Subspace Information Criterion for Model Selection

Masashi Sugiyama Hidemitsu Ogawa

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

Abstract

The problem of model selection is considerably important for acquiring higher levels of generalization capability in supervised learning. In this paper, we propose a new criterion for model selection called the subspace information criterion (SIC), which is a generalization of Mallows' C_L. It is assumed that the learning target function belongs to a specified functional Hilbert space and the generalization error is defined as the Hilbert space squared norm of the difference between the learning result function and target function. SIC gives an unbiased estimate of the generalization error so defined. SIC assumes the availability of an unbiased estimate of the target function and the noise covariance matrix, which are generally unknown. A practical calculation method of SIC for least mean squares learning is provided under the assumption that the dimension of the Hilbert space is less than the number of training examples. Finally, computer simulations in two examples show that SIC works well even when the number of training examples is small.

Keywords

Supervised learning, generalization capability, model selection, C_L, Akaike's information criterion (AIC).

1 Introduction

Supervised learning is obtaining an underlying rule from training examples made up of sample points and corresponding sample values. If the rule is successfully acquired, then appropriate output values corresponding to unknown input points can be estimated. This ability is called the *generalization capability*.

So far, many supervised learning methods have been developed, including the stochastic gradient descent method (Amari, 1967), the back-propagation algorithm (Rumelhart *et al.*, 1986a, 1986b), regularization learning (Tikhonov & Arsenin, 1977; Poggio & Girosi, 1990), Bayesian inference (Savage, 1954; MacKay, 1992), projection learning (Ogawa, 1987), and support vector machines (Vapnik, 1995; Schölkopf *et al.*, 1998). In these learning methods, the quality of the learning results depends heavily on the choice of *models*. Here, models refer to, for example, Hilbert spaces to which the learning target function belongs in the projection learning and support vector regression cases, multi-layer perceptrons (MLPs) with different numbers of hidden units in the back-propagation case, pairs of the regularization term and regularization parameter in the regularization learning case, and families of probabilistic distributions in the Bayesian inference case.

If the model is too complicated, then learning results tend to over-fit noisy training examples. In contrast, if the model is too simple, then it is not capable of fitting training examples causing learning results become under-fitted. In general, both over- and under-fitted learning results have lower levels of generalization capability. Therefore, the problem of finding an appropriate model, referred to as *model selection*, is considerably important for acquiring higher levels of generalization capability.

The problem of model selection has been studied mainly in the field of statistics. Mallows (1964) proposed C_P for the selection of subset-regression models (see also Gorman & Toman, 1966; Mallows, 1973). C_P gives an unbiased estimate of the predictive training error, i.e., the error between estimated and true values at sample points contained in the training set. Mallows (1973) extended the range of application of C_P to the selection of arbitrary linear regression models. It is called C_L or the unbiased risk estimate (Wahba, 1990). C_L may require a good estimate of the noise variance. In contrast, the generalized cross-validation (Craven & Wahba, 1979; Wahba, 1990), which is an extension of the traditional cross-validation (Mosteller & Wallace, 1963; Allen, 1974; Stone, 1974; Wahba, 1990), is the criterion for finding the model minimizing the predictive training error without the knowledge of noise variance. Li (1986) showed the asymptotic optimality of C_L and the generalized cross-validation, i.e., they asymptotically select the model minimizing the predictive training error (see also Wahba, 1990). However, these methods do not explicitly evaluate the error for unknown input points.

In contrast, model selection methods explicitly evaluating the generalization error have been studied from various standpoints: information statistics (Akaike, 1974; Takeuchi, 1976; Konishi & Kitagawa, 1996), Bayesian statistics (Schwarz, 1978; Akaike, 1980; MacKay, 1992), stochastic complexity (Rissanen, 1978, 1987, 1996; Yamanishi, 1998), and structural risk minimization (Vapnik, 1995; Cherkassky *et al.*, 1999). Particularly, information-statistics-based methods have been extensively studied. Akaike's information

criterion (AIC) (Akaike, 1974) is one of the most eminent methods of this type. Many successful applications of AIC to real world problems have been reported (e.g. Bozdogan, 1994; Akaike & Kitagawa, 1994, 1995; Kitagawa & Gersch, 1996). AIC assumes that models are *faithful*[1]. Takeuchi (1976) extended AIC to be applicable to unfaithful models. This criterion is called Takeuchi's modification of AIC (TIC) (see also Stone, 1977; Shibata, 1989). The learning method with which TIC can deal is restricted to the maximum likelihood estimation. Konishi and Kitagawa (1996) relaxed the restriction and derived the generalized information criterion for a class of learning methods represented by statistical functionals.

The common characteristic of AIC and its derivatives described above is to give an asymptotic unbiased estimate of the expected log-likelihood. This implies that when the number of training examples is small, these criteria are no longer valid. To overcome this weakness, two approaches have been taken. One is to calculate an exact unbiased estimate of the expected log-likelihood for each model. This type of modification can be found in many articles (e.g. Sugiura, 1978; Hurvich & Tsai, 1989, 1991, 1993; Noda *et al.*, 1996; Fujikoshi & Satoh, 1997; Satoh *et al.*, 1997; Hurvich *et al.*, 1998; Simonoff, 1998; McQuarrie & Tsai, 1998). The other approach is to use the bootstrap method (Efron, 1979; Efron & Tibshirani, 1993) for numerically evaluating the bias when the expected log-likelihood is estimated by the log-likelihood. The idea of the bootstrap bias correction is first introduced by Wong (1983) and Efron (1986), and then it is formalized as a model selection criterion by Ishiguro *et al.* (1997) (see also Davison & Hinkley, 1992; Cavanaugh & Shumway, 1997; Shibata, 1997).

In the neural network community, AIC has been extended to a different direction. Murata *et al.* (1994) generalized the loss function of TIC, and proposed the network information criterion (NIC). NIC assumes that the quasi-optimal estimator minimizing the empirical error, say the maximum likelihood estimator when the log loss is adopted as the loss function, has been exactly obtained. However, when we are concerned with MLP learning, it is difficult to obtain the quasi-optimal estimator in real time since MLP learning is generally performed by iterative methods such as the stochastic gradient descent method (Amari, 1967) and the back-propagation algorithm (Rumelhart *et al.*, 1986a, 1986b). To cope with this problem, information criteria taking the discrepancy between the quasi-optimal estimator and the obtained estimator into account have been devised (Wada & Kawato, 1991; Onoda, 1995).

In this paper, we propose a new criterion for model selection from the functional analytic viewpoint. We call this criterion the *subspace information criterion* (SIC). SIC is mainly different from AIC-type criteria in three respects. The first is the generalization measure. In AIC-type criteria, the averaged generalization error over all training sets is adopted as the generalization measure, and the averaged terms are replaced with particular values calculated by one given training set. In contrast, the generalization measure adopted in SIC is not averaged over training sets. The fact that the replacement of the averaged terms is unnecessary for SIC is expected to result in better selection

[1] A model is said to be *faithful* if the learning target can be expressed by the model (Murata *et al.*, 1994).

than AIC-type methods. The second is the approximation method. AIC-type criteria use asymptotic approximation and give an asymptotic unbiased estimate of the generalization error. In contrast, SIC uses the noise characteristics and gives an exact unbiased estimate of the generalization error. Our computer simulations show that SIC works well even when the number of training examples is small. The third is the restriction of models. Takeuchi (1983) pointed out that AIC-type criteria are effective only in the selection of nested models (see also Murata *et al.*, 1994). In SIC, no restriction is imposed on models.

This paper is organized as follows. Section 2 formulates the problem of model selection. In Section 3, our main result, SIC is derived. In Section 4, SIC is compared with Mallows' C_L and AIC-type criteria. In Section 5, SIC is applied to the selection of least mean squares learning models and a complete algorithm of SIC is described. Finally, Section 6 is devoted to computer simulations demonstrating the effectiveness of SIC.

2 Mathematical foundation of model selection

Let us consider the supervised learning problem of obtaining an approximation to a target function from a set of *training examples*. Let the learning target function be $f(x)$ of L variables defined on a subset $\mathcal{D}$ of the L-dimensional Euclidean space $\mathbf{R}^L$. The training examples are made up of *sample points* x_m in $\mathcal{D}$ and corresponding *sample values* y_m in $\mathbf{C}$:

$$\{(x_m, y_m) \mid y_m = f(x_m) + \epsilon_m\}_{m=1}^M, \tag{1}$$

where y_m is degraded by additive noise ϵ_m. Let θ be a set of factors determining learning results, e.g., the type and number of basis functions, and parameters in learning algorithms. We call θ a *model*. Let $\hat{f}_\theta$ be a learning result obtained with a model θ. Assuming that f and $\hat{f}_\theta$ belong to a Hilbert space H, the problem of model selection is described as follows.

Definition 1 (Model selection) *From a given set of models, find the model minimizing the generalization error defined as*

$$E_\epsilon \|\hat{f}_\theta - f\|^2, \tag{2}$$

where E_ϵ denotes the ensemble average over the noise, and $\|\cdot\|$ denotes the norm.

3 Subspace information criterion

In this section, we derive a model selection criterion named the *subspace information criterion* (SIC). SIC gives an unbiased estimate of the generalization error.

Let y, z, and ϵ be M-dimensional vectors whose m-th elements are y_m, $f(x_m)$, and ϵ_m, respectively:

$$y = z + \epsilon. \tag{3}$$

y and z are called a *sample value vector* and an *ideal sample value vector*, respectively. Let X_θ be a mapping from y to $\hat{f}_\theta$:

$$\hat{f}_\theta = X_\theta y. \tag{4}$$

X_θ is called a *learning operator*.

In the derivation of SIC, we assume the following conditions.

1. The learning operator X_θ is linear.
2. The mean noise is zero:
$$E_\epsilon \epsilon = 0. \tag{5}$$
3. A linear operator X_u which gives an unbiased learning result $\hat{f}_u$ is available:
$$E_\epsilon \hat{f}_u = f, \tag{6}$$
where
$$\hat{f}_u = X_u y. \tag{7}$$

Assumption 1 implies that the range of X_θ becomes a subspace of H. Linear learning operators include various learning methods such as least mean squares learning (Ogawa, 1992), regularization learning (Nakashima & Ogawa, 1999), projection learning (Ogawa, 1987), and parametric projection learning (Oja & Ogawa, 1986). It follows from Eqs.(7), (3), and (5) that

$$E_\epsilon \hat{f}_u = E_\epsilon X_u y = X_u z + E_\epsilon X_u \epsilon = X_u z. \tag{8}$$

Hence, Assumption 3 yields

$$X_u z = f. \tag{9}$$

Note that as discussed in Section 5, Assumption 3 holds if the dimension of the Hilbert space H is not larger than the number M of training examples.

Based on the above setting, we shall first give an estimation method of the generalization error of $\hat{f}_\theta$. The unbiased learning result $\hat{f}_u$ and the learning operator X_u are used for this purpose.

The generalization error of $\hat{f}_\theta$ is decomposed into the *bias* and *variance* (see e.g. Takemura, 1991; Geman *et al.*, 1992; Efron & Tibshirani, 1993):

$$E_\epsilon \|\hat{f}_\theta - f\|^2 = \|E_\epsilon \hat{f}_\theta - f\|^2 + E_\epsilon \|\hat{f}_\theta - E_\epsilon \hat{f}_\theta\|^2. \tag{10}$$

It follows from Eqs.(4) and (3) that Eq.(10) yields

$$\begin{aligned} E_\epsilon \|\hat{f}_\theta - f\|^2 &= \|X_\theta z - f\|^2 + E_\epsilon \|X_\theta \epsilon\|^2 \\ &= \|X_\theta z - f\|^2 + \mathrm{tr}\,(X_\theta Q X_\theta^*), \end{aligned} \tag{11}$$

where $\mathrm{tr}\,(\cdot)$ denotes the trace of an operator, Q is the noise covariance matrix, and X_θ^* denotes the adjoint operator of X_θ. Let X_0 be an operator defined as

$$X_0 = X_\theta - X_u. \tag{12}$$

Then the bias of $\hat{f}_\theta$ can be expressed by using $\hat{f}_u$ as

$$\begin{aligned}
\|X_\theta z - f\|^2 &= \|\hat{f}_\theta - \hat{f}_u\|^2 - \|\hat{f}_\theta - \hat{f}_u\|^2 + \|X_\theta z - f\|^2 \\
&= \|\hat{f}_\theta - \hat{f}_u\|^2 - \|X_\theta z + X_\theta \epsilon - (X_u z + X_u \epsilon)\|^2 + \|X_\theta z - X_u z\|^2 \\
&= \|\hat{f}_\theta - \hat{f}_u\|^2 - \|X_0 z + X_0 \epsilon\|^2 + \|X_0 z\|^2 \\
&= \|\hat{f}_\theta - \hat{f}_u\|^2 - \|X_0 z\|^2 - 2\mathrm{Re}\langle X_0 z, X_0 \epsilon\rangle - \|X_0 \epsilon\|^2 + \|X_0 z\|^2 \\
&= \|\hat{f}_\theta - \hat{f}_u\|^2 - 2\mathrm{Re}\langle X_0 z, X_0 \epsilon\rangle - \|X_0 \epsilon\|^2, \qquad (13)
\end{aligned}$$

where 'Re' stands for the real part of a complex number and $\langle \cdot, \cdot \rangle$ denotes the inner product. The second and third terms of the right-hand side of Eq.(13) can not be directly calculated since z and ϵ are unknown. Here, we shall replace them with the averages of them over the noise. Then the second term vanishes because of Eq.(5), and the third term yields

$$E_\epsilon \|X_0 \epsilon\|^2 = \mathrm{tr}\,(X_0 Q X_0^*)\,. \qquad (14)$$

This approximation immediately gives the following criterion.

Definition 2 (Subspace information criterion) *The following functional is called the subspace information criterion.*

$$\mathrm{SIC} = \|\hat{f}_\theta - \hat{f}_u\|^2 - \mathrm{tr}\,(X_0 Q X_0^*) + \mathrm{tr}\,(X_\theta Q X_\theta^*)\,. \qquad (15)$$

The model minimizing SIC is called the *minimum SIC model* (MSIC model) and the learning result obtained by the MSIC model is called the *MSIC learning result*. The generalization capability of the MSIC learning result measured by Eq.(2) is expected to be the best, the expectation is theoretically supported by the fact that SIC gives an unbiased estimate of the generalization error since it follows from Eqs.(15), (13), (5), (14), and (11) that

$$\begin{aligned}
E_\epsilon\ \mathrm{SIC} &= E_\epsilon \left(\|\hat{f}_\theta - \hat{f}_u\|^2 - \mathrm{tr}\,(X_0 Q X_0^*) + \mathrm{tr}\,(X_\theta Q X_\theta^*) \right) \\
&= E_\epsilon \left(\|X_\theta z - f\|^2 + 2\mathrm{Re}\langle X_0 z, X_0 \epsilon\rangle + \|X_0 \epsilon\|^2 - \mathrm{tr}\,(X_0 Q X_0^*) + \mathrm{tr}\,(X_\theta Q X_\theta^*) \right) \\
&= \|X_\theta z - f\|^2 + \mathrm{tr}\,(X_\theta Q X_\theta^*) \\
&= E_\epsilon \|\hat{f}_\theta - f\|^2. \qquad (16)
\end{aligned}$$

Since the bias is always non-negative from the definition, we can also consider the following corrected SIC (cSIC):

$$\mathrm{cSIC} = \left[\|\hat{f}_\theta - \hat{f}_u\|^2 - \mathrm{tr}\,(X_0 Q X_0^*) \right]_+ + \mathrm{tr}\,(X_\theta Q X_\theta^*)\,. \qquad (17)$$

where $[\cdot]_+$ is defined as

$$[t]_+ = \max(0, t). \qquad (18)$$

4 Discussion

In this section, SIC is compared with C_L and AIC-type criteria.

4.1 Comparison with C_L

1. The idea of estimation used in SIC generalizes C_L by Mallows (1973). The problem considered in Mallows' paper is to estimate the ideal sample value vector z from training examples $\{(x_m, y_m)\}_{m=1}^M$. The predictive training error is adopted as the error measure:
$$E_\epsilon \sum_{m=1}^{M} \left|\hat{f}_\theta(x_m) - f(x_m)\right|^2 = E_\epsilon \|B_\theta y - z\|^2, \tag{19}$$
where B_θ is a mapping from y to an M-dimensional vector whose m-th element is $\hat{f}_\theta(x_m)$. Analogous to Eqs.(10) and (11), the predictive training error can be decomposed into the bias and variance:
$$\begin{aligned} E_\epsilon \|B_\theta y - z\|^2 &= \|E_\epsilon B_\theta y - z\|^2 + E_\epsilon \|B_\theta y - E_\epsilon B_\theta y\|^2 \\ &= \|B_\theta z - z\|^2 + \mathrm{tr}\left(B_\theta Q B_\theta^*\right). \end{aligned} \tag{20}$$
Then C_L is given as
$$C_L = \|B_\theta y - y\|^2 - \mathrm{tr}\left((B_\theta - I)Q(B_\theta - I)^*\right) + \mathrm{tr}\left(B_\theta Q B_\theta^*\right). \tag{21}$$

2. Although the problem considered in Mallows' paper was to estimate z, acquiring a higher level of generalization capability is implicitly expected. However, minimizing the predictive training error does not generally mean to minimize the generalization error defined by Eq.(2). In contrast, we consider the problem of minimizing the generalization error and SIC directly gives an unbiased estimate of the generalization error.

3. Mallows employed the sample value vector y as an unbiased estimate of the target z. In contrast, we assumed the availability of the unbiased learning result $\hat{f}_u$ of the target function f. $\hat{f}_u$ plays a similar role to y in Mallows' case.

4.2 Comparison with AIC-type methods

1. In AIC-type criteria, the relation between the generalization error and the empirical error is first evaluated in the sense of the average over all training sets $\{x_m\}_{m=1}^M$. Then the averaged terms are replaced with particular values calculated by one given training set. In contrast, the generalization measure adopted in SIC (see Eq.(2)) is not averaged over training sets. The fact that the replacement of the averaged terms is unnecessary for SIC is expected to result in better selection than AIC-type methods.

2. AIC-type criteria give an asymptotic unbiased estimate of the generalization error. In contrast, SIC gives an exact unbiased estimate of the generalization error (see Eq.(16)). Therefore, SIC is expected to work well even when the number of training examples is small. Indeed, computer simulations performed in Section 6 support this claim.

3. Takeuchi (1983) pointed out that AIC-type criteria are effective only in the selection of nested models (see also Murata *et al.*, 1994):

$$S_1 \subset S_2 \subset \cdots. \tag{22}$$

   In SIC, no restriction is imposed on models except that the range of X_θ is included in H.

4. AIC-type methods compare models under the learning method minimizing the empirical error. In contrast, SIC can consistently compare models with different learning methods, e.g., least mean squares learning (Ogawa, 1992), regularization learning (Nakashima & Ogawa, 1999), projection learning (Ogawa, 1987), and parametric projection learning (Oja & Ogawa, 1986). Namely, the type of learning methods is also included in the model.

5. AIC-type criteria do not explicitly require a priori information on the class to which the target function belongs. In contrast, SIC requires a priori information on the Hilbert space H to which the target function f and a learning result $\hat{f}_\theta$ belong. If H is unknown, then a Hilbert space including all models is practically adopted as H (see the experiments in Section 6.2).

6. SIC requires the noise covariance matrix Q while AIC-type criteria do not. However, as shown in Section 5, we can cope with the case where Q is not available.

7. In the derivation of AIC-type methods, terms which are not dominant for model selection are neglected (see e.g. Murata *et al.*, 1994). This implies that the value of the AIC-type criteria is not an estimate of the generalization error itself. In contrast, SIC gives an estimate of the generalization error. This difference can be clearly seen in the top graph in Fig.4 (see Section 6.1).

8. AIC-type methods assume that training examples are independently and identically distributed (*i.i.d.*). In contrast, SIC can deal with the correlated noise if the noise covariance matrix Q is available.

5 SIC for least mean squares learning

In this section, SIC is applied to the selection of least mean squares (LMS) learning models.

LMS learning is to obtain a learning result $\hat{f}_\theta(x)$ in a subspace S of H minimizing the training error

$$\sum_{m=1}^{M} \left| \hat{f}_\theta(x_m) - y_m \right|^2 \tag{23}$$

from training examples $\{(x_m, y_m)\}_{m=1}^M$. In the LMS learning case, a model θ refers to a subspace S of H. Here, we assume that S has the reproducing kernel (see Aronszajn,

1950; Bergman, 1970; Wahba, 1990; Saitoh, 1988, 1997). Let $K_S(x, x')$ be the reproducing kernel of S, and $\mathcal{D}$ be the domain of functions in S. Then $K_S(x, x')$ satisfies the following conditions.

- For any fixed x' in $\mathcal{D}$, $K_S(x, x')$ is a function of x in S.
- For any function f in S and for any x' in $\mathcal{D}$, it holds that

$$\langle f(\cdot), K_S(\cdot, x') \rangle = f(x'). \tag{24}$$

Note that the reproducing kernel is unique if it exists. Let μ be the dimension of S. Then, for any orthonormal basis $\{\varphi_j\}_{j=1}^{\mu}$ in S, $K_S(x, x')$ is expressed as

$$K_S(x, x') = \sum_{j=1}^{\mu} \varphi_j(x) \overline{\varphi_j(x')}. \tag{25}$$

Let A_S be an operator from H to the M-dimensional unitary space $\mathbf{C}^M$ defined as

$$A_S = \sum_{m=1}^{M} \left(e_m \otimes \overline{K_S(x, x_m)} \right), \tag{26}$$

where $(\cdot \otimes \bar{\cdot})$ denotes the *Neumann-Schatten product*[2], and e_m is the m-th vector of the so-called standard basis in $\mathbf{C}^M$. A_S is called a *sampling operator* since it follows from Eq.(24) that

$$A_S f = z \tag{27}$$

for any f in S. Then LMS learning is rigorously defined as follows.

Definition 3 (Least mean squares learning) *(Ogawa, 1992) An operator X is called the LMS learning operator for the model θ if X minimizes the functional*

$$J[X] = \sum_{m=1}^{M} \left| \hat{f}_\theta(x_m) - y_m \right|^2 = \| A_S X y - y \|^2. \tag{28}$$

Let $A^\dagger$ be the *Moore-Penrose generalized inverse*[3] of A. Then the following proposition holds.

[2]For any fixed g in a Hilbert space H_1 and any fixed f in a Hilbert space H_2, the *Neumann-Schatten product* $(f \otimes \overline{g})$ is an operator from H_1 to H_2 defined by using any $h \in H_1$ as (Schatten, 1970)

$$(f \otimes \overline{g})\, h = \langle h, g \rangle f.$$

[3]An operator X is called the *Moore-Penrose generalized inverse* of an operator A if X satisfies the following four conditions (see Albert, 1972; Ben-Israel & Greville, 1974).

$$AXA = A, \;\; XAX = X, \;\; (AX)^* = AX, \;\; \text{and} \;\; (XA)^* = XA.$$

The Moore-Penrose generalized inverse is unique and denoted as $A^\dagger$.

Proposition 1 *(Ogawa, 1992) The LMS learning operator for the model θ is given by*

$$X_\theta = A_S^\dagger. \tag{29}$$

SIC requires an unbiased learning result $\hat{f}_u$ and an operator X_u providing $\hat{f}_u$. Here, we show a method of obtaining $\hat{f}_u$ and X_u by LMS learning. Let us assume that H is also a reproducing kernel Hilbert space and regard H as a model. Let A_H be a sampling operator defined with the reproducing kernel of H:

$$A_H = \sum_{m=1}^{M} \left(e_m \otimes \overline{K_H(x, x_m)} \right). \tag{30}$$

Since f belongs to H, it follows from Eq.(24) that the ideal sample value vector z is expressed as

$$z = A_H f. \tag{31}$$

Let $\hat{f}_H$ be a learning result obtained with the model H:

$$\hat{f}_H = X_H y, \tag{32}$$

where X_H is the LMS learning operator with the model H:

$$X_H = A_H^\dagger. \tag{33}$$

Now let us assume that the range of A_H^* agrees with H[4]. This holds only if the number M of training examples is larger than or equal to the dimension of H. Then it follows from Eqs.(32), (3), (5), (31), and (33) that

$$E_\epsilon \hat{f}_H = E_\epsilon X_H y = X_H z + E_\epsilon X_H \epsilon = X_H z = X_H A_H f = A_H^\dagger A_H f = f, \tag{34}$$

which implies that $\hat{f}_H$ is unbiased. Hence, we can put

$$\hat{f}_u = \hat{f}_H, \tag{35}$$

$$X_u = X_H. \tag{36}$$

For evaluating SIC, the noise covariance matrix Q is required (see Eq.(15)). One of the measures is to use

$$\hat{Q} = \hat{\sigma}^2 I \tag{37}$$

as an estimate of the noise covariance matrix, and $\hat{\sigma}^2$ is estimated from training examples $\{(x_m, y_m)\}_{m=1}^M$ as

$$\hat{\sigma}^2 = \frac{\sum_{m=1}^{M} \left| \hat{f}_u(x_m) - y_m \right|^2}{M - \dim(H)}, \tag{38}$$

[4]This condition is sufficient for $\hat{f}_H$ to be unbiased, not necessary.

where $M > \dim(H)$ is implicitly assumed. When Q is really in the form $Q = \sigma^2 I$ and σ^2 is estimated by Eq.(38), SIC still gives an unbiased estimate of the generalization error since Eq.(38) is an unbiased estimate of σ^2 (see Theorem 1.5.1 in Fedorov (1972)).

Based on the above discussions, we shall show a calculation method of LMS learning results and the value of SIC by matrix operations. Let T_S and T_H be M-dimensional matrices whose (m, m')-th elements are $K_S(x_m, x_{m'})$ and $K_H(x_m, x_{m'})$, respectively. Then the following theorem holds.

Theorem 1 (Calculation of LMS learning results) *LMS learning results $\hat{f}_\theta(x)$ and $\hat{f}_u(x)$ can be calculated as*

$$\hat{f}_\theta(x) = \sum_{m=1}^{M} \langle T_S^\dagger y, e_m \rangle K_S(x, x_m), \tag{39}$$

$$\hat{f}_u(x) = \sum_{m=1}^{M} \langle T_H^\dagger y, e_m \rangle K_H(x, x_m). \tag{40}$$

A proof of Theorem 1 is provided in Appendix A. Let T be an M-dimensional matrix defined as

$$T = T_S^\dagger - T_H^\dagger T_S T_S^\dagger - T_S^\dagger T_S T_H^\dagger + T_H^\dagger. \tag{41}$$

Then the following theorem holds.

Theorem 2 (Calculation of SIC for LMS learning) *When the noise covariance matrix Q is estimated by Eqs.(37) and (38), SIC for LMS learning is given as*

$$\mathrm{SIC} = \langle Ty, y \rangle - \hat{\sigma}^2 \mathrm{tr}\,(T) + \hat{\sigma}^2 \mathrm{tr}\left(T_S^\dagger\right), \tag{42}$$

where $\hat{\sigma}^2$ is given as

$$\hat{\sigma}^2 = \frac{\|y\|^2 - \langle T_H T_H^\dagger y, y \rangle}{M - \dim(H)}. \tag{43}$$

A proof of Theorem 2 is given in Appendix B. In practice, the calculation of the Moore-Penrose generalized inverse is sometimes unstable. To overcome the unstableness, we recommend using *Tikhonov's regularization* (Tikhonov & Arsenin, 1977):

$$T_S^\dagger \longleftarrow T_S(T_S^2 + \alpha I)^{-1}, \tag{44}$$

where α is a small constant, say $\alpha = 10^{-3}$. Then a complete algorithm of the MSIC procedure, i.e., finding the best model from $\{S_n\}_n$, is described in a pseudo code in Fig.1.

It is said that the dimension of the subspace S required for obtaining an approximation in a certain level of precision grows exponentially with the dimension L of the input space $\mathcal{D}$, a concept referred to as the *curse of dimensionality* (Bishop, 1995). This phenomenon generally results in large computational complexity, so that learning procedures are infeasible to compute in real time. However, thanks to good properties of the reproducing kernel, the computational complexity does not exponentially increase with the dimension of the input space if the reproducing kernel can be expressed in a closed form. Examples of such nice reproducing kernels are described in Girosi (1998), including polynomials, trigonometric polynomials, multi-layer perceptrons, radial basis function networks with fixed width, and B-splines.

---

input $\{(x_m, y_m)\}_{m=1}^M$, $K_H(x, x')$, and $\{K_{S_n}(x, x')\}_n$;
$\alpha \longleftarrow 10^{-3}$;
$T_H \longleftarrow [K_H(x_m, x_{m'})]_{mm'}$;
if $(M \leq \dim(H))$ or $(\mathrm{rank}\,(T_H) < \dim(H))$
    print('H is too complicated.');
    exit;
end
$T_H^\dagger \longleftarrow T_H(T_H^2 + \alpha I)^{-1}$;
$\hat{\sigma}^2 \longleftarrow (\|y\|^2 - \langle T_H T_H^\dagger y, y\rangle)/\,(M - \dim(H))$;
for all n
    $T_{S_n} \longleftarrow [K_{S_n}(x_m, x_{m'})]_{mm'}$;
    $T_{S_n}^\dagger \longleftarrow T_{S_n}(T_{S_n}^2 + \alpha I)^{-1}$;
    $T \longleftarrow T_{S_n}^\dagger - T_H^\dagger T_{S_n} T_{S_n}^\dagger - T_{S_n}^\dagger T_{S_n} T_H^\dagger + T_H^\dagger$;
    $\mathrm{SIC}_n \longleftarrow \langle Ty, y\rangle - \hat{\sigma}^2 \mathrm{tr}\,(T) + \hat{\sigma}^2 \mathrm{tr}\left(T_{S_n}^\dagger\right)$;
end
$\hat{n} \longleftarrow \mathrm{argmin}_n\{\mathrm{SIC}_n\}$;
$\hat{f}(x) \longleftarrow \Sigma_{m=1}^M \langle T_{S_{\hat{n}}}^\dagger y, e_m\rangle K_{S_{\hat{n}}}(x, x_m)$;

Figure 1: MSIC procedure in a pseudo code — Find the best model from $\{S_n\}_n$.

6 Computer simulations

In this section, computer simulations are performed to demonstrate the effectiveness of SIC compared with the network information criterion (NIC) (Murata *et al.*, 1994), which is a generalized AIC.

6.1 Illustrative example

Let the target function $f(x)$ be

$$\begin{aligned} f(x) &= \sqrt{2}\sin x + 2\sqrt{2}\cos x - \sqrt{2}\sin 2x - 2\sqrt{2}\cos 2x + \sqrt{2}\sin 3x - \sqrt{2}\cos 3x \\ &\quad + 2\sqrt{2}\sin 4x - \sqrt{2}\cos 4x + \sqrt{2}\sin 5x - \sqrt{2}\cos 5x, \end{aligned} \tag{45}$$

and training examples $\{(x_m, y_m)\}_{m=1}^M$ be

$$x_m = -\pi - \frac{\pi}{M} + \frac{2\pi m}{M}, \tag{46}$$

$$y_m = f(x_m) + \epsilon_m, \tag{47}$$

where the noise ϵ_m is independently subject to the normal distribution with mean 0 and variance 3:

$$\epsilon_m \sim N(0, 3). \tag{48}$$

Let us consider the following models:

$$\{S_n\}_{n=1}^{20} \tag{49}$$

where S_n is a trigonometric polynomial space of order n, i.e., a Hilbert space spanned by

$$\{1, \sin px, \cos px\}_{p=1}^n, \tag{50}$$

and the inner product is defined as

$$\langle f, g \rangle_{S_n} = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x)\overline{g(x)}dx. \tag{51}$$

The reproducing kernel of S_n is expressed as

$$K_{S_n}(x, x') = \begin{cases} \sin\dfrac{(2n+1)(x-x')}{2} \Big/ \sin\dfrac{x-x'}{2} & \text{if } x \neq x', \\ 2n+1 & \text{if } x = x'. \end{cases} \tag{52}$$

Our task is to find the best model S_n minimizing

$$\text{Error} = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left| \hat{f}_{S_n}(x) - f(x) \right|^2 dx, \tag{53}$$

where $\hat{f}_{S_n}(x)$ is the learning result obtained with the model S_n. Let us consider the following two model selection methods.

(A) SIC: We use cSIC given by Eq.(17). LMS learning is adopted. The largest model S_{20} including all models is employed as H. The unbiased learning result $\hat{f}_u$ and the estimate $\hat{\sigma}^2$ of the noise variance are obtained by Eqs.(35) and (38), respectively.

(B) NIC: The squared loss is adopted as the loss function. In this case, learning results obtained by the stochastic gradient descent method (Amari, 1967) converge to the LMS estimator. The distribution of sample points given by Eq.(46) is regarded as a uniform distribution.

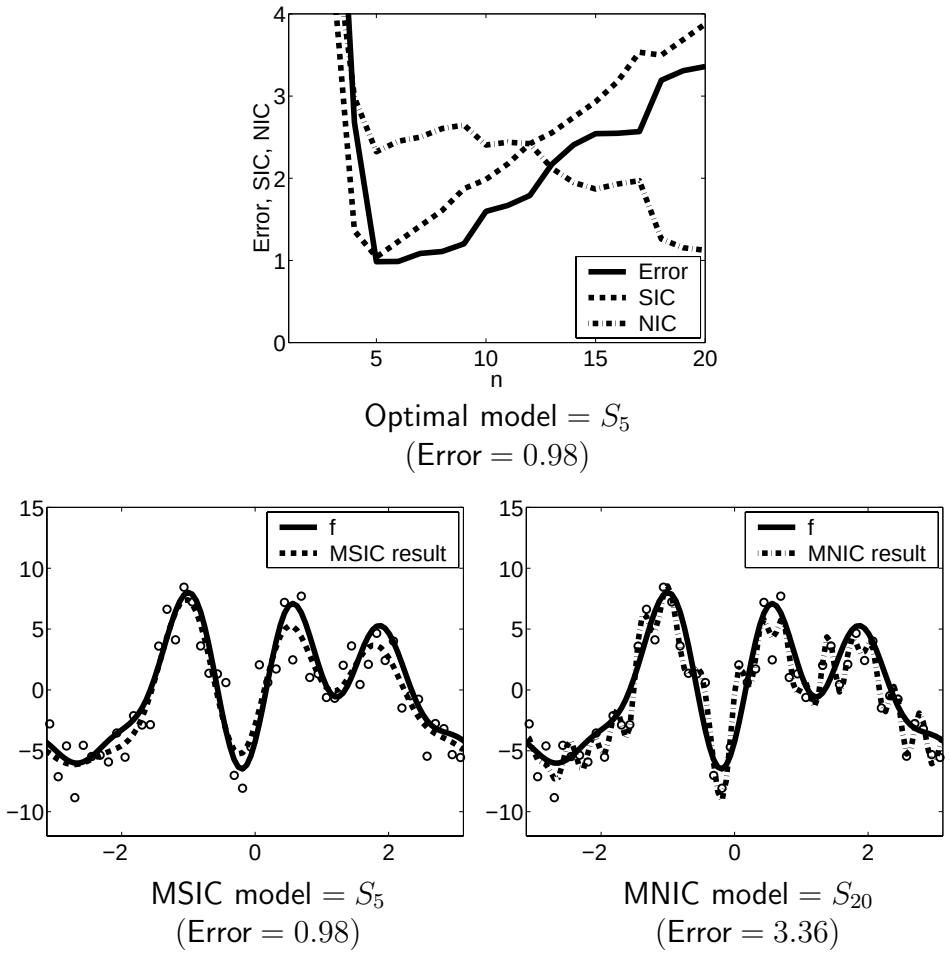

Figure 2: Simulation results when the number M of training examples is 50. The top graph shows the values of the error measured by Eq.(53), SIC, and NIC in each model, denoted by the solid, dashed, and dash-dotted lines, respectively. The horizontal axis denotes the highest order n of trigonometric polynomials in the model S_n. The bottom-left graph shows the target function $f(x)$, training examples, and the MSIC learning result, denoted by the solid line, '$\circ$', and the dashed line, respectively. The bottom-right graph shows the target function $f(x)$, training examples, and the MNIC learning result, denoted by the solid line, '$\circ$', and the dash-dotted line, respectively. The minimum value of the error is 0.98 attained by the model S_5. The MSIC model is S_5 and the error is 0.98. The MNIC model is S_{20} and the error is 3.36.

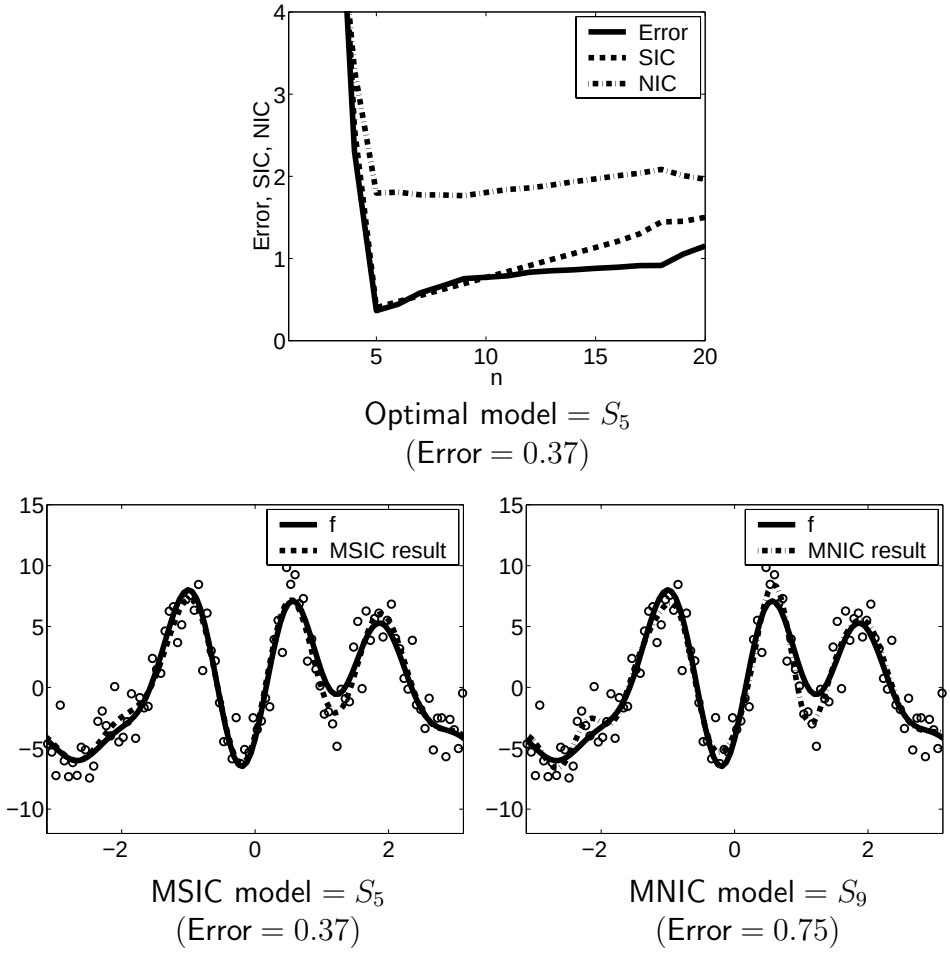

Figure 3: Simulation results when the number M of training examples is 100. The minimum value of the error is 0.37 attained by the model S_5. The MSIC model is S_5 and the error is 0.37. The MNIC model is S_9 and the error is 0.75.

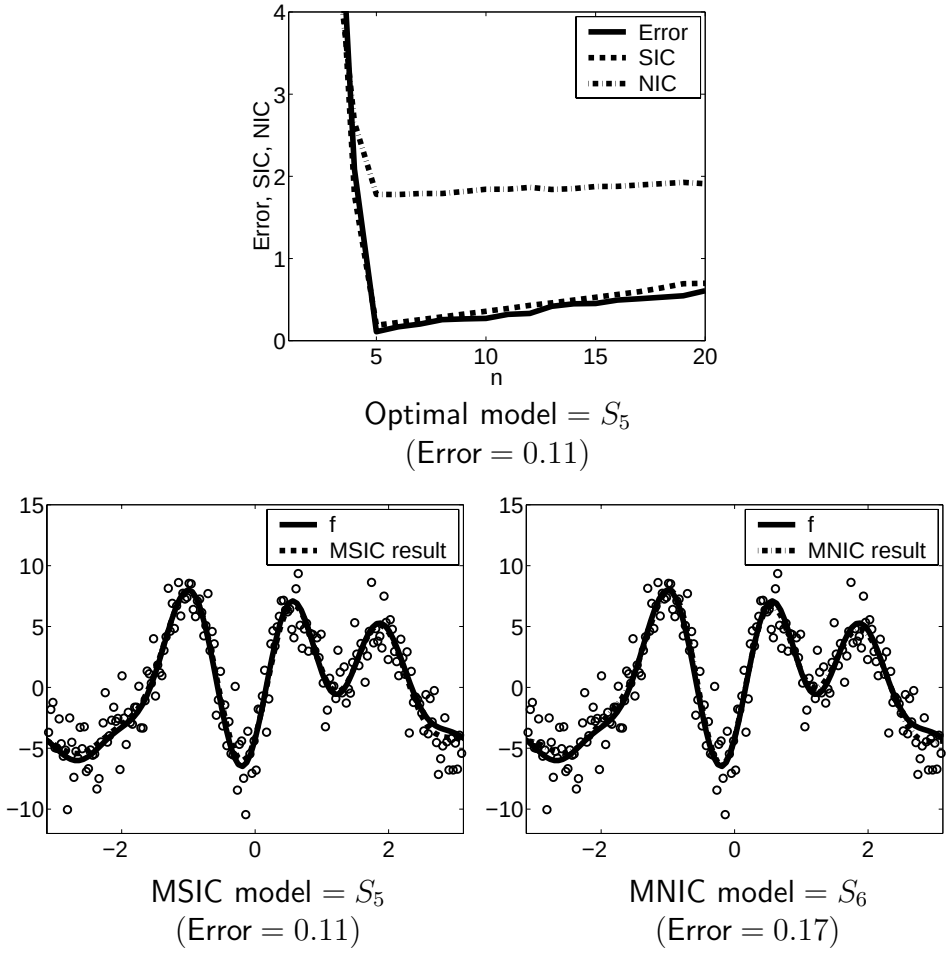

Figure 4: Simulation results when the number M of training examples is 200. The minimum value of the error is 0.11 attained by the model S_5. The MSIC model is S_5 and the error is 0.11. The MNIC model is S_6 and the error is 0.17.

In both (A) and (B), no a priori information is used and the LMS estimator is commonly adopted. Hence, the efficiency of SIC and NIC can be fairly compared by this simulation.

Figs.2, 3, and 4 show the simulation results when the numbers of training examples are 50, 100, and 200, respectively. The top graphs show the values of the error measured by Eq.(53), SIC, and NIC by each model. They are shown by the solid, dashed, and dash-dotted lines, respectively. The horizontal axis denotes the highest order n of trigonometric polynomials in the model S_n. The bottom-left graphs show the target function $f(x)$, training examples, and the MSIC learning result, denoted by the solid line, '$\circ$', and the dashed line, respectively. The bottom-right graphs show the target function $f(x)$, training examples, and the minimum NIC (MNIC) learning result, denoted by the solid line, '$\circ$', and the dash-dotted line, respectively.

When $M = 50$, the minimum value of the error measured by Eq.(53) is 0.98 attained by the model S_5. The MSIC model is S_5 and the error of the MSIC learning result is 0.98. The MNIC model is S_{20} and the error of the MNIC learning result is 3.36. When $M = 100$, the minimum value of the error is 0.37 attained by the model S_5. The MSIC model is S_5 and the error is 0.37, while the MNIC model is S_9 and the error is 0.75. When $M = 200$, the minimum value of the error is 0.11 attained by the model S_5. The MSIC model is S_5 and the error is 0.11, while the MNIC model is S_6 and the error is 0.17.

These results show that when M is large, both SIC and NIC give reasonable learning results (see Figs.3 and 4). However, when it comes to the case where $M = 50$, SIC outperforms NIC (see Fig.2). This implies that SIC works well even when the number of training examples is small.

As mentioned in Section 4.2.7, NIC neglects terms which are not dominant for model selection while SIC gives an unbiased estimate of the generalization error. The top graph in Fig.4 clearly shows this difference. SIC well approximates the true error while the value of NIC is larger than the true error. Generally speaking, neglecting non-dominant terms does not affect the performance of model selection. However, this graph shows that, for $5 \leq n \leq 20$, the slope of NIC is gentler than the true error while that of SIC is in good agreement with it. This implies that SIC is expected to be resistant to the noise since the discrimination of models by SIC is clearer than NIC.

6.2 Interpolation of chaotic series

Let us consider the problem of interpolating the following chaotic series created by the Mackey-Glass delay-difference equation (see e.g. Platt, 1991):

$$g(t+1) = \begin{cases} (1-b)g(t) + \dfrac{a\, g(t-\tau)}{1+g(t-\tau)^{10}} & \text{for } t \geq \tau+1, \\ 0.3 & \text{for } 1 \leq t \leq \tau, \end{cases} \tag{54}$$

where $a = 0.2$, $b = 0.1$, and $\tau = 17$. Let $\{h_t\}_{t=1}^{200}$ be

$$h_t = g(t+\tau+1). \tag{55}$$

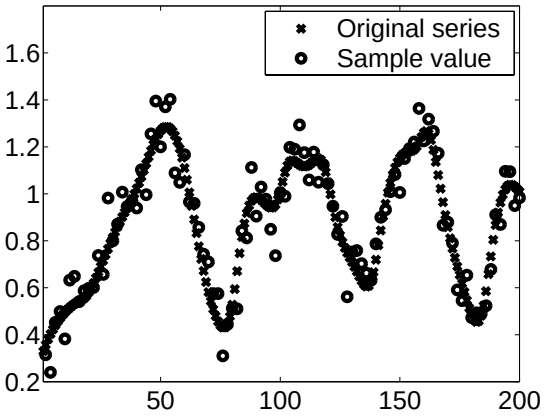

Figure 5: Chaotic series created by the Mackey-Glass delay-difference equation and 100 sample values.

Our task is to estimate $\{h_t\}_{t=1}^{200}$ from M sample values $\{y_m\}_{m=1}^{M}$:

$$y_m = h_p + \epsilon_m \ : \ p = \left\lceil 200 \times \frac{m}{M} \right\rceil, \tag{56}$$

where $\lceil c \rceil$ denotes the minimum integer larger than or equal to c and $\{\epsilon_m\}_{m=1}^{M}$ are noises independently subject to the normal distribution:

$$\epsilon_m \sim N\left(0, \frac{1}{100} \exp\left\{ -\left(\frac{p-100.5}{500}\right)^2 \right\}\right). \tag{57}$$

True values $\{h_t\}_{t=1}^{200}$ and an example of sample values $\{y_m\}_{m=1}^{M}$ are shown in Fig.5.

Let us consider sample points $\{x_m\}_{m=1}^{M}$ corresponding to sample values $\{y_m\}_{m=1}^{M}$:

$$x_m = -0.995 + \frac{2}{200}(p-1) \ : \ p = \left\lceil 200 \times \frac{m}{M} \right\rceil. \tag{58}$$

Then $\hat{f}(-0.995 + \frac{2}{200}(t-1))$ can be regarded as an estimate of h_t for $1 \leq t \leq 200$, where $\hat{f}(x)$ is a learning result from $\{(x_m, y_m)\}_{m=1}^{M}$. Let us consider the following models:

$$\{S_{15}, S_{20}, S_{25}, S_{30}, S_{35}, S_{40}\}, \tag{59}$$

where S_n is a polynomial space of order n, i.e., a Hilbert space spanned by

$$\{x^p\}_{p=0}^{n}, \tag{60}$$

and the inner product is defined by

$$\langle f, g \rangle_{S_n} = \int_{-1}^{1} f(x)\overline{g(x)}dx. \tag{61}$$

The reproducing kernel of S_n is expressed by using the Christoffel-Darboux formula (see e.g. Szegö, 1939; Abramowitz & Segun, 1964; Freud, 1966) as

$$K_{S_n}(x,x') = \begin{cases} \frac{n+1}{2(x-x')}[P_{n+1}(x)P_n(x') - P_n(x)P_{n+1}(x')] & \text{if } x \neq x', \\ \frac{(n+1)^2}{2(1-x^2)}[P_n(x)^2 - 2xP_n(x)P_{n+1}(x) + P_{n+1}(x)^2] & \text{if } x = x', \end{cases} \tag{62}$$

where $P_n(x)$ is the Legendre polynomial of order n defined as

$$P_n(x) = \frac{1}{2^n n!}\frac{d^n}{dx^n}(x^2-1)^n. \tag{63}$$

We again compare SIC and NIC. S_{40} is adopted as H in SIC. The error is measured by

$$\text{Error} = \frac{1}{200}\sum_{t=1}^{200}\left|\hat{f}\left(-0.995 + \frac{2}{200}(t-1)\right) - h_t\right|^2. \tag{64}$$

The distributions of errors by 1000 trials are shown in Figs.6, 7, and 8, where the numbers of training examples are 50, 150, and 250, respectively. The horizontal axis denotes the error measured by Eq.(64), while the vertical axis denotes the number of trials in which the corresponding generalization error is given. The distributions of the errors by SIC and NIC are almost the same when the number M of training examples is 250 (see Fig.8). However, when it comes to the case where $M = 50$ and $M = 150$, SIC tends to outperform NIC (see Figs.6 and 7). This simulation again shows that SIC works well even when the number of training examples is small.

7 Conclusion

In this paper, we proposed a new model selection criterion called the subspace information criterion (SIC). It is assumed that the learning target function belongs to a specified functional Hilbert space and the generalization error is defined as the Hilbert space squared norm of the difference between the learning result function and target function. SIC gives an unbiased estimate of the generalization error. SIC assumed the availability of an unbiased estimate of the target function and the noise covariance matrix, which are generally unknown. A practical calculation method of SIC for least mean squares learning was provided under the assumption that the dimension of the Hilbert space is less than the number of training examples. The range of application of SIC for LMS learning is limited to the finite dimensional Hilbert space case. A practical estimation methods of an unbiased learning result and the noise covariance matrix for the infinite dimensional Hilbert space case is our important future work.

8 Acknowledgements

We would like to thank anonymous referees for bringing C_L to our attention and for their valuable comments.

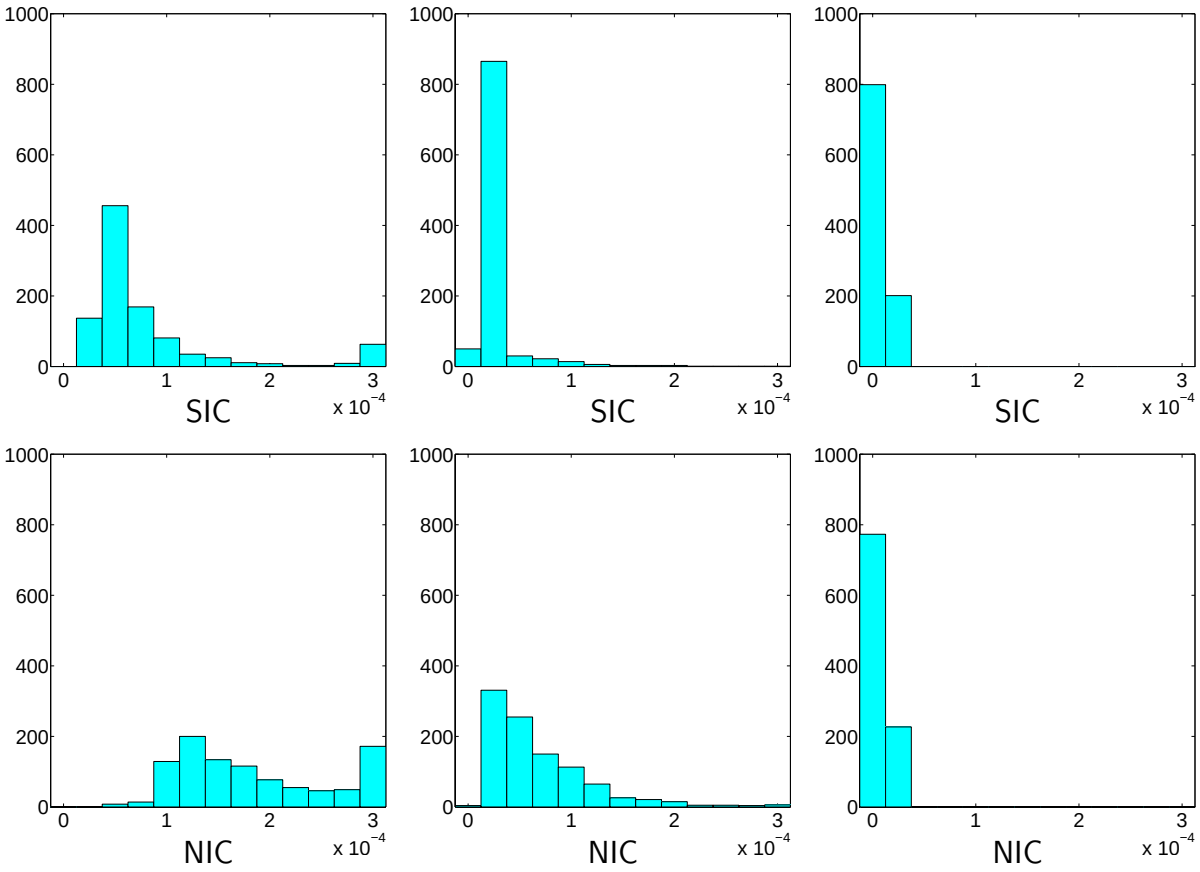

Figure 6: Distributions of errors by 1000 trials when the number M of training examples is 50. The horizontal axis denotes the error measured by Eq.(64) while the vertical axis denotes the number of trials.

Figure 7: Distributions of errors by 1000 trials when the number M of training examples is 150.

Figure 8: Distributions of errors by 1000 trials when the number M of training examples is 250.

A Proof of Theorem 1

It follows from Eq.(24) that

$$\langle K(\cdot, x_{m'}), K(\cdot, x_m)\rangle = K(x_m, x_{m'}). \tag{65}$$

Hence, Eqs.(26) and (30) yield

$$A_S A_S^* = \sum_{m=1}^{M}\sum_{m'=1}^{M} K_S(x_m, x_{m'})\,(e_m \otimes \overline{e_{m'}}) = T_S, \tag{66}$$

$$A_H A_H^* = \sum_{m=1}^{M}\sum_{m'=1}^{M} K_H(x_m, x_{m'})\,(e_m \otimes \overline{e_{m'}}) = T_H. \tag{67}$$

From Eqs.(4), (29), (7), (36), and (33), we have

$$\hat{f}_\theta = X_\theta y = A_S^\dagger y = A_S^*(A_S A_S^*)^\dagger y = A_S^* T_S^\dagger y, \tag{68}$$

$$\hat{f}_u = X_u y = A_H^\dagger y = A_H^*(A_H A_H^*)^\dagger y = A_H^* T_H^\dagger y, \tag{69}$$

which imply Eqs.(39) and (40) because of Eqs.(26) and (30). ■

B Proof of Theorem 2

Since S is a subspace of H, it follows from Eqs.(30), (26), (65), and (66) that

$$\begin{aligned} A_H A_S^* &= \sum_{m=1}^{M}\sum_{m'=1}^{M} \langle K_S(\cdot, x_{m'}), K_H(\cdot, x_m)\rangle\,(e_m \otimes \overline{e_{m'}}) \\ &= \sum_{m=1}^{M}\sum_{m'=1}^{M} K_S(x_m, x_{m'})\,(e_m \otimes \overline{e_{m'}}) \\ &= T_S. \end{aligned} \tag{70}$$

It follows from Eqs.(68), (69), (70), and (41) that

$$\begin{aligned} \|\hat{f}_\theta - \hat{f}_u\|^2 &= \|A_S^* T_S^\dagger y - A_H^* T_H^\dagger y\|^2 \\ &= \langle (A_S^* T_S^\dagger - A_H^* T_H^\dagger)^*(A_S^* T_S^\dagger - A_H^* T_H^\dagger) y, y\rangle \\ &= \langle Ty, y\rangle. \end{aligned} \tag{71}$$

It follows from Eqs.(12), (29), (36), (33), (37), (66), (67), (70), and (41) that

$$\begin{aligned} \mathrm{tr}\,(X_0 Q X_0^*) &= \hat{\sigma}^2 \mathrm{tr}\left((A_S^\dagger - A_H^\dagger)(A_S^\dagger - A_H^\dagger)^*\right) \\ &= \hat{\sigma}^2 \mathrm{tr}\left((A_S^\dagger - A_H^\dagger)^*(A_S^\dagger - A_H^\dagger)\right) \\ &= \hat{\sigma}^2 \mathrm{tr}\,(T)\,. \end{aligned} \tag{72}$$

It follows from Eqs.(29), (37), and (66) that

$$\mathrm{tr}\,(X_\theta Q X_\theta^*) = \hat{\sigma}^2 \mathrm{tr}\left(A_S^\dagger (A_S^\dagger)^*\right) = \hat{\sigma}^2 \mathrm{tr}\left((A_S^\dagger)^* A_S^\dagger\right) = \hat{\sigma}^2 \mathrm{tr}\left((A_S A_S^*)^\dagger\right) = \hat{\sigma}^2 \mathrm{tr}\left(T_S^\dagger\right). \quad (73)$$

From Eqs.(38), (24), (69), and (67), we have

$$\hat{\sigma}^2 = \frac{\|A_H \hat{f}_u - y\|^2}{M - \dim(H)} = \frac{\|A_H A_H^* T_H^\dagger y - y\|^2}{M - \dim(H)} = \frac{\|y\|^2 - \langle T_H T_H^\dagger y, y\rangle}{M - \dim(H)}. \quad (74)$$

Substituting Eqs.(71), (72), (73), and (74) into Eq.(15), we have Eqs.(42) and (43). ■

References

Abramowitz, M., & Segun, I. A. (Eds.) (1964). *Handbook of mathematical functions with formulas, graphs, and mathematical tables*. New York: Dover Publications.

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

Akaike, H. (1980). Likelihood and the Bayes procedure. In N. J. Bernardo, M. H. DeGroot, D. V. Lindley, & A. F. M. Smith (Eds.), *Bayesian Statistics* (pp. 141–166). Valencia: University Press.

Akaike, H., & Kitagawa, G. (Eds.) (1994). *The practice of time series analysis I*. Tokyo: Asakura Syoten. (In Japanese)

Akaike, H., & Kitagawa, G. (Eds.) (1995). *The practice of time series analysis II*. Tokyo: Asakura Syoten. (In Japanese)

Albert, A. (1972). *Regression and the Moore-Penrose pseudoinverse*. New York and London: Academic Press.

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

Amari, S. (1967). Theory of adaptive pattern classifiers. *IEEE Transactions on Electronic Computers, EC-16(3)*, 299–307.

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society, 68*, 337–404.

Ben-Israel, A., & Greville, T. N. E. (1974). *Generalized inverses: Theory and applications*. New York: John Wiley & Sons.

Bergman, S. (1970). *The kernel function and conformal mapping*. Providence, Rhode Island: American Mathematical Society.

Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford: Clarendon Press.

Bozdogan, H. (Ed.) (1994). *Proceedings of the first US/Japan conference on the frontiers of statistical modeling: An informational approach*. Netherlands: Kluwer Academic Publishers.

Cavanaugh, J. E., & Shumway, R. H. (1997). A bootstrap variant of AIC for state space model selection. *Statistica Sinica, 7*, 473–496.

Cherkassky, V., Shao, X., Mulier, F. M., & Vapnik, V. N. (1999). Model complexity control for regression using VC generalization bounds. *IEEE Transactions on Neural Networks, 10(5)*, 1075–1089.

Craven, P., & Wahba, G. (1979). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik, 31*, 377–403.

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application*. Cambridge University Press.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics, 7(1)*, 1–26.

Efron, B. (1986). How biased is the apparent error rate of a prediction rule? *Journal of the American Statistical Association, 81(394)*, 461–470.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Fedorov, V. V. (1972). *Theory of optimal experiments*. New York: Academic Press.

Freud, G. (1966). *Orthogonal polynomials*. Oxford: Pergamon Press.

Fujikoshi, Y., & Satoh, K. (1997). Modified AIC and C_P in multivariate linear regression. *Biometrika, 84*, 707–716.

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

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

Gorman, J. W., & Toman, R. J. (1966). Selection of variables for fitting equations to data. *Technometrics, 8(1)*, 27–51.

Hurvich, C. M., Simonoff, J. S., & Tsai, C. L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society, Series B, 60*, 271–293.

Hurvich, C. M., & Tsai, C. L. (1989). Regression and time series model selection in small samples. *Biometrika, 76*, 297–307.

Hurvich, C. M., & Tsai, C. L. (1991). Bias of the corrected AIC criterion for under-fitted regression and time series models. *Biometrika, 78*, 499–509.

Hurvich, C. M., & Tsai, C. L. (1993). A corrected Akaike information criterion for vector autoregressive model selection. *Journal of Time Series Analysis, 14*, 271–279.

Ishiguro, M., Sakamoto, Y., & Kitagawa, G. (1997). Bootstrapping log likelihood and EIC, an extension of AIC. *Annals of the Institute of Statistical Mathematics, 49*, 411–434.

Kitagawa, G., & Gersch, W. (1996). *Smoothness priors analysis of time series*. Lecture Notes in Statistics, 116. New York: Springer-Verlag.

Konishi, S., & Kitagawa, G. (1996). Generalized information criterion in model selection. *Biometrika, 83*, 875–890.

Li, K. (1986). Asymptotic optimality of C_L and generalized cross-validation in ridge regression with application to spline smoothing. *The Annals of Statistics, 14(3)*, 1101–1112.

MacKay, D. (1992). Bayesian interpolation. *Neural Computation, 4(3)*, 415–447.

Mallows, C. L. (1964). Choosing variables in a linear regression: A graphical aid. Presented at *the Central Regional Meeting of the Institute of Mathematical Statistics*, Manhattan, Kansas.

Mallows, C. L. (1973). Some comments on C_P. *Technometrics, 15(4)*, 661–675.

McQuarrie, A. D. R., & Tsai, C. L. (1998). *Regression and time series model selection*. Singapore, New Jersey: World Scientific.

Mosteller, F., & Wallace, D. (1963). Inference in an authorship problem. A comparative study of discrimination methods applied to the authorship of the disputed Federalist papers. *Journal of the American Statistical Association, 58*, 275–309.

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

Nakashima, A., & Ogawa, H. (1999). How to design a regularization term for improving generalization. In *Proceedings of ICONIP'99, the 6th International Conference on Neural Information Processing, 1* (pp. 222–227), Perth, Australia.

Noda, K., Miyaoka, E., & Itoh, M. (1996). On bias correction of the Akaike information criterion in linear models. *Communications in Statistics. Theory and Methods, 25*, 1845–1857.

Ogawa, H. (1987). Projection filter regularization of ill-conditioned problem. In *Proceedings of SPIE, Inverse Problems in Optics, 808* (pp. 189–196).

Ogawa, H. (1992). Neural network learning, generalization and over-learning. In *Proceedings of the ICIIPS'92, International Conference on Intelligent Information Processing & System, 2* (pp. 1–6), Beijing, China.

Oja, E., & Ogawa, H. (1986). Parametric projection filter for image and signal restoration. *IEEE Transactions on Acoustics, Speech, and Signal Processing, ASSP-34(6)*, 1643–1653.

Onoda, T. (1995). Neural network information criterion for the optimal number of hidden units. In *Proceedings of IEEE International Conference on Neural Networks, ICNN'95* (pp.275–280), Perth, Australia.

Platt, J. (1991). A resource-allocating network for function interpolation. *Neural Computation, 3(2)*, 213–225.

Poggio, T., & Girosi, F. (1990). Networks for approximation and learning. *Proceedings of the IEEE, 78(9)*, 1481–1497.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica, 14*, 465–471.

Rissanen, J. (1987). Stochastic complexity. *Journal of the Royal Statistical Society, Series B, 49(3)*, 223–239.

Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Transactions on Information Theory, IT-42(1)*, 40–47.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986a). Learning representations by back-propagating errors. *Nature, 323*, 533–536.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986b). Learning internal representations by error propagation. In D. E. Rumelhart, J. L. McClelland, & the PDP Research Group (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, 1* (pp. 318–362). Cambridge, MA: The MIT Press.

Saitoh, S. (1988). *Theory of reproducing kernels and its applications*. Pitsman Research Notes in Mathematics Series, 189. UK: Longman Scientific & Technical.

Saitoh, S. (1997). *Integral transform, reproducing kernels and their applications*. Pitsman Research Notes in Mathematics Series, 369. UK: Longman.

Sato, K., Kobayashi, M., & Fujikoshi, Y. (1997). Variable selection for the growth curve model. *Journal of Multivariate Analysis, 60*, 277–292.

Savage, L. J. (1954). *The foundation of statistics*. New York: Wiley.

Schatten, R. (1970). *Norm ideals of completely continuous operators*. Berlin: Springer-Verlag.

Schölkopf, B., Burges, C. J. C., & Smola, A. J. (1998). *Advances in kernel methods: Support vector machines*. Cambridge, MA: The MIT Press.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461–464.

Shibata, R. (1989). Statistical aspects of model selection. In J. C. Willems (Ed.), *From Data to Model* (pp. 375–394). New York: Springer-Verlag.

Shibata, R. (1997). Bootstrap estimate of Kullback-Leibler information for model selection. *Statistica Sinica, 7*, 375–394.

Simonoff, J. S. (1998). Three sides of smoothing: Categorical data smoothing, nonparametric regression, and density estimation. *International Statistical Review, 66*, 137–156.

Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics. Theory and Methods, 7(1)*, 13–26.

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

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society, Series B, 39*, 44–47.

Szegö, G. (1939). *Orthogonal polynomials*. Providence, Rhode Island: American Mathematical Society.

Takemura, A. (1991). *Modern mathematical statistics*. Tokyo: Sobunsya. (In Japanese)

Takeuchi, K. (1976). Distribution of information statistics and validity criteria of models. *Mathematical Science, 153*, 12–18. (In Japanese)

Takeuchi, K. (1983). On the selection of statistical models by AIC. *Journal of the Society of Instrument and Control Engineering, 22(5)*, 445–453. (In Japanese)

Tikhonov, A. N., & Arsenin, V. Y. (1977). *Solutions of ill-posed problems*. Washington DC: V. H. Winston.

Vapnik, V. N. (1995). *The nature of statistical learning theory*. Berlin: Springer-Verlag.

Wada, Y., & Kawato, M. (1991). Estimation of generalization capability by combination of new information criterion and cross validation. *The Transactions of the IEICE, J74-D-II(7)*, 955–965. (In Japanese)

Wahba, H. (1990). *Spline model for observational data*. Philadelphia and Pennsylvania: Society for Industrial and Applied Mathematics.

Wong, W. (1983). A note on the modified likelihood for density estimation. *Journal of the American Statistical Association, 78(382)*, 461–463.

Yamanishi, K. (1998). A decision-theoretic extension of stochastic complexity and its application to learning. *IEEE Transactions on Information Theory, IT-44*, 1424–1439.