Model Selection with Small Samples[1]

Masashi Sugiyama* [2], Hidemitsu Ogawa*

*Department of Computer Science, Tokyo Institute of Technology, Tokyo, Japan.

Abstract

Recently, a new model selection criterion called the subspace information criterion (SIC) was proposed. SIC 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. Theoretical evaluation includes the comparison of the generalization measure, approximation method, and restriction on model candidates and learning methods. The simulations show that SIC outperforms existing techniques especially when the number of training examples is small and the noise variance is large.

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*, which indicates, for example, the number and type 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 from various standpoints: information statistics [1][2][3][4], Bayesian statistics [5], stochastic complexity [6], and structural risk minimization principle [7][8].

Recently, a new model selection criterion called the *subspace information criterion* (SIC) was proposed by the authors [9]. SIC 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.

2 Subspace information criterion (SIC) for subset regression

Let us consider the regression problem of obtaining, from a set of M training examples, an approximation to a target function $f(x)$ of L variables defined on $\mathcal{D} \subset \mathbf{R}^L$. The training examples are made up of *sample points* $x_m \in \mathcal{D}$ and corresponding *sample values* $y_m \in \mathbf{R}$. We suppose that y_m is degraded by additive noise ϵ_m, i.e., $y_m = f(x_m) + \epsilon_m$.

Let θ be a set of factors which determine 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 measure the generalization error of $\hat{f}_\theta(x)$ by

$$J_G[\theta] = \mathrm{E}_\epsilon \int (\hat{f}_\theta(u) - f(u))^2 p(u) du, \qquad (1)$$

where E_ϵ denotes the ensemble average over the noise and $p(\cdot)$ is the probability density function of future (test) input points u. Then the problem of model selection considered in this paper is to select, from a set $\mathcal{M}$ of model candidates, the best model $\hat{\theta}$ that minimizes the generalization error J_G.

The model selection criterion called the *subspace information criterion* (SIC) [9] gives an unbiased estimate of the generalization error J_G. Here, we briefly review SIC for subset regression.

The following conditions are assumed:

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

(b) The $M \times \mu$-dimensional *design matrix* A with (m,p)-th element being $\varphi_p(x_m)$ has the rank μ.

(c) The number M of training examples is larger than the number μ of basis functions.

(d) The mean noise is zero and the noise covariance matrix is given as $\sigma^2 I_M$ where $\sigma^2 > 0$ and I_M is the M-dimensional identity matrix.

(e) The μ-dimensional covariance matrix U with (p,p')-element being $\int \varphi_{p'}(u)\varphi_p(u)p(u)du$ is known.

(f) A model θ indicates a subset of indices $\{1, 2, \ldots, \mu\}$, and the learning result function $\hat{f}_\theta(x)$ is defined as a minimizer of the training error $\frac{1}{M}\sum_{m=1}^{M}(\hat{f}(x_m) - y_m)^2$ in a subspace spanned by

[1] A detailed version of this article is to appear in *Machine Learning, Special Issue on New Methods for Model Selection and Model Combination*, 2001.

[2] E-mail: sugi@og.cs.titech.ac.jp, URL: http://ogawa-www.cs.titech.ac.jp/~sugi, FAX: +81-3-5734-2949.

$\{\varphi_p(x)\}_{p\in\theta}$. In this case, $\hat{f}_\theta(x)$ is given as

$$\hat{f}_\theta(x) = \textstyle\sum_{p\in\theta}[A_\theta^\dagger y]_p \varphi_p(x), \tag{2}$$

where A_θ is an $M \times \mu$ matrix with (m,p)-th element being $\varphi_p(x_m)$ if $p \in \theta$ otherwise 0. $A_\theta^\dagger$ denotes the *Moore-Penrose generalized inverse* of A_θ and $[\cdot]_p$ denotes the p-th element of a vector.

Under the above assumptions, SIC is given as

$$\begin{aligned}\mathrm{SIC}[\theta] = {} & \langle U(A_\theta^\dagger - A^\dagger)y, (A_\theta^\dagger - A^\dagger)y\rangle \\ & - \hat{\sigma}^2 \mathrm{tr}(U(A_\theta^\dagger - A^\dagger)(A_\theta^\dagger - A^\dagger)^\top) \\ & + \hat{\sigma}^2 \mathrm{tr}(UA_\theta^\dagger(A_\theta^\dagger)^\top),\end{aligned} \tag{3}$$

where $\hat{\sigma}^2 = \langle y - AA^\dagger y, y\rangle/(M-\mu)$ and $\top$ denotes the transpose of a matrix. It is shown that SIC is an unbiased estimate of J_G [9]:

$$\mathrm{E}_\epsilon \ \mathrm{SIC}[\theta] = J_G[\theta]. \tag{4}$$

3 Theoretical evaluation of SIC

In this section, SIC is compared with the traditional leave-one-out cross-validation (CV), Mallows's C_P [10], Akaike's information criterion (AIC) [1], Sugiura's corrected AIC (cAIC) [2], Schwarz's Bayesian information criterion (BIC) [5], Rissanen's minimum description length criterion (MDL) [6], and Vapnik's measure (VM) [8].

3.1 Generalization measure

SIC can adopt any generalization measure expressed as $\mathrm{E}_\epsilon\|\hat{f}_\theta - f\|^2$ as long as it is computable (e.g. the covariance matrix U is known). $\|\cdot\|$ denotes the norm in the functional Hilbert space spanned by $\{\varphi_p(x)\}_{p=1}^{\mu}$. The derivatives of the functions $\hat{f}_\theta(x)$ and $f(x)$ can also be included in the generalization measure (with the Sobolev norm).

C_P adopts the predictive training error $\frac{1}{M}\mathrm{E}_\epsilon\sum_{m=1}^{M}(\hat{f}_\theta(x_m) - f(x_m))^2$ as the error measure, which is equivalent to Eq.(1) with $p(u)$ being replaced by the empirical distribution. Note that the predictive training error does not evaluate the error at future sample points u.

CV adopts the so-called leave-one-out error $\frac{1}{M}\sum_{m=1}^{M}(\hat{f}_\theta^{(m)}(x_m) - y_m)^2$ as the error measure, where $\hat{f}_\theta^{(m)}$ denotes the learning result function obtained with the training examples without (x_m, y_m). The leave-one-out error also does not directly evaluate the error at future sample points u. The relation between the leave-one-out error and Eq.(1) is not well recognized yet.

AIC and cAIC adopt the expected Kullback-Leibler information over all possible training sets $\{(x_m, y_m)\}_{m=1}^{M}$ as the generalization measure, which is conceptually similar to the expectation of Eq.(1) over training sample points $\{x_m\}_{m=1}^{M}$ [3]. 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, the purpose of model selection is to obtain the model that gives good learning result functions on average. In contrast, if one adopts the generalization measure which is *not* averaged over training sample points, the purpose of model selection is to obtain the model that gives the optimal learning result function from a given, particular training set. This implies that the latter standpoint is suitable for acquiring the best prediction performance from given training examples.

BIC gives an estimate of the posterior probability of parameters, and MDL gives an estimate of the description length of the model and data. The relation between the posterior probability, description length of the model and data, and generalization error is not clear.

The generalization measure of VM is a probabilistic upper bound of the risk functional $\int(\hat{f}_\theta(u) - f(u))^2 p(u)du$, 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.

3.2 Approximation methods

C_P, AIC, cAIC, BIC, MDL, and VM are expressed with the training error $\frac{1}{M}\sum_{m=1}^{M}(\hat{f}_\theta(x_m) - y_m)^2$. In contrast, CV and SIC directly evaluate the error measures.

C_P is an unbiased estimate of the predictive training error with finite samples. Since the predictive training error asymptotically agrees with the generalization error Eq.(1) if training sample points $\{x_m\}_{m=1}^{M}$ are subject to $p(\cdot)$, it can be regarded as an approximation of Eq.(1). Although asymptotic optimality of C_P is shown, its effectiveness with small samples is not theoretically sure.

In CV, the leave-one-out error 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. 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 models (i.e., models which include the learning target function). However, its performance for unfaithful models is not sure. VM gives a probabilistic upper bound of the risk functional based on the VC theory [7]. 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. In order to calculate SIC, rather restrictive conditions should be assumed (see Sec. 2). However, these conditions do not have to be rigorously satisfied in practice. For example, when basis functions $\{\varphi_p(x)\}_{p=1}^{\mu}$ which include the learning target function $f(x)$ are unknown (see Assumption (a) in Sec .2), basis functions $\{\varphi_p(x)'\}_{p=1}^{\mu'}$ with the following properties are practically adopted:

(i) $\{\varphi_p(x)'\}_{p=1}^{\mu'}$ approximately include the learning target function $f(x)$.

(ii) The number μ' of basis functions is less than the number M of training examples.

When the covariance matrix U (see Assumption (e) in Sec. 2) is unknown, it can be estimated by using unlabeled sample points $\{x'_m\}_{m=1}^{M'}$ (i.e., sample points without sample values $\{y'_m\}_{m=1}^{M'}$) as $[\hat{U}]_{p,p'} = \frac{1}{M'}\sum_{m=1}^{M'}\varphi_{p'}(x'_m)\varphi_p(x'_m)$. 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 [9]).

3.3 Restriction on model candidates

AIC and cAIC are valid only when model candidates in the set $\mathcal{M}$ are nested [11][3], 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 models.

3.4 Restriction on learning methods

AIC, cAIC, BIC, and MDL are specialized for maximum likelihood estimation. A generalized AIC [3][4] relaxed the restriction of maximum likelihood estimation. C_P is specialized for the training error minimization learning with linear regression models. An extension of C_P called C_L [10], VM, and SIC are applicable to various learning methods expressed by linear mapping ($A_\theta^\dagger$ in Eq.(2)), including regularization learning with quadratic regularizers (ridge regression). Note that in VM, the VC-dimension [7] of models should be explicitly calculated.

4 Experimental evaluation of SIC

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

Let the learning target function $f(x)$ be $f(x) = \frac{1}{10}\sum_{p=1}^{50}(\sin px + \cos px)$ defined on $[-\pi, \pi]$. Let us consider a set of 201 basis functions $\{1, \sin px, \cos px\}_{p=1}^{100}$ which includes $f(x)$. Let the set $\mathcal{M}$ of model candidates be $\mathcal{M} = \{\theta_0, \theta_{10}, \theta_{20}, \ldots, \theta_{100}\}$, where θ_n indicates a regression model with $\{1, \sin px, \cos px\}_{p=1}^{n}$. 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. We compare SIC, CV, C_P, AIC, cAIC, BIC (which is the same as MDL), and VM. Note that the covariance matrix U (see Assumption (e) in Sec. 2) is the identity matrix in the above setting. We shall measure the error of a learning result function $\hat{f}_{\theta_n}(x)$ by $\frac{1}{2\pi}\int_{-\pi}^{\pi}(\hat{f}_{\theta_n}(x) - f(x))^2 dx$.

The simulation is performed 100 times with changing the noise $\{\epsilon_m\}_{m=1}^{M}$ in each trial. Fig. 1 show the distributions of the selected order n of models (upper) and error obtained by the selected model (lower) by 100 trials. 'OPT' indicates the optimal model that minimizes the error. When $(M, \sigma^2) = (500, 0.2)$, all model selection criteria work well. When $(M, \sigma^2) = (250, 0.2)$, 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 $(M, \sigma^2) = (500, 0.6)$, 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

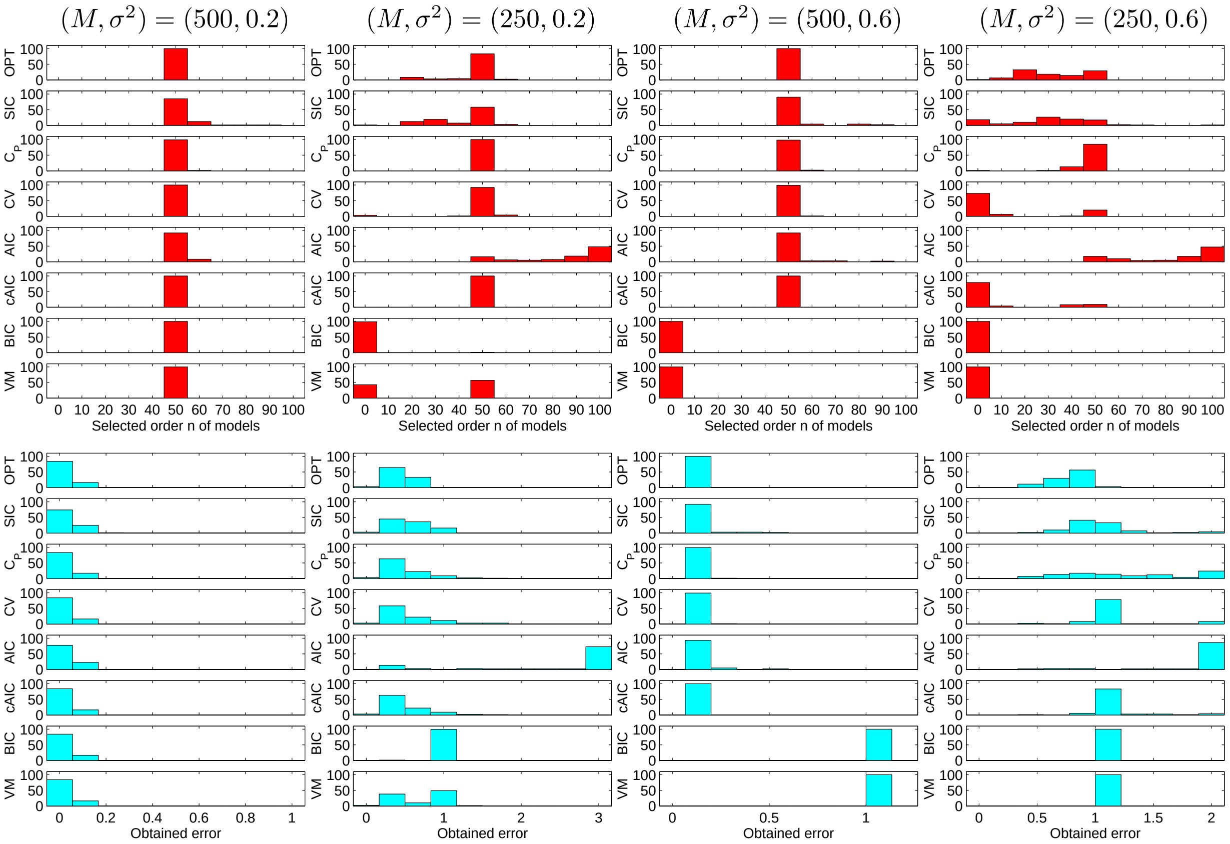

Fig. 1. Distributions of the selected order and error by 100 trials.

$(M, \sigma^2) = (250, 0.6)$, 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. 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.

References

[1] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. AC-19, no. 6, pp. 716–723, 1974.

[2] N. Sugiura, "Further analysis of the data by Akaike's information criterion and the finite corrections," *Communications in Statistics. Theory and Methods*, vol. 7, no. 1, pp. 13–26, 1978.

[3] N. Murata, S. Yoshizawa, and S. Amari, "Network information criterion—determining the number of hidden units for an artificial neural network model," *IEEE Transactions on Neural Networks*, vol. 5, no. 6, pp. 865–872, 1994.

[4] S. Konishi and G. Kitagawa, "Generalized information criterion in model selection," *Biometrika*, vol. 83, pp. 875–890, 1996.

[5] G. Schwarz, "Estimating the dimension of a model," *Annals of Statistics*, vol. 6, pp. 461–464, 1978.

[6] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, pp. 465–471, 1978.

[7] V. N. Vapnik, *The Nature of Statistical Learning Theory*. Berlin: Springer-Verlag, 1995.

[8] V. Cherkassky, X. Shao, F. M. Mulier, and V. N. Vapnik, "Model complexity control for regression using VC generalization bounds," *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 1075–1089, 1999.

[9] M. Sugiyama and H. Ogawa, "Subspace information criterion for model selection," *Neural Computation*, 2001. (to appear).

[10] C. L. Mallows, "Some comments on C_P," *Technometrics*, vol. 15, no. 4, pp. 661–675, 1973.

[11] K. Takeuchi, "On the selection of statistical models by AIC," *Journal of the Society of Instrument and Control Engineering*, vol. 22, no. 5, pp. 445–453, 1983. (in Japanese).