{-1} \boldsymbol{K}, \tag{30}$$

where $\boldsymbol{I}$ denotes the identity matrix. For the above ridge estimation, we have the following theorem.

Theorem 4 *The learning matrix of the ridge estimation given by Eq.(30) satisfies Eq.(24).*

(Proof) Let $\{d_i\}_{i=1}^n$ be the eigenvalues of $\boldsymbol{K}$. Since the kernel matrix $\boldsymbol{K}$ is non-negative, $d_i \geq 0$ for all i. Let us diagonalize $\boldsymbol{K}$ by

$$\boldsymbol{K} = \boldsymbol{T}\boldsymbol{D}\boldsymbol{T}^\top, \tag{31}$$

where $\boldsymbol{T}$ is the orthogonal matrix and $\boldsymbol{D}$ is the diagonal matrix with diagonal elements $\{d_i\}_{i=1}^n$. Then Eq.(30) yields

$$\boldsymbol{X} = \boldsymbol{T}(\boldsymbol{D}^2 + \lambda \boldsymbol{I})^{-1} \boldsymbol{D}\boldsymbol{T}^\top. \tag{32}$$

This implies that the eigenvalues of $\boldsymbol{X}$ are given by $\{\frac{d_i}{d_i^2+\lambda}\}_{i=1}^n$, which are all non-negative. Then we have

$$\|\boldsymbol{X}\| = \max_i \frac{d_i}{d_i^2 + \lambda}. \tag{33}$$

Since $\frac{d}{d^2+\lambda} \leq \frac{1}{2\sqrt{\lambda}}$ for any $d \geq 0$, we have

$$\|\boldsymbol{X}\| \leq \frac{1}{2\sqrt{\lambda}} = \mathcal{O}(1) = o(1/\delta) \quad \text{as } \delta \to 0. \tag{34}$$

■

Theorem 4 means that for the ridge estimation, $|\Delta J|$ always converges to zero in the limit of small input noise. Therefore, we may use SIC without concern even in the presence of small input noise.

4 Computer Simulations

In this section, we experimentally investigate the influence of the input noise.

Let the dimension d of the input vector be 1. We use the Gaussian RKHS with width $c = 1$ (see Eq.(19)). We use Eq.(20) as the learning target function. Let the number n of training examples be 25. The noiseless sample points $\{v_i\}_{i=1}^n$ are independently drawn from the uniform distribution on $(-\pi, \pi)$. The input noise $\{\xi_i\}_{i=1}^n$ are independently drawn from the normal distribution with mean zero and standard deviation σ_x. The sample values $\{y_i\}_{i=1}^n$ are created as $y_i = f(v_i) + \epsilon_i$, where the output noise $\{\epsilon_i\}_{i=1}^n$ are independently drawn from the normal distribution with mean zero and standard deviation $\sigma = 0.05$. We consider the following three cases as the standard deviation σ_x of the input noise.

$$\sigma_x = 0, 0.1, 0.2. \tag{35}$$

Examples of the training set are depicted in Figure 2.

We use the kernel regression model (3), and the parameters $\{\alpha_i\}_{i=1}^n$ in the model are learned by ridge regression, i.e., the learning matrix is given by Eq.(30). The accuracy of SIC is investigated as a function of the ridge parameter λ, using the following values:

$$\lambda \in \{10^{-5}, 10^{-4.5}, 10^{-4}, \ldots, 10^{1}\}. \tag{36}$$

We estimate the noise variance σ^2 in SIC by

$$\hat{\sigma}^2 = \frac{\|\boldsymbol{K}\boldsymbol{X}\boldsymbol{y} - \boldsymbol{y}\|^2}{n - \mathrm{tr}\,(\boldsymbol{K}\boldsymbol{X})}. \tag{37}$$

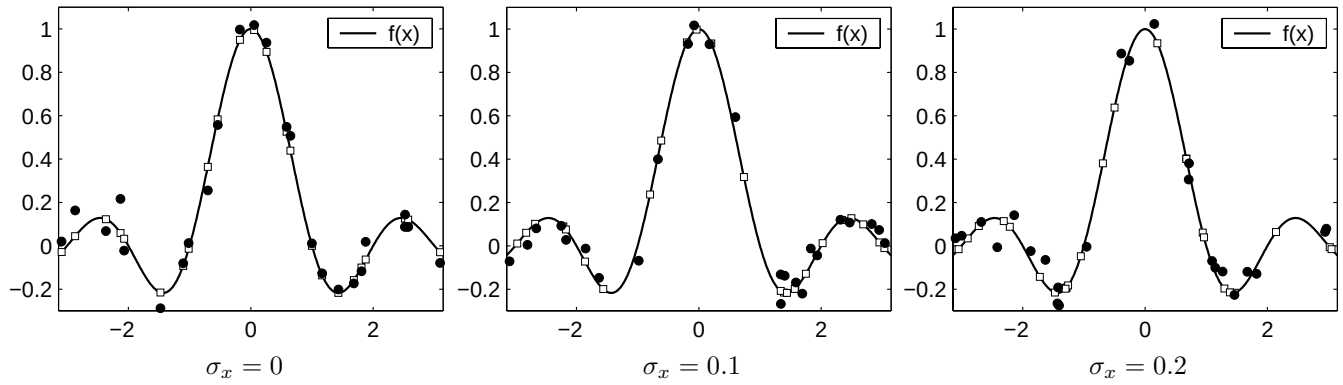

Figure 2. Learning target function $f(x)$ and 25 training examples. '□' denotes a noiseless training example $(v_i, f(v_i))$, while '•' denotes a noisy training example (x_i, y_i).

The simulations are repeated 100 times for each σ_x in Eq.(35), changing the noiseless sample points $\{v_i\}_{i=1}^n$, input noise $\{\xi_i\}_{i=1}^n$, and output noise $\{\epsilon_i\}_{i=1}^n$ in each trial. Note that in theory, we fixed the noiseless sample points $\{v_i\}_{i=1}^n$ and input noise $\{\xi_i\}_{i=1}^n$, and only changed the output noise $\{\epsilon_i\}_{i=1}^n$. However, in this experiment, we also change the noiseless sample points $\{v_i\}_{i=1}^n$ and input noise $\{\xi_i\}_{i=1}^n$ because we are interested in investigating the accuracy of SIC for various training sets. For this reason, we measure the generalization error by the following criterion in this experiment (cf. Eq.(8)):

$$\mathrm{Error}(\lambda) = \|\hat{f}_\lambda - f\|_{\mathcal{H}}^2 - \|f\|_{\mathcal{H}}^2, \tag{38}$$

where $\hat{f}_\lambda$ denotes the learned function with a ridge parameter λ.

Figure 3 displays the values of Error and SIC. The horizontal axis denotes the values of λ in log-scale. The mean is taken over 100 trials, and the error bar denotes the standard deviation over 100 trials. In order to clearly compare the mean curves, the mean Error is also drawn by the dashed line in the bottom graphs.

When $\sigma_x = 0$, the mean SIC approximates the mean Error very well. When $\sigma_x = 0.1$, the mean SIC slightly overestimates the mean Error. However, the difference is comparatively small, so SIC may be regarded as a reasonably accurate estimator even when the small input noise exists. Finally, when $\sigma_x = 0.2$, the mean SIC overestimates the mean Error and the difference is rather large.

5 Conclusions

We investigated the influence of input noise on a generalization error estimator called the subspace information criterion (SIC). Intuitively, small input noise does not seem to have serious effect on the accuracy of SIC if the learning target function is continuous. However, we constructed a counterexample showing that this intuition is not always true. This fact casts doubt on the use of SIC in the presence of input noise. For resolving this concern, we showed that if the learning method satisfies a mild condition, SIC is roughly robust against small input noise. We also showed that the standard ridge estimation satisfies this condition.

In experiments, we confirmed that SIC with ridge regression is still reasonably accurate even when the small input noise exists. However, as expected, SIC is no longer accurate in the presence of large input noise. An important future work is therefore to improve the accuracy of SIC in the presence of large input noise. Furthermore, the simulation results also showed that the variance of SIC tends to be large in the large input noise cases. Recently, methods to suppress the variance of SIC have been proposed [14, 15]. It would be interesting to see whether these or other schemes work for suppressing the variance of SIC even in the presence of input noise.

Acknowledgement

The authors would like to thank Gilles Blanchard and Motoaki Kawanabe for their comments. We also thank partial financial support from MEXT, Grants-in-Aid for Scientific Research, 14380158. A part of this work is also supported by the Alexander von Humboldt Foundation.

References

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

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

[3] C. M. Bishop. *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford, 1995.

[4] R. J. Carroll, D. Ruppert, and L. A. Stefanski. *Measurement Error in Nonlinear Models*. Chapman & Hall, London, 1995.

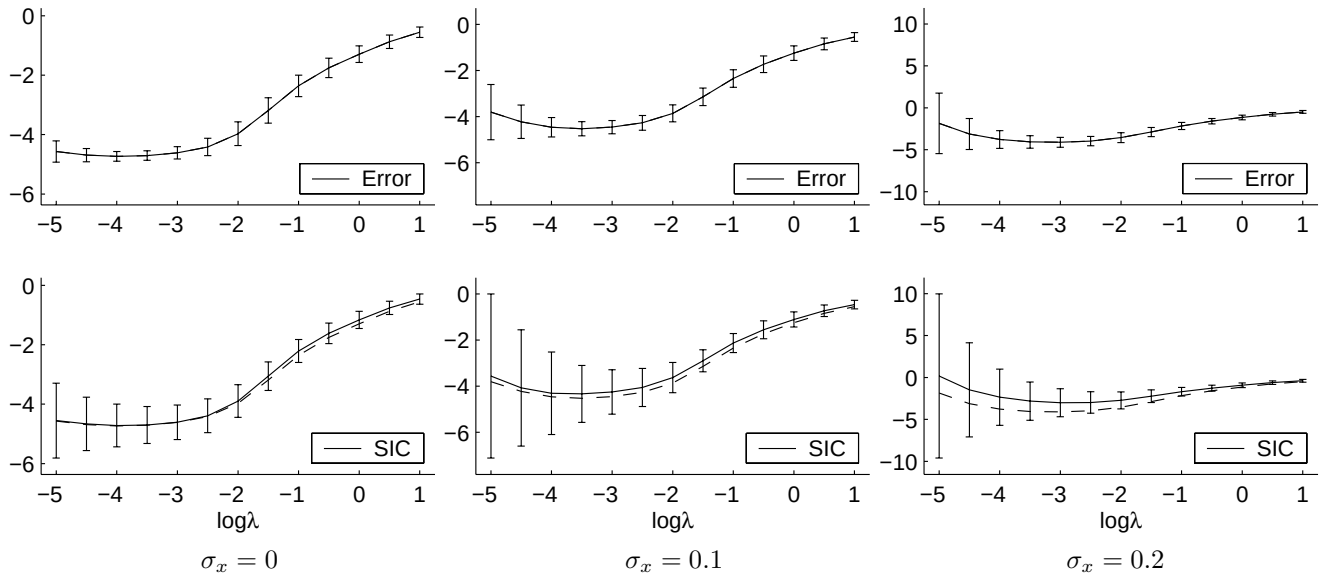

Figure 3. Values of Error and SIC. The horizontal axis denotes the value of the ridge parameter λ in log-scale. Dashed curves in the bottom graphs are the mean Error (same as the curves in the top graphs).

[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] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Number 31 in Applications of mathematics. Springer, New York, 1996.

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

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

[10] W. A. Fuller. *Measurement Error Models*. John Wiley & Sons, New York, 1987.

[11] A. C. Girard, C. E. Rasmussen, and R. Murray-Smith. Multiple-step ahead prediction for non linear dynamic systems — A Gaussian process treatment with propagation of the uncertainty. In *Advances in Neural Information Processing Systems 15*, Cambridge, MA, 2003. MIT Press.

[12] A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(3):55–67, 1970.

[13] 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.

[14] M. Sugiyama. Improving accuracy of subspace information criterion. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, E86-A(7):1885–1895, 2003.

[15] M. Sugiyama, M. Kawanabe, and K.-R. Müller. Trading variance reduction with unbiasedness — The regularized subspace information criterion for robust model selection in kernel regression. *Neural Computation*. to appear.

[16] 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.

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

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

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