Machine Learning, vol.48, no.1/2/3, pp.25–50, 2002.
(Special Issue on New Methods for Model Selection and Model Combination)

Theoretical and Experimental Evaluation of Subspace Information Criterion

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

Recently, a new model selection criterion called the subspace information criterion (SIC) was proposed. SIC works well with small samples since it gives an unbiased estimate of the generalization error with finite samples. In this paper, we theoretically and experimentally evaluate the effectiveness of SIC in comparison with existing model selection techniques including the traditional leave-one-out cross-validation (CV), Mallows's C_P, Akaike's information criterion (AIC), Sugiura's corrected AIC (cAIC), Schwarz's Bayesian information criterion (BIC), Rissanen's minimum description length criterion (MDL), and Vapnik's measure (VM). Theoretical evaluation includes the comparison of the generalization measure, approximation method, and restriction on model candidates and learning methods. Experimentally, the performance of SIC in various situations is investigated. The simulations show that SIC outperforms existing techniques especially when the number of training examples is small and the noise variance is large.

Keywords

supervised learning, generalization capability, model selection, subspace information criterion, small samples

1 Introduction

Supervised learning is estimating unknown input-output dependency from available input-output examples. Once the dependency has been accurately estimated, it can be used for predicting output values corresponding to novel input points. This ability is called the *generalization capability.*

The level of the generalization capability depends heavily on the choice of the *model.* The model indicates, for example, the type and number of basis functions used for learning. The problem of choosing the model that provides the optimal generalization capability is called *model selection.*

Model selection has been extensively studied in the field of statistics. A classic approach to model selection is to find the learning result that minimizes the predictive training error, i.e., the mean error between estimated and true values at sample points contained in the training set. C_P by Mallows (1964, 1973), the generalized cross-validation (GCV) by Craven and Wahba (1979), and the criterion by Mikhal'skii (1987) are based on this approach. Asymptotic optimality of C_P and GCV are shown by Li (1986) (see also Wahba, 1990). However, with finite samples, the optimal generalization capability is not guaranteed since they do not evaluate the generalization error itself.

In contrast, model selection methods which explicitly evaluate the generalization error have been studied from various standpoints, e.g. the information statistics (Akaike, 1974; Takeuchi, 1976; Sugiura, 1978; Konishi & Kitagawa, 1996; Ishiguro *et al.*, 1997) and structural risk minimization (Vapnik, 1995; Cherkassky *et al.*, 1999). Model selection has also been studied from the viewpoints of Bayesian statistics (Schwarz, 1978; Akaike, 1980; MacKay, 1992) and stochastic complexity (Rissanen, 1978, 1987, 1996; Yamanishi, 1998).

Recently, a new model selection criterion called the *subspace information criterion* (SIC) was proposed by Sugiyama and Ogawa (1999, 2001). SIC works well with small samples since it gives an unbiased estimate of the generalization error with finite samples. In this paper, we evaluate the effectiveness of SIC in comparison with existing model selection techniques including the traditional leave-one-out cross-validation (CV), C_P by Mallows (1964, 1973), Akaike's information criterion (AIC) by Akaike (1974), corrected AIC (cAIC) by Sugiura (1978), the Bayesian information criterion (BIC) by Schwarz (1978), the minimum description length criterion (MDL) by Rissanen (1978, 1987, 1996), and Vapnik's measure (VM) by Cherkassky *et al.* (1999).

This paper is organized as follows. In Section 2, the problem of model selection is mathematically formulated. Within this formulation, the derivation of SIC is briefly reviewed in Section 3, and a practical calculation method of SIC for subset regression is given in Section 4. In Section 5, SIC is theoretically compared with existing model selection techniques from points of view of the generalization measure, approximation method, and restriction on model candidates and learning methods. Finally, Section 6 is devoted to computer simulations for experimentally comparing SIC with existing model selection techniques. The simulations will demonstrate that SIC outperforms existing methods especially when the number of training examples is small and the noise variance is large.

2 Problem formulation

In this section, the problem of model selection is mathematically formulated.

Let us consider the supervised learning problem of obtaining an approximation to a learning target function from a set of *training examples.* Let the learning target function $f(x)$ be a complex function 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. The purpose of supervised learning is to obtain the learning result function that minimizes a certain generalization error.

Let θ be a set of factors which determines learning result functions, for example, the type and number of basis functions. We call θ a *model.* Let $\hat{f}_\theta(x)$ be a learning result function obtained with a model θ. We assume that $f(x)$ and $\hat{f}_\theta(x)$ belong to a *reproducing kernel Hilbert space*[1] H. If such a functional Hilbert space H is unknown, then a functional Hilbert space which approximately includes the learning target function $f(x)$ is practically adopted (see Sections 5.2 and 6.5 for detail). We measure the generalization error of $\hat{f}_\theta(x)$ by

$$J_G = \mathrm{E}_\epsilon \|\hat{f}_\theta - f\|^2, \tag{2}$$

where E_ϵ denotes the ensemble average over the noise and $\|\cdot\|$ denotes the norm in H. In the regression case, J_G is typically expressed as

$$J_G = \mathrm{E}_\epsilon \int \left|\hat{f}_\theta(u) - f(u)\right|^2 p(u)du, \tag{3}$$

where $p(\cdot)$ is the probability density function of future (test) input points u. Note that the generalization measure J_G is not averaged over training sample points $\{x_m\}_{m=1}^M$, i.e., we consider a fixed design $\{x_m\}_{m=1}^M$ (see Section 5.1 for detail). Then the problem of model selection considered in this paper is formulated as follows.

[1] The *reproducing kernel* of H, denoted by $K_H(x,x')$, is a bivariate function defined on $\mathcal{D}\times\mathcal{D}$ which satisfies the following conditions (see Aronszajn, 1950; Bergman, 1970; Wahba, 1990; Saitoh, 1988, 1997):

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

  $$\langle f(\cdot), K_H(\cdot,x')\rangle = f(x'),$$

  where $\langle\cdot,\cdot\rangle$ stands for the inner product in H.

The reproducing kernel of H can be expressed by using an orthonormal basis $\{\varphi_p(x)\}_{p=1}^{\mu}$ in H as

$$K_H(x,x') = \sum_{p=1}^{\mu} \varphi_p(x)\overline{\varphi_p(x')},$$

where μ is the dimension of H.

Definition 1 (Model selection) *From a set $\mathcal{M}$ of model candidates, select the best model $\hat{\theta}$ that minimizes the generalization error J_G:*

$$\hat{\theta} = \underset{\theta \in \mathcal{M}}{\operatorname{argmin}} J_G[\theta]. \tag{4}$$

3 Subspace information criterion (SIC)

In this section, we briefly review the derivation of a model selection criterion called the *subspace information criterion* (SIC).

Let y, z, and ϵ be M-dimensional vectors consisting of $\{y_m\}_{m=1}^M$, $\{f(x_m)\}_{m=1}^M$, and $\{\epsilon_m\}_{m=1}^M$, respectively:

$$y = z + \epsilon. \tag{5}$$

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

1. The learning result function $\hat{f}_\theta(x)$ obtained with the model θ is given by using a linear operator X_θ as
$$\hat{f}_\theta = X_\theta y. \tag{6}$$

2. The mean noise is zero:
$$\mathrm{E}_\epsilon \epsilon = 0. \tag{7}$$

3. A linear operator X_u which gives an unbiased learning result function $\hat{f}_u(x)$ is available:
$$\mathrm{E}_\epsilon \hat{f}_u = f, \tag{8}$$
where
$$\hat{f}_u = X_u y. \tag{9}$$

Assumption 1 implies that the range of X_θ becomes a subspace of H. The unbiased learning result function $\hat{f}_u$ is used for estimating the generalization error of $\hat{f}_\theta$.

It follows from Eqs.(6), (8), (5), and (7) that the generalization error of $\hat{f}_\theta$ can be exactly expressed by using $\hat{f}_u$ and the noise covariance matrix Q as

$$\begin{aligned}
J_G[\theta] &= \|\mathrm{E}_\epsilon \hat{f}_\theta - f\|^2 + \mathrm{E}_\epsilon \|\hat{f}_\theta - \mathrm{E}_\epsilon \hat{f}_\theta\|^2 \\
&= \|\hat{f}_\theta - \hat{f}_u\|^2 - \|\hat{f}_\theta - \hat{f}_u\|^2 + \|\mathrm{E}_\epsilon \hat{f}_\theta - f\|^2 + \mathrm{E}_\epsilon \|X_\theta y - \mathrm{E}_\epsilon X_\theta y\|^2 \\
&= \|\hat{f}_\theta - \hat{f}_u\|^2 - \|\mathrm{E}_\epsilon(\hat{f}_\theta - \hat{f}_u) - \mathrm{E}_\epsilon(\hat{f}_\theta - \hat{f}_u) + \hat{f}_\theta - \hat{f}_u\|^2 \\
&\quad + \|\mathrm{E}_\epsilon(\hat{f}_\theta - \hat{f}_u)\|^2 + \mathrm{E}_\epsilon \|X_\theta(z+\epsilon) - \mathrm{E}_\epsilon X_\theta(z+\epsilon)\|^2 \\
&= \|\hat{f}_\theta - \hat{f}_u\|^2 - \|\mathrm{E}_\epsilon(\hat{f}_\theta - \hat{f}_u)\|^2 \\
&\quad - 2\mathrm{Re}\langle \mathrm{E}_\epsilon(\hat{f}_\theta - \hat{f}_u), -\mathrm{E}_\epsilon(\hat{f}_\theta - \hat{f}_u) + \hat{f}_\theta - \hat{f}_u \rangle \\
&\quad - \|\mathrm{E}_\epsilon(\hat{f}_\theta - \hat{f}_u) - (\hat{f}_\theta - \hat{f}_u)\|^2 + \|\mathrm{E}_\epsilon(\hat{f}_\theta - \hat{f}_u)\|^2 + \mathrm{E}_\epsilon \|X_\theta \epsilon\|^2 \\
&= \|\hat{f}_\theta - \hat{f}_u\|^2 - 2\mathrm{Re}\langle \mathrm{E}_\epsilon(\hat{f}_\theta - \hat{f}_u), -\mathrm{E}_\epsilon(\hat{f}_\theta - \hat{f}_u) + \hat{f}_\theta - \hat{f}_u \rangle \\
&\quad - \|\mathrm{E}_\epsilon(\hat{f}_\theta - \hat{f}_u) - (\hat{f}_\theta - \hat{f}_u)\|^2 + \mathrm{tr}\left(X_\theta Q X_\theta^*\right),
\end{aligned} \tag{10}$$

where 'Re', $\langle\cdot,\cdot\rangle$, $\mathrm{tr}\,(\cdot)$, and X_θ^* stand for the real part of a complex number, the inner product in H, the trace of an operator, and the adjoint operator of X_θ, respectively. The second and third terms in Eq.(10) can not be directly evaluated since $\mathrm{E}_\epsilon(\hat{f}_\theta - \hat{f}_u)$ is unknown, so we shall average these terms out over the noise. Then the second term vanishes:

$$\mathrm{E}_\epsilon \mathrm{Re}\langle \mathrm{E}_\epsilon(\hat{f}_\theta - \hat{f}_u), -\mathrm{E}_\epsilon(\hat{f}_\theta - \hat{f}_u) + \hat{f}_\theta - \hat{f}_u\rangle = 0. \tag{11}$$

And it follows from Eqs.(6), (9), (5), and (7) that the third term yields

$$\begin{aligned}
&\mathrm{E}_\epsilon\|\mathrm{E}_\epsilon(\hat{f}_\theta - \hat{f}_u) - (\hat{f}_\theta - \hat{f}_u)\|^2 \\
&\qquad = \mathrm{E}_\epsilon\|\mathrm{E}_\epsilon(X_\theta - X_u)y - (X_\theta - X_u)y\|^2 \\
&\qquad = \mathrm{E}_\epsilon\|\mathrm{E}_\epsilon(X_\theta - X_u)(z+\epsilon) - (X_\theta - X_u)(z+\epsilon)\|^2 \\
&\qquad = \mathrm{E}_\epsilon\|(X_\theta - X_u)\epsilon\|^2 \\
&\qquad = \mathrm{tr}\left((X_\theta - X_u)Q(X_\theta - X_u)^*\right).
\end{aligned} \tag{12}$$

Then we have the following criterion.

Definition 2 (Subspace information criterion) *(Sugiyama & Ogawa, 1999, 2001) The following functional* $\mathrm{SIC}[\theta]$ *is called the subspace information criterion for a model* θ*.*

$$\mathrm{SIC}[\theta] = \|\hat{f}_\theta - \hat{f}_u\|^2 - \mathrm{tr}\left((X_\theta - X_u)Q(X_\theta - X_u)^*\right) + \mathrm{tr}\left(X_\theta Q X_\theta^*\right). \tag{13}$$

The model that minimizes SIC is called the *minimum SIC (MSIC) model*. An important property of SIC is that it is an unbiased estimate of the generalization error:

$$\mathrm{E}_\epsilon\ \mathrm{SIC}[\theta] = J_G[\theta] \tag{14}$$

since it follows from Eqs.(13), (11), (12), and (10) that

$$\begin{aligned}
\mathrm{E}_\epsilon\ \mathrm{SIC}[\theta] &= \mathrm{E}_\epsilon\|\hat{f}_\theta - \hat{f}_u\|^2 - \mathrm{tr}\left((X_\theta - X_u)Q(X_\theta - X_u)^*\right) \\
&\quad + \mathrm{tr}\left(X_\theta Q X_\theta^*\right) \\
&= \mathrm{E}_\epsilon J_G[\theta] \\
&= J_G[\theta].
\end{aligned} \tag{15}$$

Eq.(14) supports the effectiveness of SIC as a model selection criterion.

4 Practical calculation of SIC for subset regression

Although SIC does not include the unknown learning target function $f(x)$, it requires other information which is often unknown, e.g. an unbiased estimate $\hat{f}_u$ and the noise covariance matrix Q. In this section, we give a practical calculation method of SIC for linear regression models under the following conditions.

(a) The learning target function $f(x)$ is expressed by using a given set $\{\varphi_p(x)\}_{p=1}^{\mu}$ of μ linearly independent functions as

$$f(x) = \sum_{p=1}^{\mu} a_p \varphi_p(x). \tag{16}$$

This implies that $f(x)$ is included in a μ-dimensional functional Hilbert space H spanned by $\{\varphi_p(x)\}_{p=1}^{\mu}$. If such a set $\{\varphi_p(x)\}_{p=1}^{\mu}$ of basis functions is unknown, basis functions which approximately express the learning target function $f(x)$ are practically adopted (see Sections 5.2 and 6.5 for detail).

(b) The so-called *design matrix* B, which is a $M \times \mu$ matrix with (m, p)-th element being $\varphi_p(x_m)$ (see e.g. Efron & Tibshirani, 1993; Orr, 1996), has the rank μ:

$$\mathrm{rank}\ B = \mu, \tag{17}$$

where

$$[B]_{m,p} = \varphi_p(x_m). \tag{18}$$

$[\cdot]_{m,p}$ denotes the (m, p)-th element of a matrix.

(c) The number of training examples is larger than the number of basis functions:

$$M > \mu. \tag{19}$$

(d) The noise covariance matrix Q is given as

$$Q = \sigma^2 I_M \tag{20}$$

where $\sigma^2 > 0$ and I_M is the M-dimensional identity matrix.

(e) The generalization measure J_G is defined by Eq.(3).

(f) The following μ-dimensional covariance matrix U is known:

$$[U]_{p,p'} = \int \varphi_{p'}(u)\overline{\varphi_p(u)}p(u)du, \tag{21}$$

where $\overline{\varphi_p(u)}$ is the complex conjugate of $\varphi_p(u)$ and $p(u)$ is the probability density function of future sample points u (see Sections 5.2 and 6.4 if U is unknown).

(g) Least mean squares (LMS) learning (i.e., the training error minimization learning) in a subspace S spanned by a subset of $\{\varphi_p(x)\}_{p=1}^{\mu}$ is adopted:

$$\min_{g \in S} \frac{1}{M} \sum_{m=1}^{M} |g(x_m) - y_m|^2 . \tag{22}$$

Under the assumptions (a) and (b), the *best linear unbiased estimate (BLUE)* of f, which is the linear unbiased estimate of f with minimum variance (see e.g. Albert, 1972), can be obtained. We adopt BLUE as $\hat{f}_u$, which is given as

$$\hat{f}_u(x) = \sum_{p=1}^{\mu} [B^\dagger y]_p \varphi_p(x), \tag{23}$$

where $B^\dagger$ is the *Moore-Penrose generalized inverse*[2] of B and $[\cdot]_p$ denotes the p-th element of a vector. Under the assumptions (a), (b), (c), and (d), an unbiased estimate of the noise variance σ^2 is given as (see e.g. Fedorov, 1972)

$$\hat{\sigma}^2 = \frac{\langle y - BB^\dagger y, y\rangle}{M - \mu}. \tag{24}$$

Let θ be a subset of indices $\{1, 2, \ldots, \mu\}$ and B_θ be an $M \times \mu$ matrix whose (m, p)-th element is

$$[B_\theta]_{m,p} = \begin{cases} \varphi_p(x_m) & \text{if } p \in \theta, \\ 0 & \text{otherwise.} \end{cases} \tag{25}$$

Then the LMS estimate (i.e., the minimum training error estimate) of f in a subspace spanned by a subset $\{\varphi_p(x)\}_{p\in\theta}$ is given as (see e.g. Efron & Tibshirani, 1993)

$$\hat{f}_\theta(x) = \sum_{p\in\theta} [B_\theta^\dagger y]_p \varphi_p(x). \tag{26}$$

Let W be an operator from $\mathbf{C}^\mu$ to H defined as

$$W = \sum_{p=1}^{\mu} \left(\varphi_p \otimes \overline{e_p}\right), \tag{27}$$

where $(\cdot \otimes \bar{\cdot})$ denotes the *Neumann-Schatten product*[3] and e_p is the p-th vector of the so-called standard basis in $\mathbf{C}^\mu$. From Eqs.(23), (26), and (27), operators X_u and X_θ in Eqs.(9) and (6) are expressed as

$$\begin{aligned} X_u &= WB^\dagger, & (28) \\ X_\theta &= WB_\theta^\dagger. & (29) \end{aligned}$$

[2] A matrix A is called the *Moore-Penrose generalized inverse* of a matrix B if A satisfies the following four conditions (see e.g. Albert, 1972; Ben-Israel & Greville, 1974).

$$BAB = B, \quad ABA = A, \quad (BA)^* = BA, \quad \text{and} \quad (AB)^* = AB.$$

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

[3] 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 (see Schatten, 1970)

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

Since it follows from Eqs.(21), (27), and (3) that $U = W^*W$, SIC is reduced to as

$$\begin{aligned}\mathrm{SIC}[\theta] &= \langle U(B_\theta^\dagger - B^\dagger)y, (B_\theta^\dagger - B^\dagger)y\rangle \\ &\quad - \hat{\sigma}^2 \mathrm{tr}\left(U(B_\theta^\dagger - B^\dagger)(B_\theta^\dagger - B^\dagger)^*\right) \\ &\quad + \hat{\sigma}^2 \mathrm{tr}\left(UB_\theta^\dagger (B_\theta^\dagger)^*\right). \qquad (30)\end{aligned}$$

Note that Eq.(30) is an unbiased estimate of the generalization error, i.e., Eq.(14) holds. Practically, we recommend using *Tikhonov's regularization* (Tikhonov & Arsenin, 1977) for calculating the Moore-Penrose generalized inverse:

$$B^\dagger \longleftarrow (B^*B + \gamma I_\mu)^{-1} B^*, \qquad (31)$$

where γ is a small positive constant and I_μ is the μ-dimensional identity matrix.

5 Theoretical evaluation of SIC

In this section, SIC is compared with the traditional leave-one-out cross-validation (CV), C_P by Mallows (1964, 1973), Akaike's information criterion (AIC) by Akaike (1974), corrected AIC (cAIC) by Sugiura (1978), the Bayesian information criterion (BIC) by Schwarz (1978), the minimum description length criterion (MDL) by Rissanen (1978, 1987, 1996), and Vapnik's measure (VM) by Cherkassky *et al.* (1999).

5.1 Generalization measure

SIC can adopt any norm in the functional Hilbert space as the generalization measure (see Eq.(2)) as long as it can be calculated. It is, for example, expressed as Eq.(3). The derivatives of the functions $\hat{f}_\theta(x)$ and $f(x)$ can also be included in the generalization measures (with the Sobolev norm).

C_P adopts the predictive training error as the error measure:

$$\frac{1}{M}\mathrm{E}_\epsilon \sum_{m=1}^{M} \left|\hat{f}_\theta(x_m) - f(x_m)\right|^2. \qquad (32)$$

Note that Eq.(32) does not evaluate the error at future sample points u, and it is equivalent to Eq.(3) with $p(u)$ replaced by the empirical distribution. It is shown that Eq.(32) converges to Eq.(3) when training sample points $\{x_m\}_{m=1}^M$ are subject to $p(\cdot)$ and the number M of training examples tends to be infinity.

CV adopts the leave-one-out error as the error measure:

$$\sum_{m=1}^{M} \left|\hat{f}_\theta^{(m)}(x_m) - y_m\right|^2, \qquad (33)$$

where $\hat{f}_\theta^{(m)}$ denotes the learning result function obtained from the training examples without (x_m, y_m). Eq.(33) also does not directly evaluate the error at future sample points u.

The relation between Eq.(33) and the generalization measure Eq.(3) is not well recognized yet.

AIC and cAIC adopt the expected Kullback-Leibler information (Kullback & Leibler, 1951) over all training sets $\{(x_m, y_m)\}_{m=1}^{M}$ as the generalization measure, which is conceptually similar to the expectation of Eq.(3) over training sample points $\{x_m\}_{m=1}^{M}$ (Murata *et al.*, 1994):

$$\mathrm{E}_{\{x_m\}}\mathrm{E}_{\epsilon}\int\left|\hat{f}_{\theta}(u)-f(u)\right|^2 p(u)du. \tag{34}$$

Although $p(\cdot)$ can be unknown in AIC and cAIC, instead training sample points $\{x_m\}_{m=1}^{M}$ and future sample points u are assumed to be independently subject to the same probability density function $p(\cdot)$ and the generalization measure is further averaged over training sample points. If one adopts the generalization measure averaged over training sample points (Eq.(34)), the purpose of model selection is to obtain the model that gives good learning result functions on average (i.e., *sample-point-independent model selection*). In contrast, if one adopts the generalization measure which is *not* averaged over training sample points (Eq.(3)), the purpose of model selection is to obtain the model that gives the optimal learning result function from a given, particular training set (i.e., *sample-point-dependent model selection*). This implies that the latter standpoint is suitable for acquiring the best prediction performance from given training examples.

BIC gives an estimate of the evidence (MacKay, 1992) and MDL gives an estimate of the description length of the model and data. The relation between the evidence, description length of the model and data, and generalization error is not clear.

The generalization measure adopted in VM is a probabilistic upper bound of the risk functional:

$$\int\left|\hat{f}_{\theta}(u)-f(u)\right|^2 p(u)du, \tag{35}$$

where $p(\cdot)$ can be unknown but training sample points $\{x_m\}_{m=1}^{M}$ and future sample points u are assumed to be independently subject to the same probability density function $p(\cdot)$ instead.

5.2 Approximation methods

C_P, AIC, cAIC, BIC, MDL, and VM are based on the training error:

$$\frac{1}{M}\sum_{m=1}^{M}\left|\hat{f}_{\theta}(x_m)-y_m\right|^2. \tag{36}$$

In contrast, CV and SIC do not use the training error and they directly evaluate the error measures.

C_P is an unbiased estimate of the predictive training error given by Eq.(32) with finite samples. Since the predictive training error asymptotically agrees with the generalization error given by Eq.(3), it can be regarded as an approximation of Eq.(3). Although asymptotic optimality of C_P is shown by Li (1986), its effectiveness with small samples is not theoretically sure.

In CV, the leave-one-out error given by Eq.(33) can be regarded as an approximation of the generalization error (i.e., the error at future sample points u) since it is shown that the model selection by CV is asymptotically equivalent to that by AIC (Stone, 1977; see also Amari *et al.*, 1997 for asymptotic analysis). Although it is known that CV practically works well, its mechanism in small sample cases is not well recognized yet.

Although AIC directly evaluates the generalization error, it is assumed in the derivation that the number of training examples is very large. This means that when the number of training examples is small, the approximation is no longer valid. BIC and MDL also use asymptotic approximation so they have the same drawback.

cAIC, VM, and SIC do not assume the availability of a large number of training examples for evaluating the generalization error. Therefore, they will work well with small samples. cAIC is a modified AIC with consideration of small sample effect for *faithful*[4] models. However, its performance for unfaithful models is not sure. VM gives a probabilistic upper bound of Eq.(35) based on the VC theory (Vapnik, 1995). Although VM is derived under general setting, some heuristics are used in its derivation and the tightness of the upper bound is not evaluated yet.

SIC utilizes only the noise characteristics in its derivation, and it gives an unbiased estimate of the generalization error J_G with finite samples. However, its variance is not theoretically investigated yet (see Section 6 for experimental evaluation). In order to calculate SIC, rather restrictive conditions should be assumed as shown in Section 4. However, these conditions do not have to be rigorously satisfied in practice. For example, when H is unknown, a functional Hilbert space $\hat{H}$ with the following properties is practically adopted.

- $\hat{H}$ approximately includes the learning target function $f(x)$.
- The dimension of $\hat{H}$ is less than the number M of training examples.
- $\hat{H}$ includes all model candidates in the set $\mathcal{M}$ (see Definition 1).

As shown in Section 6.5, such a functional Hilbert space $\hat{H}$ is practically useful. When the covariance matrix U defined by Eq.(21) is unknown, it can be estimated from unlabeled sample points $\{x'_m\}_{m=1}^{M'}$ (i.e., sample points without sample values $\{y'_m\}_{m=1}^{M'}$) as

$$[\hat{U}]_{p,p'} = \frac{1}{M'} \sum_{m=1}^{M'} \varphi_{p'}(x'_m)\overline{\varphi_p(x'_m)}. \tag{37}$$

If the training sample points $\{x_m\}_{m=1}^{M}$ are used instead of unlabeled sample points, then SIC agrees with Mallows's C_P. For this reason, SIC can be regarded as an extension of C_P (see also Sugiyama and Ogawa, 2000b, 2001).

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

5.3 Restriction on model candidates

AIC and cAIC are valid only when model candidates in the set $\mathcal{M}$ are nested (Takeuchi, 1983; Murata *et al.*, 1994). As Murata *et al.* (1994) pointed out, the fact is known to those who work on AIC, but it is still not well known to those who apply AIC in practice. In contrast, SIC imposes no restriction on model candidates in the set $\mathcal{M}$ except that the range of X_θ is included in the functional Hilbert space H.

5.4 Restriction on learning methods

AIC, cAIC, BIC, and MDL are specialized for maximum likelihood estimation, which is equivalent to LMS learning if the noise is subject to the normal distribution. cAIC is valid only for linear regression models. A generalized AIC proposed by Murata *et al.* (1994) and Konishi and Kitagawa (1996) relaxed the restriction of maximum likelihood estimation. C_P is specialized for LMS learning with linear regression models. An extension of C_P called C_L by Mallows (1973), VM, and SIC are applicable to various learning methods expressed by linear operators (see Eq.(6)) including regularization learning with quadratic regularizers (ridge regression). Note that in VM, the VC-dimension (Vapnik, 1995) should be explicitly calculated. SIC for regularization learning with quadratic regularizers is studied in Sugiyama and Ogawa (2000a). SIC can also be approximately applied to non-linear operators (see Tsuda *et al.*, 2000 for sparse regressors).

6 Experimental evaluation of SIC

In this section, SIC is experimentally compared with existing model selection techniques through computer simulations.

6.1 Basic simulations

First, we consider the case when all the assumptions in Section 4 are satisfied. Let the learning target function $f(x)$ be

$$f(x) = \frac{1}{10}\sum_{p=1}^{50}(\sin px + \cos px) \tag{38}$$

defined on $[-\pi, \pi]$. Let us consider a set of the following 201 basis functions which includes $f(x)$:

$$\{1, \sin px, \cos px\}_{p=1}^{100}. \tag{39}$$

Let the set $\mathcal{M}$ of model candidates be

$$\mathcal{M} = \{\theta_0, \theta_{10}, \theta_{20}, \ldots, \theta_{100}\}, \tag{40}$$

where θ_n indicates the following regression model:

$$\hat{f}_{\theta_n}(x) = \hat{a}_0 + \sum_{p=1}^{n}(\hat{a}_{2p-1}\sin px + \hat{a}_{2p}\cos px). \tag{41}$$

Note that the number of basis functions in the model θ_n is $2n+1$. Let us assume that the training sample points $\{x_m\}_{m=1}^{M}$ and future sample points u are independently subject to the same uniform distribution on $[-\pi, \pi]$. Let the noise ϵ_m be independently subject to the same normal distribution with mean 0 and variance σ^2.

$$\epsilon_m \sim N(0, \sigma^2). \tag{42}$$

For calculating the Moore-Penrose generalized inverse, we use Eq.(31) with $\gamma = 0.1$. The following model selection criteria are compared.

(a) Subspace information criterion (SIC) (Sugiyama & Ogawa, 1999, 2001): The model θ_{100} is regarded as H. We assume that the covariance matrix U given by Eq.(21) is exactly known. Since U is the identity matrix in the above setting, SIC for a model θ_n is given as

$$\begin{aligned}\mathrm{SIC}[\theta_n] &= \|(B_{\theta_n}^{\dagger} - B^{\dagger})y\|^2 - \hat{\sigma}^2\mathrm{tr}\left((B_{\theta_n}^{\dagger} - B^{\dagger})(B_{\theta_n}^{\dagger} - B^{\dagger})^*\right) \\ &\quad + \hat{\sigma}^2\mathrm{tr}\left(B_{\theta_n}^{\dagger}(B_{\theta_n}^{\dagger})^*\right),\end{aligned} \tag{43}$$

where $\hat{\sigma}^2$ is given by Eq.(24).

(b) Leave-one-out cross-validation (CV): A closed form expression of the leave-one-out error (Eq.(33)) for a linear regression model θ_n is given as (see Orr, 1996)

$$\mathrm{CV}[\theta_n] = \frac{\|(\mathrm{diag}(Z_{\theta_n}))^{-1}Z_{\theta_n}y\|^2}{M}, \tag{44}$$

where Z_{θ_n} is an M-dimensional matrix defined as

$$Z_{\theta_n} = I_M - B_{\theta_n}B_{\theta_n}^{\dagger}. \tag{45}$$

The matrix 'diag(Z_{θ_n})' is the same size and has the same diagonal as Z_{θ_n} but is zero off the diagonal.

(c) C_P (Mallows, 1964, 1973): C_P for a model θ_n is given as

$$C_P[\theta_n] = \frac{\|Z_{\theta_n}y\|^2}{M} + \frac{2\hat{\sigma}^2(2n+1)}{M} - \hat{\sigma}^2, \tag{46}$$

where $\hat{\sigma}^2$ is given by Eq.(24). Note that $\frac{\|Z_{\theta_n}y\|^2}{M}$ is the training error of a learning result function $\hat{f}_{\theta_n}(x)$ (see Eq.(36)).

(d) Akaike's information criterion (AIC) (Akaike, 1974): Since the noise is subject to the normal distribution, AIC is expressed as

$$\mathrm{AIC}[\theta_n] = M \log \frac{\|Z_{\theta_n} y\|^2}{M} + 2(2n + 1 + 1). \tag{47}$$

(e) Corrected AIC (cAIC) (Sugiura, 1978): Since the noise is subject to the normal distribution, cAIC is expressed as

$$\mathrm{cAIC}[\theta_n] = M \log \frac{\|Z_{\theta_n} y\|^2}{M} + \frac{2(2n + 1 + 1)M}{M - (2n + 1) - 2}. \tag{48}$$

(f) Bayesian information criterion (BIC) (Schwarz, 1978): Since the noise is subject to the normal distribution, BIC is expressed as

$$\mathrm{BIC}[\theta_n] = M \log \frac{\|Z_{\theta_n} y\|^2}{M} + (2n + 1 + 1) \log M. \tag{49}$$

Note that the minimum description length criterion (MDL) (Rissanen, 1978, 1987, 1996) is also given by Eq.(49).

(g) Vapnik's measure (VM) (Cherkassky *et al.*, 1999): Since the VC-dimension of the model θ_n is $2n + 1$ (Vapnik, 1995), VM is expressed as

$$\mathrm{VM}[\theta_n] = \frac{\|Z_{\theta_n} y\|^2}{M} \Bigg/ \max\left(0, 1 - \sqrt{p - p \log p + \frac{\log M}{2M}}\right), \tag{50}$$

where

$$p = \frac{2n + 1}{M}. \tag{51}$$

Note that in AIC, cAIC, and BIC (MDL), the information that the noise is subject to the normal distribution is used. We shall measure the error of a learning result function $\hat{f}_{\theta_n}(x)$ by

$$\mathrm{Error}[\theta_n] = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left| \hat{f}_{\theta_n}(x) - f(x) \right|^2 dx. \tag{52}$$

The simulation is performed 100 times for $(M, \sigma^2) = (500, 0.2)$, $(250, 0.2)$, $(500, 0.6)$, and $(250, 0.6)$, with changing the noise $\{\epsilon_m\}_{m=1}^{M}$ in each trial.

Figures 1, 2, 3, and 4 show the simulation results. The top eight graphs show the values of the error (Eq.(52)) and model selection criteria corresponding to the order n of models (see Eq.(40)). The box plot notation specifies marks at 95, 75, 50, 25, and 5 percentiles of values. The solid line denotes the mean values. The bottom-left eight graphs show the distribution of the selected order n of models. 'OPT' indicates the optimal model that minimizes the error. The bottom-right eight graphs show the distribution of the error obtained by the model selected by each criterion.

When $(M, \sigma^2) = (500, 0.2)$ (Figure 1), all model selection criteria work well. When $(M, \sigma^2) = (250, 0.2)$ (Figure 2), AIC tends to select larger models and BIC (MDL) is inclined to select smaller models, so they provide large errors. This may be caused since AIC and BIC (MDL) are derived under the assumption that the number M of training examples is very large. When $(M, \sigma^2) = (500, 0.6)$ (Figure 3), BIC (MDL) and VM show a tendency to select smaller models and they result in large errors. This implies that BIC (MDL) and VM are not robust against the noise. Finally, when $(M, \sigma^2) = (250, 0.6)$ (Figure 4), SIC works better than other criteria. In this case, C_P almost always selects θ_{50}, AIC tends to select larger models, and other criteria tend to select smaller models. As a result, they give large errors.

The simulation results show that SIC outperforms other model selection criteria especially when the number M of training examples is small and the noise variance σ^2 is large. Although SIC almost always gives a very good estimate of the true error on average, its variance is rather large when $M = 250$ (Figures 2 and 4). However, the large variance of SIC may be dominated by terms that are irrelevant to model selection since SIC given by Eq.(43) is expressed as

$$\begin{aligned} \mathrm{SIC}[\theta_n] &= \|B_{\theta_n}^{\dagger} y\|^2 - 2\mathrm{Re}\langle B_{\theta_n}^{\dagger} y, B^{\dagger} y\rangle + \|B^{\dagger} y\|^2 \\ &\quad + 2\hat{\sigma}^2 \mathrm{Re}\ \mathrm{tr}\left(B_{\theta_n}^{\dagger}(B^{\dagger})^*\right) - \hat{\sigma}^2 \mathrm{tr}\left(B^{\dagger}(B^{\dagger})^*\right), \end{aligned} \tag{53}$$

where the third and fifth terms are irrelevant to θ_n. It should be noted that C_P almost always selects the true model θ_{50} in any cases. This implies that C_P is more suitable for finding the true model than finding the model with minimum generalization error.

6.2 Uniform noise

Let us investigate the robustness of SIC against non-Gaussian noise. We consider the same setting as Section 6.1 but the noise ϵ_m is subject to the uniform distribution on $[-0.3, 0.3]$. The simulation results for $M = 500$ and 250 are displayed in Figures 5 and 6, respectively. The results show that SIC works well even in the uniform noise case.

6.3 Changing the dimension of H

Now we investigate the influence of changing H when $(M, \sigma^2) = (250, 0.2)$. The simulation is performed with the same setting as Section 6.1 but H is changed as θ_{120}, θ_{100}, θ_{80}, or θ_{60}. Note that when H is specified by θ_{80} or θ_{60}, we only consider the model candidates whose orders are less than or equal to the order of H (see Eq.(40)). The simulation results displayed in Figure 7 show that the variance of SIC is reduced as H is small. This may be because $\hat{f}_u$ and $\hat{\sigma}^2$ tend to be accurate as H is small (see Eq.(53)). The performance of SIC is almost the same when H is specified by θ_{100}, θ_{80}, or θ_{60}. However, when H is specified by θ_{120}, the variance is rather large. This implies that when the dimension of H is very close to the number M of training examples (e.g. when H is specified by θ_{120}, $\dim H = 241$ and $M = 250$), SIC tends to be inaccurate.

6.4 Estimating U from unlabeled sample points

Let us investigate the robustness of SIC when the covariance matrix U (see Eq.(21)), which is assumed to be known in the simulations in Section 6.1, is estimated from unlabeled sample points $\{x'_m\}_{m=1}^{M'}$ as Eq.(37). The simulation is performed with the same setting as Section 6.1 but U is estimated with M' unlabeled sample points. M' is changed as $M' =$ 500, 250, 100, and 50. Figure 8 displays the simulation results when $(M, \sigma^2) = (500, 0.6)$. The simulation results show that the good performance of SIC is maintained as the number M' of unlabeled sample points is small. This implies that SIC will work well if only a rough estimate of U is available.

6.5 Unrealizable learning target function

Finally, we consider the case when the learning target function $f(x)$ is not included in H. The simulation is performed with the same setting as Section 6.1 but the learning target function $f(x)$ is the step function or $\frac{1}{1+x^2}$ defined on $[-\pi, \pi]$. Let the number M of training examples be 100. We decide the function space H following Section 5.2. Since $M = 100$, the dimension of H should be less than 100. Moreover, as shown in Section 6.3, the dimension of H should not be close to M. For this reason, we adopt θ_{20} as H. Let the set $\mathcal{M}$ of model candidates be $\{\theta_0, \theta_2, \theta_4, \ldots, \theta_{20}\}$, which are included in H. We measure the error of a learning result function $\hat{f}_{\theta_n}(x)$ from 1000 future sample points $\{u_j\}_{j=1}^{1000}$ randomly generated in $[-\pi, \pi]$ as

$$\mathrm{Error}[\theta_n] = \frac{1}{1000} \sum_{j=1}^{1000} \left| \hat{f}_{\theta_n}(u_j) - f(u_j) \right|^2 . \tag{54}$$

The simulation results with the learning target function being the step function for $(M, \sigma^2) = (100, 0.1)$ are displayed in Figure 9. The simulation results with $f(x) = \frac{1}{1+x^2}$ for $(M, \sigma^2) = (100, 0.03)$ are displayed in Figure 10. These results show that SIC seems still effective even in unrealizable cases as long as the Hilbert space H approximately includes the learning target function $f(x)$. However, further experiments may be needed to confirm the robustness against unrealizable learning target functions.

7 Conclusion

In this paper, we theoretically and experimentally evaluated the effectiveness of the model selection criterion called the subspace information criterion (SIC) in comparison with the traditional leave-one-out cross-validation (CV), Mallows's C_P, Akaike's information criterion (AIC), Sugiura's corrected AIC (cAIC), Schwarz's Bayesian information criterion (BIC), Rissanen's minimum description length criterion (MDL), and Vapnik's measure (VM). Theoretical evaluation included the comparison of the generalization measure, approximation method, and restriction on model candidates and learning methods. Experimentally, SIC was shown to outperform existing techniques especially when the number of training examples is small and the noise variance is large.

Despite the outstanding performance of SIC, its variance is rather large when the number of training examples is small. Further work is needed to investigate the variance. The simulation results showed that SIC seems still effective even in unrealizable cases as long as the functional Hilbert space approximately includes the learning target function. However, further experiments may be necessary for confirming the robustness. The concept of SIC given in Section 3 is valid even when the number of training examples is not larger than the dimension of the functional Hilbert space which includes the learning target function. Devising a practical calculation method of SIC for such a case is also important future work.

Acknowledgement

The authors would like to thank Dr. Hironori Fujisawa, Dr. Shun-ichi Amari, Dr. Koji Tsuda, Dr. Klaus-Robert Müller, Dr. Joachim M. Buhmann, and anonymous reviewers for their valuable comments.

References

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.

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

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

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.

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.

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.

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.

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

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics, 22*, 79–86.

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 a subset regression. 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.

Mikhal'skii, A. I. (1987). Choosing an algorithm of estimation based on samples of limited size. *Automation and Remote Control, 48*, 909–918.

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.

Orr, M. J. L. (1996). Introduction to radial basis function networks. *Technical Report, Center for Cognitive Science, University of Edinburgh*, Edinburgh, Scotland. (available from http://www.anc.ed.ac.uk/ mjo/papers/intro.ps.gz)

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.

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.

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

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

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.

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.

Sugiyama, M., & Ogawa, H. (1999). Functional analytic approach to model selection — Subspace information criterion. In *Proceedings of 1999 Workshop on Information-Based Induction Sciences (IBIS'99)* (pp. 93–98). Izu, Japan.

Sugiyama, M., & Ogawa, H. (2000a). Another look at C_P and network information criterion as approximation of subspace information criterion. *Technical Report TR00-0012, Department of Computer Science, Tokyo Institute of Technology*, Tokyo, Japan. (available from ftp://ftp.cs.titech.ac.jp/pub/TR/00/TR00-0012.pdf)

Sugiyama, M., & Ogawa, H. (2000b). Optimal design of regularization term and regularization parameter by subspace information criterion. *Technical Report TR00-0013, Department of Computer Science, Tokyo Institute of Technology*, Tokyo, Japan. (available from ftp://ftp.cs.titech.ac.jp/pub/TR/00/TR00-0013.pdf)

Sugiyama, M., & Ogawa, H. (2001). Subspace information criterion for model selection. *Neural Computation, 13(4).* (to appear)

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.

Tsuda, K., Sugiyama, M, & Müller, K.-R. (2000). Subspace information criterion for non-quadratic regularizers. *Technical Report, 120, GMD FIRST*, Berlin, Germany.

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

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

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

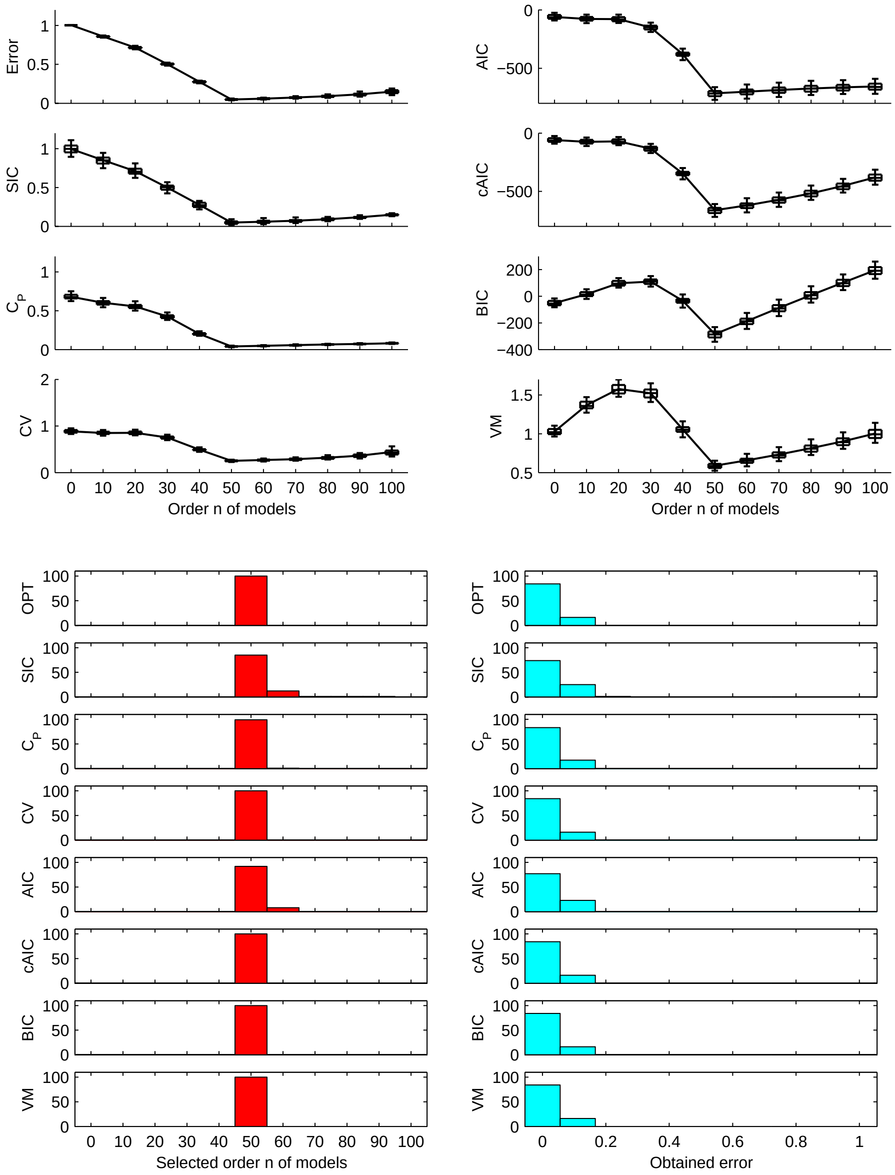

Figure 1: Simulation results when $(M, \sigma^2) = (500, 0.2)$.

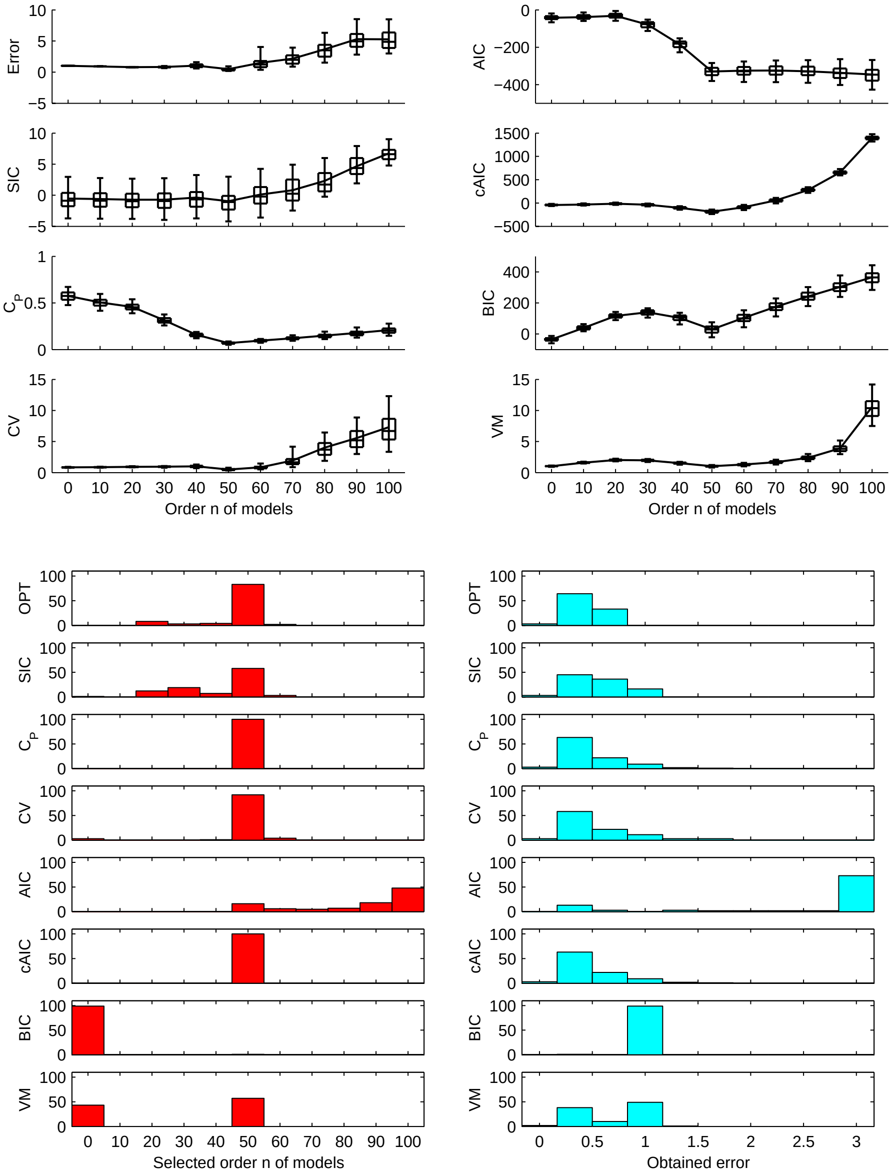

Figure 2: Simulation results when $(M, \sigma^2) = (250, 0.2)$.

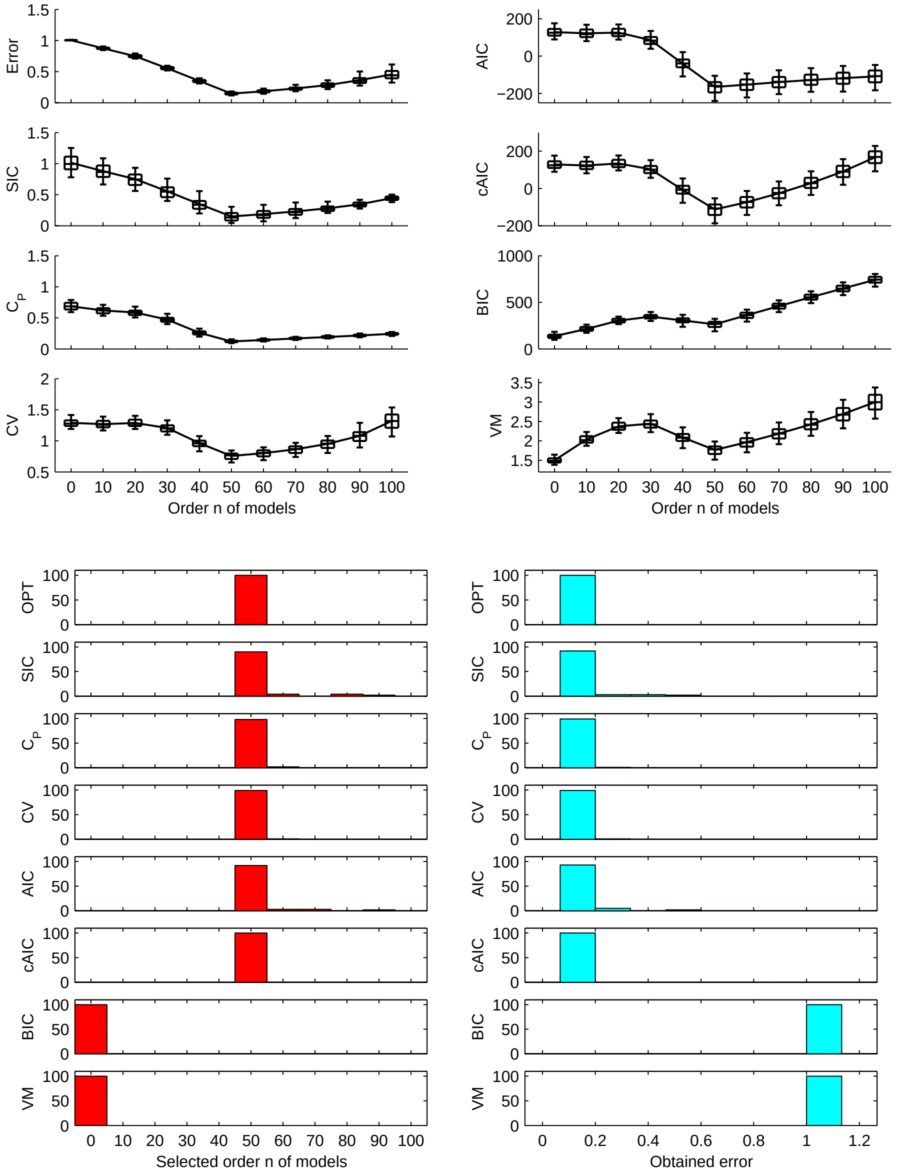

Figure 3: Simulation results when $(M, \sigma^2) = (500, 0.6)$.

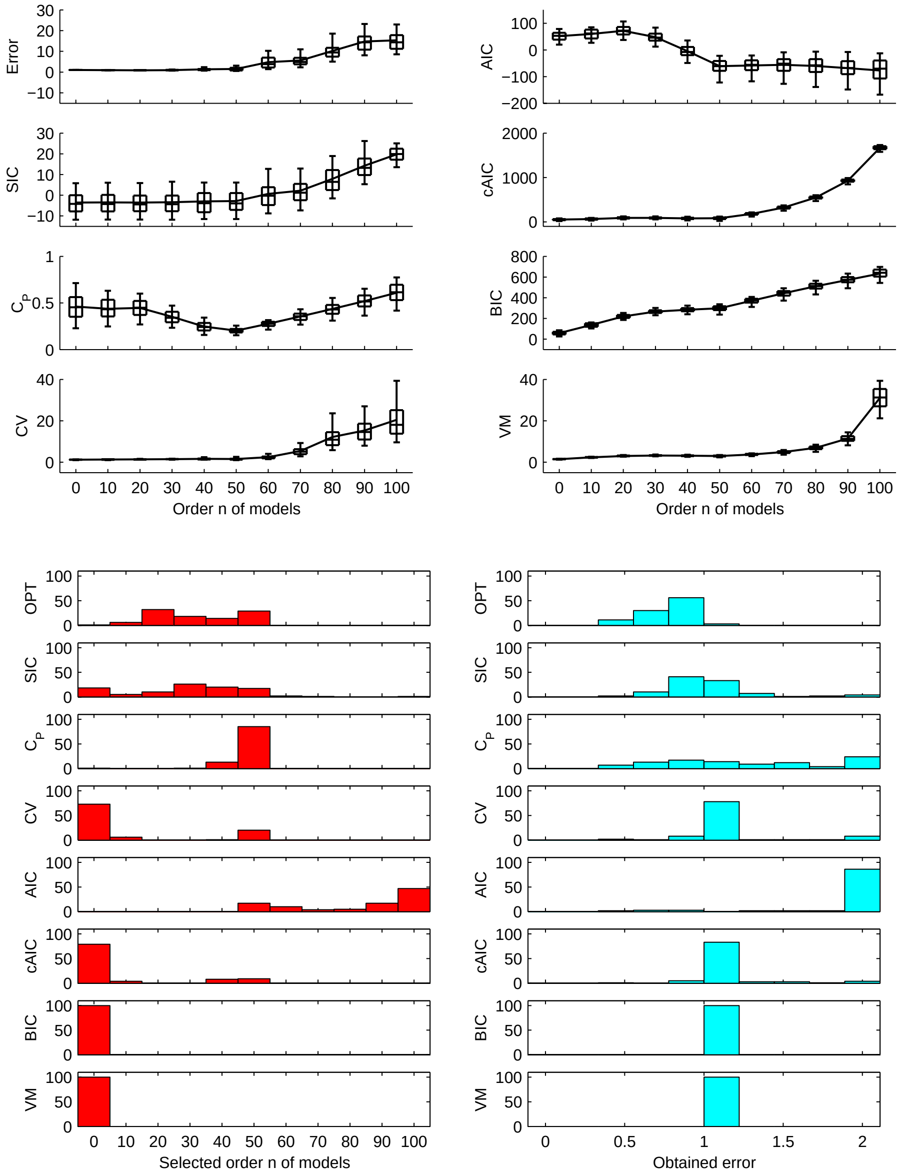

Figure 4: Simulation results when $(M, \sigma^2) = (250, 0.6)$.

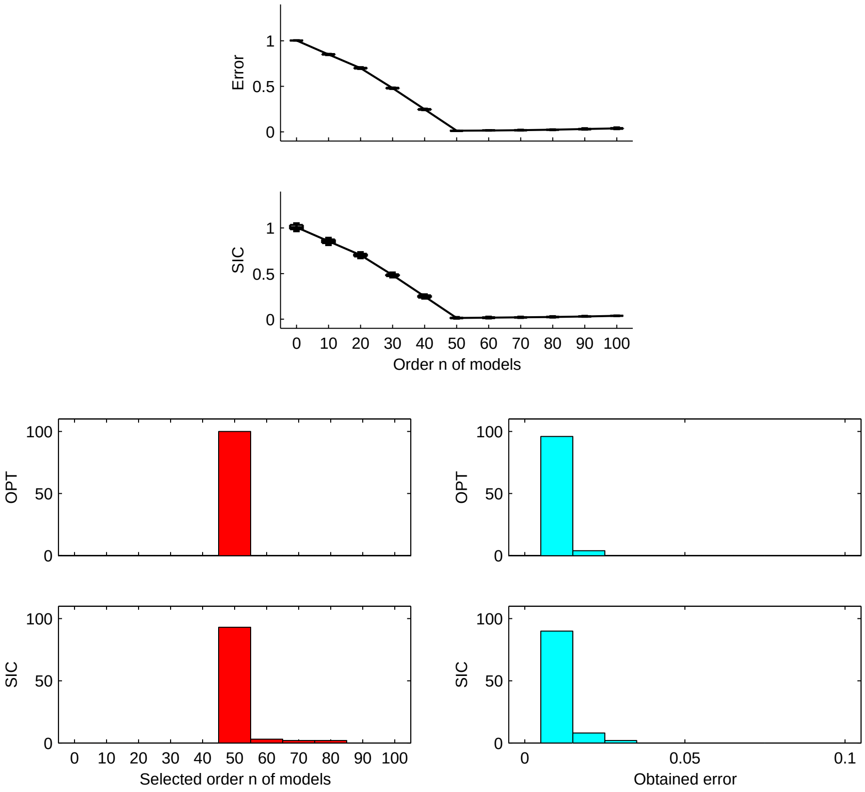

Figure 5: Simulation results when $M = 500$ with uniform noise.

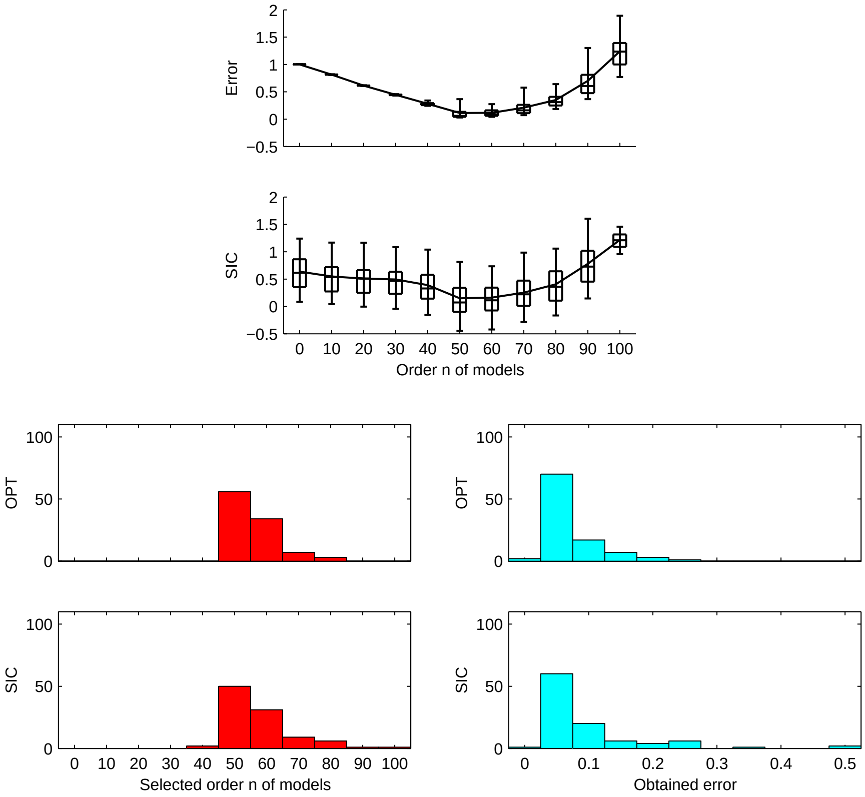

Figure 6: Simulation results when $M = 250$ with uniform noise.

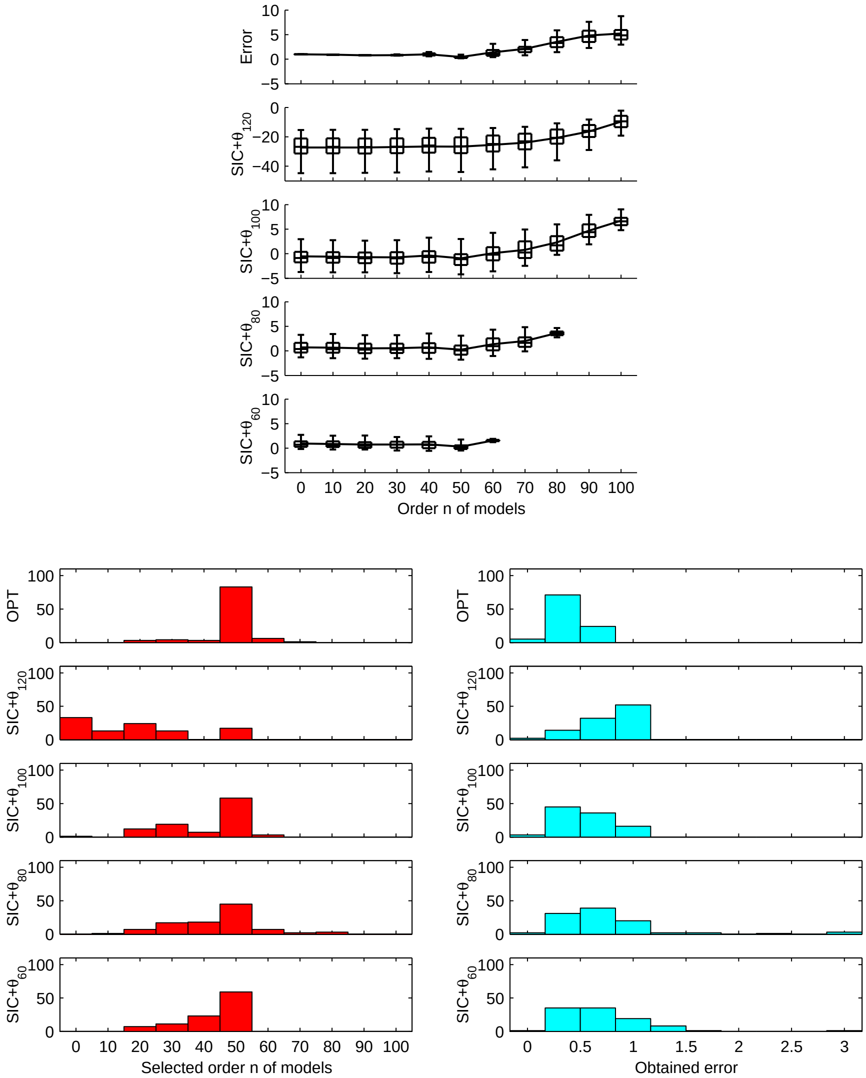

Figure 7: Simulation results when $(M, \sigma^2) = (250, 0.2)$ with changing H. 'SIC+θ_j' denotes the case when SIC is calculated with H specified by θ_j.

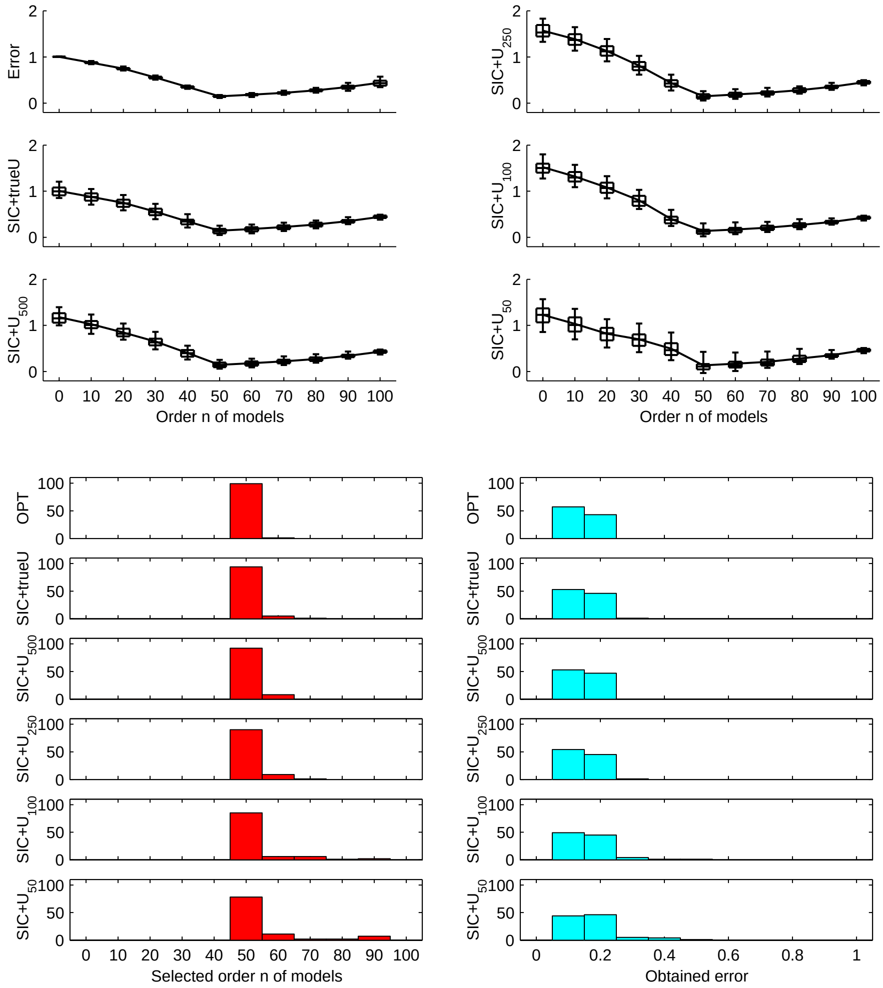

Figure 8: Simulation results when $(M, \sigma^2) = (500, 0.6)$ with the covariance matrix U (see Eq.(21)) estimated from unlabeled sample points as Eq.(37). ‘SIC+true$U$’ denotes the case when SIC is calculated with the true covariance matrix U. ‘SIC+U_j’ denotes the case when SIC is calculated with the covariance matrix U estimated with j unlabeled sample points.

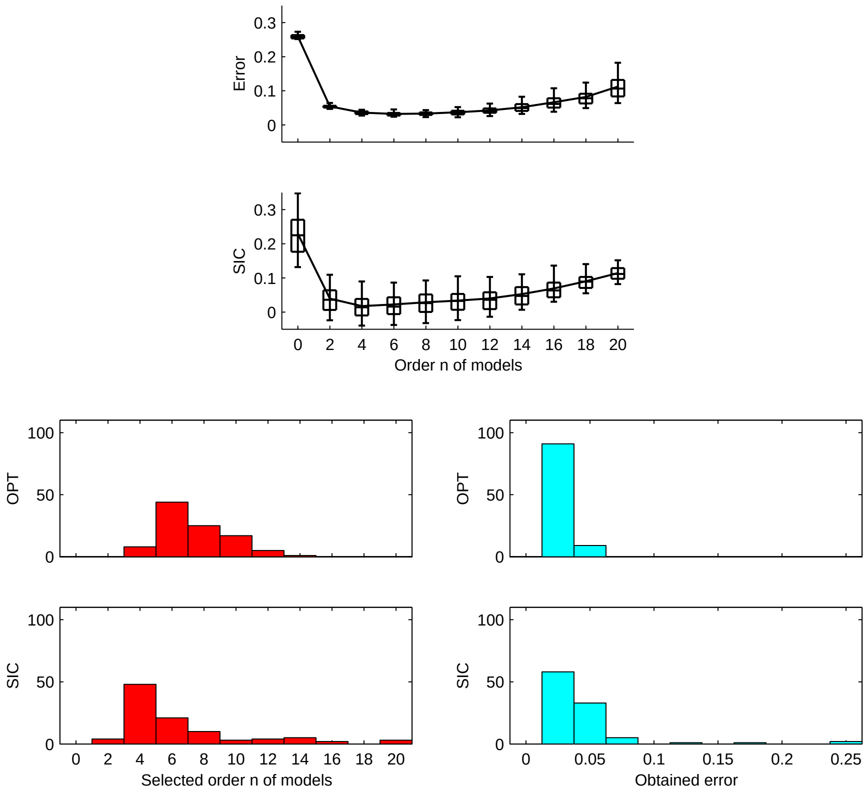

Figure 9: Simulation results when $(M, \sigma^2) = (100, 0.1)$ with unrealizable learning target function: $f(x)$ is the step function.

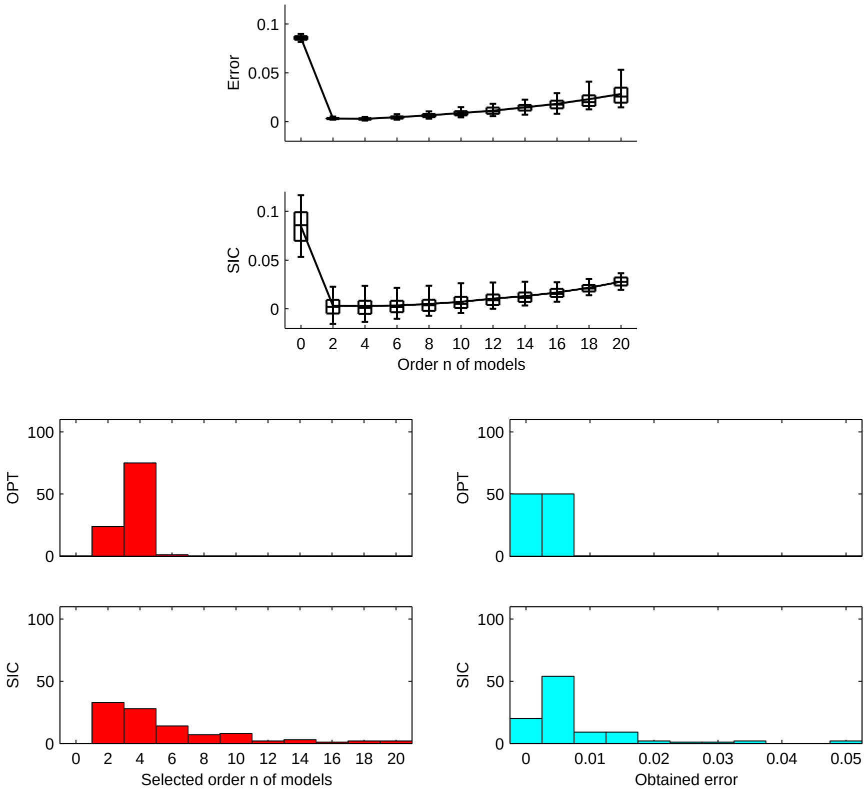

Figure 10: Simulation results when $(M, \sigma^2) = (100, 0.03)$ with unrealizable learning target function: $f(x) = \frac{1}{1+x^2}$.