Optimal Design of Regularization Term and Regularization Parameter by 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

The problem of designing the regularization term and regularization parameter for linear regression models is discussed. Previously, we derived an approximation to the generalization error called the subspace information criterion (SIC), which is an unbiased estimator of the generalization error with finite samples under certain conditions. In this paper, we apply SIC to regularization learning and use it for (a) choosing the optimal regularization term and regularization parameter from given candidates, and (b) obtaining the closed form of the optimal regularization parameter for a fixed regularization term. The effectiveness of SIC is demonstrated through computer simulations with artificial and real data.

Keywords

supervised learning, generalization error, linear regression, regularization learning, ridge regression, model selection, regularization parameter, subspace information criterion

Nomenclature

$f(\boldsymbol{x})$	:	learning target function
$\mathcal{D}$	:	domain of $f(\boldsymbol{x})$
$\boldsymbol{x}_m$	:	m-th sample point
y_m	:	m-th sample value
ϵ_m	:	m-th noise
$(\boldsymbol{x}_m, y_m)$	:	m-th training example
M	:	the number of training examples
$\boldsymbol{y}$	:	M-dimensional vector consisting of $\{y_m\}_{m=1}^M$
$\boldsymbol{\epsilon}$	:	M-dimensional vector consisting of $\{\epsilon_m\}_{m=1}^M$
$\varphi_p(\boldsymbol{x})$	:	p-th basis function
θ_p	:	p-th coefficient
μ	:	the number of basis functions
J_G	:	generalization error
J_{TE}	:	training error
J_R	:	regularized training error
T	:	regularization matrix
α	:	regularization parameter
A	:	design matrix
$X_{T,\alpha}$	:	regularization learning matrix
U	:	μ-dimensional matrix
$\boldsymbol{\theta}$	:	true parameter
$\hat{\boldsymbol{\theta}}_{T,\alpha}$	:	regularization estimate
$\hat{\boldsymbol{\theta}}_u$	:	unbiased estimate
σ^2	:	noise variance

1 Introduction

The purpose of supervised learning is acquiring a higher level of the generalization capability. *Least mean squares* (LMS) learning, aimed at minimizing the *training error*, is widely used for obtaining a learning result from training examples. However, LMS learning sometimes causes so-called *over-fitting* and yields a lower level of the generalization capability. To avoid over-fitting, *regularization learning* (which is referred to as *ridge regression* or *weight decay* in a special case) is often employed. Regularization learning is aimed at minimizing the weighted sum of the training error and a *regularization term*. The weight parameter is called the *regularization parameter*. Intuitively, the regularization term makes the learning result function *smooth* for avoiding over-fitting. A difficulty of using regularization learning is that the choice of the regularization term and regularization parameter (i.e., the type and level of smoothing) is crucial for acquiring a higher level of the generalization capability.

So far, research from various standpoints have been conducted for designing the regularization term and regularization parameter. One of the classic approaches is based on

the *discrepancy principle* (Groetsch, 1984; Morozov, 1993; Kunisch & Zou, 1998). The discrepancy principle asserts that the training error should be equal to the noise variance. A heuristic motivation for this principle is that it does not make sense to ask for an estimation with the training error less than the noise variance since only the noisy sample values are available (Groetsch, 1984). The traditional *cross-validation* is one of the well-known methods for determining the regularization term and regularization parameter. In k-fold cross-validation, training examples are divided into k disjoint sets. $k-1$ sets are used for obtaining a learning result function and the rest is used for evaluating its error. This procedure is repeated for all k combinations and the mean error is regarded as an estimate of the generalization error. If k is equal to the number of training examples, it is specially called the *leave-one-out cross-validation*. C_L (Mallows, 1973), which is also referred to as the *unbiased risk estimate* (Wahba, 1990), gives an unbiased estimate of the error between estimated and true values at training sample points. C_L requires an estimate of the noise variance. In contrast, the *generalized cross-validation* (Craven & Wahba, 1979), which is a simplified version of the leave-one-out cross-validation, can be calculated without estimating the noise variance. Within the framework of the Bayesian statistics, Akaike (1980) proposed a *Bayesian information criterion*. It is aimed at choosing the regularization term and regularization parameter so that the marginal likelihood of the regularization parameter is maximized (see also MacKay, 1992; Watanabe, 2001). Shibata (1989) extended *Akaike's information criterion* (Akaike, 1974), which evaluates the generalization error itself in the asymptotic sense, to choose the regularization parameter (see also, Murata *et al.*, 1994; Konishi & Kitagawa, 1996). From the viewpoint of the structural risk minimization principle (Vapnik, 1995), Cherkassky *et al.*, (1999) proposed *Vapnik's measure*, which is a probabilistic upper bound of the generalization error.

In this article, we apply the *subspace information criterion* (SIC) (Sugiyama & Ogawa, 2001, 2002) to the problem of designing the regularization term and regularization parameter. SIC gives an unbiased estimate of the generalization error with finite samples under certain conditions. SIC is used for (a) choosing the optimal regularization term and regularization parameter from given candidates, and (b) obtaining the closed form of the optimal regularization parameter for a fixed regularization term. The effectiveness of SIC is demonstrated through computer simulations with artificial and real data.

2 Problem formulation

In this section, the supervised learning problem is mathematically formulated.

2.1 Supervised learning

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

$\mathbb{R}$:

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

where y_m is degraded by unknown additive noise ϵ_m. We assume that ϵ_m is independently drawn from a distribution with mean zero and variance σ^2.

For the time being, we assume that the unknown learning target function $f(\boldsymbol{x})$ can be expressed by a linear combination of μ specified basis functions $\{\varphi_p(\boldsymbol{x})\}_{p=1}^\mu$:

$$f(\boldsymbol{x}) = \sum_{p=1}^{\mu} \theta_p \varphi_p(\boldsymbol{x}), \tag{2}$$

where $\{\theta_p\}_{p=1}^\mu$ are unknown. The assumption that $f(\boldsymbol{x})$ is realizable is removed in Section 3.2.3.

Let A be an $M \times \mu$ matrix with the (m, p)-th element being

$$A_{m,p} = \varphi_p(\boldsymbol{x}_m). \tag{3}$$

A is called the *design matrix* and it extracts the values of a function at sample points $\{\boldsymbol{x}_m\}_{m=1}^M$:

$$A\boldsymbol{\theta} = (f(\boldsymbol{x}_1), f(\boldsymbol{x}_2), \ldots, f(\boldsymbol{x}_M))^\top, \tag{4}$$

where $\top$ denotes the transpose of a vector (or matrix) and $\boldsymbol{\theta}$ is the μ-dimensional vector defined by

$$\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_\mu)^\top. \tag{5}$$

We assume that the rank of A is μ. The condition holds only if $M \geq \mu$.

Let $\boldsymbol{y}$ and $\boldsymbol{\epsilon}$ be M-dimensional vectors defined by

$$\boldsymbol{y} = (y_1, y_2, \ldots, y_M)^\top, \tag{6}$$

$$\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_M)^\top. \tag{7}$$

Then the relationship between $\boldsymbol{\theta}$ and $\boldsymbol{y}$ can be expressed as

$$\boldsymbol{y} = A\boldsymbol{\theta} + \boldsymbol{\epsilon}. \tag{8}$$

Let X be a matrix that gives a learning result function $\hat{f}(\boldsymbol{x})$:

$$\hat{\boldsymbol{\theta}} = X\boldsymbol{y}, \tag{9}$$

$$\hat{f}(\boldsymbol{x}) = \sum_{p=1}^{\mu} \hat{\theta}_p \varphi_p(\boldsymbol{x}). \tag{10}$$

X is called a *learning matrix*.

The purpose of supervised learning is to obtain the optimal approximation $\hat{f}(\boldsymbol{x})$ that minimizes a *generalization error* J_G. In this article, we define J_G by

$$J_G = \mathrm{E}_{\boldsymbol{\epsilon}} \int_{\mathcal{D}} \left(\hat{f}(\boldsymbol{x}) - f(\boldsymbol{x})\right)^2 q(\boldsymbol{x}) d\boldsymbol{x}, \tag{11}$$

where $\mathrm{E}_{\boldsymbol{\epsilon}}$ denotes the expectation over the noise $\boldsymbol{\epsilon}$ and $q(\boldsymbol{x})$ denotes the probability density function of future (test) input points. In Eq.(11), we suppose that $\hat{f}$ is obtained from noisy training examples $\{(\boldsymbol{x}_m, y_m)\}_{m=1}^M$, i.e., it depends on the noise.

2.2 Regularization learning

In general, the learning matrix X is obtained on the basis of a learning criterion. *Least mean squares* (LMS) learning is often used as the learning criterion.

Definition 1 (Least mean squares learning) *A matrix X is called the LMS learning matrix if X minimizes the training error J_{TE}:*

$$J_{TE} = \frac{1}{M}\sum_{m=1}^{M}\left(\hat{f}(\boldsymbol{x}_m) - y_m\right)^2. \tag{12}$$

LMS learning sometimes causes over-fitting and yields a lower level of the generalization capability. To avoid over-fitting, we will use *regularization learning*:

Definition 2 (Regularization learning) *Let T be a $\mu' \times \mu$ matrix and α be a positive constant. A matrix X is called the regularization learning matrix if X minimizes the regularized training error J_R:*

$$J_R = \sum_{m=1}^{M}\left(\hat{f}(\boldsymbol{x}_m) - y_m\right)^2 + \alpha\|T\hat{\boldsymbol{\theta}}\|^2. \tag{13}$$

T is called the *regularization matrix* and α is called the *regularization parameter*. $\|T\hat{\boldsymbol{\theta}}\|^2$ is called the *regularization term*. When T is the identity matrix, regularization learning is called *ridge regression* or *weight decay*.

Let $X_{T,\alpha}$ be the regularization learning matrix obtained with T and α. $X_{T,\alpha}$ is expressed as

$$X_{T,\alpha} = (B + \alpha T^\top T)^{-1}A^\top, \tag{14}$$

where B is defined by

$$B = A^\top A. \tag{15}$$

A difficulty of using regularization learning is that the choice of T and α is crucial for acquiring a higher level of the generalization capability. In the following sections, we will discuss the problem of designing T and α for optimal generalization.

3 Subspace information criterion for regularization learning

Since the generalization error J_G defined by Eq.(11) includes the unknown learning target function $f(\boldsymbol{x})$, it is not possible to directly optimize the regularization matrix T and regularization parameter α. In this section, we give an unbiased estimator of J_G called the *subspace information criterion* (SIC) for regularization learning.

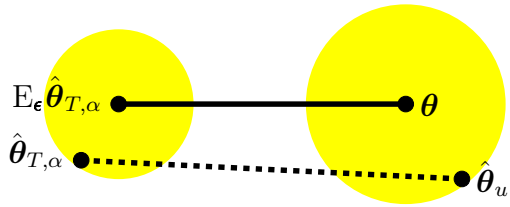


Figure 1: Intuitive idea of SIC. The solid line denotes the bias of $\hat{\boldsymbol{\theta}}_{T,\alpha}$. It can be roughly estimated by the dotted line, which can be calculated.

3.1 Subspace information criterion

Let U be a μ-dimensional matrix with the (p,p')-th element being

$$U_{p,p'} = \int_{\mathcal{D}} \varphi_p(\boldsymbol{x})\varphi_{p'}(\boldsymbol{x})q(\boldsymbol{x})d\boldsymbol{x}. \tag{16}$$

Then the generalization error of $\hat{f}_{T,\alpha}(\boldsymbol{x})$ is expressed as

$$J_G[T,\alpha] = \mathrm{E}_{\boldsymbol{\epsilon}}\|\hat{\boldsymbol{\theta}}_{T,\alpha} - \boldsymbol{\theta}\|_U^2, \tag{17}$$

where $\|\cdot\|_U^2$ is the norm weighted by U:

$$\|\boldsymbol{\theta}\|_U^2 = \boldsymbol{\theta}^\top U\boldsymbol{\theta}. \tag{18}$$

It is known that J_G can be decomposed into the *bias* and *variance*:

$$J_G[T,\alpha] = \|\mathrm{E}_{\boldsymbol{\epsilon}}\hat{\boldsymbol{\theta}}_{T,\alpha} - \boldsymbol{\theta}\|_U^2 + \sigma^2\mathrm{tr}\left(UX_{T,\alpha}X_{T,\alpha}^\top\right). \tag{19}$$

Let X_u be a learning matrix that gives an unbiased estimate $\hat{\boldsymbol{\theta}}_u$ of the true $\boldsymbol{\theta}$:

$$\hat{\boldsymbol{\theta}}_u = X_u\boldsymbol{y}, \tag{20}$$
$$\mathrm{E}_{\boldsymbol{\epsilon}}\hat{\boldsymbol{\theta}}_u = \boldsymbol{\theta}. \tag{21}$$

In the present setting, X_u is given by

$$X_u = B^{-1}A^\top. \tag{22}$$

For the time being, we assume that the correlation matrix U and the noise variance σ^2 are available. We will discuss their estimation methods in Section 3.2.

An intuitive idea of SIC is that the bias term $\|\mathrm{E}_{\boldsymbol{\epsilon}}\hat{\boldsymbol{\theta}}_{T,\alpha} - \boldsymbol{\theta}\|_U^2$ can be roughly approximated by $\|\hat{\boldsymbol{\theta}}_{T,\alpha} - \hat{\boldsymbol{\theta}}_u\|_U^2$ (see Fig.1). However, $\|\hat{\boldsymbol{\theta}}_{T,\alpha} - \hat{\boldsymbol{\theta}}_u\|_U^2$ is larger than the bias term on average. According to Sugiyama and Ogawa (2001), an unbiased estimate of the bias term is given by

$$\|\hat{\boldsymbol{\theta}}_{T,\alpha} - \hat{\boldsymbol{\theta}}_u\|_U^2 - \sigma^2\mathrm{tr}\left(U(X_{T,\alpha} - X_u)(X_{T,\alpha} - X_u)^\top\right). \tag{23}$$

Then the subspace information criterion (SIC) is given as follows.

Definition 3 (Subspace information criterion for regularization learning) *The following functional SIC is called the subspace information criterion for regularization learning:*

$$SIC[T,\alpha] = \|\hat{\boldsymbol{\theta}}_{T,\alpha} - \hat{\boldsymbol{\theta}}_u\|_U^2 - \sigma^2 \mathrm{tr}\left(U(X_{T,\alpha} - X_u)(X_{T,\alpha} - X_u)^\top\right) + \sigma^2 \mathrm{tr}\left(U X_{T,\alpha} X_{T,\alpha}^\top\right). \quad (24)$$

The name *subspace information criterion* originated in Sugiyama and Ogawa (2001), where the criterion is used for selecting the subspace for LMS learning (see also Sugiyama and Ogawa (2002) for its evaluation).

The validity of SIC as an approximation to the generalization error J_G is theoretically substantiated by the following proposition.

Proposition 1 *(Sugiyama & Ogawa, 2001) For any T and α, SIC is an unbiased estimator of the generalization error J_G:*

$$\mathrm{E}_{\boldsymbol{\epsilon}} SIC[T,\alpha] = J_G[T,\alpha]. \quad (25)$$

3.2 SIC in practice

Although SIC does not include the unknown learning target function $f(\boldsymbol{x})$, it still includes factors which are sometimes unknown: the noise variance σ^2, the correlation matrix U defined by Eq.(16), and basis functions $\{\varphi_p(\boldsymbol{x})\}_{p=1}^{\mu}$ whose span includes $f(\boldsymbol{x})$. Here, we show their practical estimation methods.

3.2.1 Noise variance σ^2

If $M > \mu$, an unbiased estimate of σ^2 is given as follows (Fedorov, 1972).

$$\hat{\sigma}^2 = \frac{\|AB^{-1}A^\top \boldsymbol{y} - \boldsymbol{y}\|^2}{M - \mu}. \quad (26)$$

Note that the unbiasedness of SIC is still maintained (i.e. Proposition 1 holds) if σ^2 is estimated by Eq.(26). Also, the following estimate is often used in practice (Wahba, 1990):

$$\hat{\sigma}^2 = \frac{\|AX_{T,\alpha}\boldsymbol{y} - \boldsymbol{y}\|^2}{M - \mathrm{tr}\,(AX_{T,\alpha})}. \quad (27)$$

3.2.2 Correlation matrix U

The correlation matrix U is estimated in practice as follows.

(a) Unlabeled sample points: When unlabeled sample points $\{\boldsymbol{x}'_m\}_{m=1}^{M'}$ (i.e., sample points without sample values) are available, U is estimated by

$$\hat{U}_{p,p'} = \frac{1}{M'} \sum_{m=1}^{M'} \varphi_p(\boldsymbol{x}'_m)\varphi_{p'}(\boldsymbol{x}'_m). \quad (28)$$

Note that in some practical problems, a large number of unlabeled sample points are available (see e.g. Schuurmans & Southey, 2000), so U can be accurately estimated by Eq.(28). Furthermore, if $\{\boldsymbol{x}'_m\}_{m=1}^{M'}$ are the future test input points, SIC gives an unbiased estimate of the error at $\{\boldsymbol{x}'_m\}_{m=1}^{M'}$.

(b) Uniform distribution: When a bounded region $\mathcal{D}'$, from which future test sample points are drawn, is known, U may be estimated by

$$\hat{U}_{p,p'} = \frac{1}{|\mathcal{D}'|} \int_{\mathcal{D}'} \varphi_p(\boldsymbol{x}) \varphi_{p'}(\boldsymbol{x}) d\boldsymbol{x}. \tag{29}$$

(c) Empirical distribution: When no knowledge on future test input points is available, training sample points $\{\boldsymbol{x}_m\}_{m=1}^{M}$ may be used instead of unlabeled sample points $\{\boldsymbol{x}'_m\}_{m=1}^{M'}$ in Eq.(28). In this case, an estimate of U is given by

$$\hat{U} = \frac{1}{M} B. \tag{30}$$

Note that in this case, SIC essentially agrees with C_L proposed by Mallows (1973), which is a generalization of Mallows's C_P (Mallows, 1964). It is known that C_P is essentially equivalent to Akaike's information criterion (AIC) proposed by Akaike (1974) (see e.g. Konishi & Kitagawa, 1996). Similarly, for regularization learning, C_L is essentially equivalent to the criteria by Shibata (1989), Murata *et al.* (1994), and Konishi and Kitagawa (1996). Therefore, SIC can be regarded as an extension of these methods since additional information such as unlabeled sample points can be effectively utilized in SIC. Comparison of SIC to C_L and AIC is given in Sugiyama and Ogawa (2001), and the analysis can be carried over into the present paper.

(d) Vicinal distribution: The empirical distribution may be improved by

$$\hat{U}_{p,p'} = \frac{1}{M} \sum_{m=1}^{M} \int_{\mathcal{D}} \varphi_p(\boldsymbol{x}) \varphi_{p'}(\boldsymbol{x}) \phi(\boldsymbol{x}; \boldsymbol{x}_m) d\boldsymbol{x}, \tag{31}$$

where $\phi(\boldsymbol{x}; \boldsymbol{x}_m)$ is a probability density function centered on $\boldsymbol{x}_m$, e.g., the normalized Gaussian function with mean $\boldsymbol{x}_m$.

(e) Identity matrix: If you want to save the computational cost, U may be just estimated by

$$\hat{U} = I. \tag{32}$$

3.2.3 Basis functions $\{\varphi_p(\boldsymbol{x})\}_{p=1}^{\mu}$ whose span includes $f(\boldsymbol{x})$

Let us consider an unrealizable learning target function $f(\boldsymbol{x})$ given by

$$f(\boldsymbol{x}) = \sum_{p=1}^{\mu} \theta_p \varphi_p(\boldsymbol{x}) + g(\boldsymbol{x}), \tag{33}$$

where $g(\boldsymbol{x})$ is an unknown function. Without loss of generality, we assume that $g(\boldsymbol{x})$ is orthogonal to $\{\varphi_p(\boldsymbol{x})\}_{p=1}^{\mu}$:

$$\int_{\mathcal{D}} g(\boldsymbol{x})\varphi_p(\boldsymbol{x})q(\boldsymbol{x})d\boldsymbol{x} = 0 \quad \text{for } p = 1, 2, \ldots, \mu. \tag{34}$$

Then the generalization error of $\hat{f}_{T,\alpha}(\boldsymbol{x})$ is expressed as

$$\begin{aligned} J_G &= \mathrm{E}_{\boldsymbol{\epsilon}} \int_{\mathcal{D}} \left(\hat{f}_{T,\alpha}(\boldsymbol{x}) - f(\boldsymbol{x})\right)^2 q(\boldsymbol{x})d\boldsymbol{x} \\ &= \mathrm{E}_{\boldsymbol{\epsilon}} \|\hat{\boldsymbol{\theta}}_{T,\alpha} - \boldsymbol{\theta}\|_U^2 + \int_{\mathcal{D}} \left(g(\boldsymbol{x})\right)^2 q(\boldsymbol{x})d\boldsymbol{x}. \end{aligned} \tag{35}$$

Since the second term is irrelevant to T and α, we focus on the first term.

If a learning matrix X_u that gives an unbiased estimate of $\boldsymbol{\theta}$ is available, an unbiased estimate of the first term in Eq.(35) can be obtained in the same manner as Section 3.1. Therefore, the concept of SIC is still valid in unrealizable scenarios. However, it is not generally possible to obtain an unbiased learning matrix X_u for an unrealizable learning target function.

In practice, Eq.(22) may be used even in unrealizable scenarios. Then it holds that

$$\mathrm{E}_{\boldsymbol{\epsilon}} B^{-1}A^\top \boldsymbol{y} = \boldsymbol{\theta} + B^{-1}A^\top \boldsymbol{z}, \tag{36}$$

where $\boldsymbol{z}$ is an M-dimensional vector defined as

$$\boldsymbol{z} = (g(\boldsymbol{x}_1), g(\boldsymbol{x}_2), \ldots, g(\boldsymbol{x}_M))^\top. \tag{37}$$

This means that if $f(\boldsymbol{x})$ is approximately included in the span of $\{\varphi_p(\boldsymbol{x})\}_{p=1}^{\mu}$ (i.e., $g(\boldsymbol{x})$ is nearly a zero function), Eq.(36) gives a nearly unbiased estimate of $\boldsymbol{\theta}$. In this case, SIC is almost an unbiased estimator of the generalization error.

Furthermore, when sample points $\{\boldsymbol{x}_m\}_{m=1}^{M}$ are independently drawn from the probability density function $q(\boldsymbol{x})$ in Eq.(35), the central limit theorem asserts that $\frac{1}{M}B$ converges to U and $\frac{1}{M}A^\top \boldsymbol{z}$ converges to zero as M tends to infinity, with asymptotic bias $\mathcal{O}_p(M^{-1/2})$. This implies that Eq.(36) is consistent for any unrealizable learning target functions so SIC is consistent in unrealizable scenarios.

4 Optimal selection of regularization matrix and regularization parameter from given pairs

If a set $\mathcal{M}$ of finite pairs of the regularization matrix T and regularization parameter α are given as candidates, SIC can be used for choosing the optimal pair $(\hat{T}, \hat{\alpha})$ from the set $\mathcal{M}$. That is, calculate the value of SIC for each pair (T, α), and choose the one that minimizes SIC:

$$(\hat{T}, \hat{\alpha}) = \underset{(T,\alpha)\in\mathcal{M}}{\operatorname{argmin}}\, SIC[T, \alpha]. \tag{38}$$

SIC includes terms which are irrelevant to T and α. If such terms are ignored, SIC is reduced to

$$SIC[T,\alpha] = \hat{\boldsymbol{\theta}}_{T,\alpha}^{\top} U(\hat{\boldsymbol{\theta}}_{T,\alpha} - 2\hat{\boldsymbol{\theta}}_u) + 2\sigma^2 \mathrm{tr}\left(U(A^{\top}A + \alpha T^{\top}T)^{-1}\right). \tag{39}$$

5 Active design of optimal regularization parameter

In this section, we give a method for actively determining the optimal regularization parameter for a fixed regularization matrix.

5.1 Second order approximation

Let us consider the case when the regularization matrix T is non-singular. Especially, we focus on

$$T = I \tag{40}$$

for simplicity. However, all the discussion in Section 5.1 can be easily scaled to any non-singular T by replacing A, B, and U with

$$A \longleftarrow AT^{-1}, \tag{41}$$
$$B \longleftarrow (T^{-1})^{\top} B T^{-1}, \tag{42}$$
$$U \longleftarrow (T^{-1})^{\top} U T^{-1}. \tag{43}$$

Let X_α be the regularization learning matrix obtained with the regularization matrix I and regularization parameter α:

$$X_\alpha = X_{I,\alpha} = (B + \alpha I)^{-1} A^{\top}. \tag{44}$$

Then the following lemma holds.

Lemma 1 *For a positive integer n, the regularization learning matrix X_α is expressed as*

$$X_\alpha = \sum_{j=1}^{n} (-\alpha)^{j-1} B^{-j} A^{\top} + (-\alpha)^n B^{-(n+1)} (I + \alpha B^{-1})^{-1} A^{\top}. \tag{45}$$

All proofs are provided in Appendix. When sample points $\{\boldsymbol{x}_m\}_{m=1}^{M}$ are independently drawn from a probability density function $r(\boldsymbol{x})$, the eigenvalues of B^{-1} are $\mathcal{O}\left(\frac{1}{M}\right)$ as M tends to infinity. Based on the fact, we have the following lemma.

Lemma 2 *Let $\widehat{SIC}$ be defined by*

$$\begin{aligned}\widehat{SIC}[\alpha] = \alpha^2 \left(\|B^{-2}A^{\top}\boldsymbol{y}\|_U^2 + 2\sigma^2 \mathrm{tr}\left(UB^{-3}\right)\right) \\ - 2\alpha\sigma^2 \mathrm{tr}\left(UB^{-2}\right) + \sigma^2 \mathrm{tr}\left(UB^{-1}\right).\end{aligned} \tag{46}$$

$\widehat{SIC}$ is an approximation to SIC with precision $\mathcal{O}\left(\left(\frac{\alpha}{M}\right)^3\right)$ as M tends to infinity:

$$\widehat{SIC}[\alpha] - SIC[\alpha] = \mathcal{O}\left(\left(\frac{\alpha}{M}\right)^3\right). \tag{47}$$

Note that U may be replaced by $\frac{1}{M}B$ if $q(\boldsymbol{x})$, from which future (test) input points are drawn, agrees with $r(\boldsymbol{x})$, from which training sample points are drawn. Lemma 2 immediately gives the following theorem.

Theorem 1 *Let $\alpha_{\widehat{SIC}}$ be defined by*

$$\alpha_{\widehat{SIC}} = \frac{\sigma^2 \mathrm{tr}\left(UB^{-2}\right)}{\|B^{-2}A^\top \boldsymbol{y}\|_U^2 + 2\sigma^2 \mathrm{tr}\left(UB^{-3}\right)}. \tag{48}$$

$\widehat{SIC}$ is minimized with respect to α if and only if $\alpha = \alpha_{\widehat{SIC}}$.

Theorem 1 is clear from Lemma 2, so its proof is omitted. The validity of $\alpha_{\widehat{SIC}}$ is assessed by the following lemmas.

Lemma 3 *Let $\widehat{J_G}$ be defined by*

$$\begin{aligned}\widehat{J_G}[\alpha] = \alpha^2 \left(\|B^{-1}\boldsymbol{\theta}\|_U^2 + 3\sigma^2 \mathrm{tr}\left(UB^{-3}\right)\right) \\ - 2\alpha\sigma^2 \mathrm{tr}\left(UB^{-2}\right) + \sigma^2 \mathrm{tr}\left(UB^{-1}\right).\end{aligned} \tag{49}$$

$\widehat{J_G}$ is an approximation to J_G with precision $\mathcal{O}\left(\left(\frac{\alpha}{M}\right)^3\right)$ as M tends to infinity:

$$\widehat{J_G}[\alpha] - J_G[\alpha] = \mathcal{O}\left(\left(\frac{\alpha}{M}\right)^3\right). \tag{50}$$

Lemma 4 *Let $\alpha_{\widehat{OPT}}$ be defined by*

$$\alpha_{\widehat{OPT}} = \frac{\sigma^2 \mathrm{tr}\left(UB^{-2}\right)}{\|B^{-1}\boldsymbol{\theta}\|_U^2 + 3\sigma^2 \mathrm{tr}\left(UB^{-3}\right)}. \tag{51}$$

$\widehat{J_G}$ is minimized with respect to α if and only if $\alpha = \alpha_{\widehat{OPT}}$.

Lemma 5 *It holds that*

$$\mathrm{E}_{\boldsymbol{\epsilon}}\|B^{-2}A^\top \boldsymbol{y}\|_U^2 = \|B^{-1}\boldsymbol{\theta}\|_U^2 + \sigma^2 \mathrm{tr}\left(UB^{-3}\right). \tag{52}$$

Proof of Lemmas 4 and 5 are omitted since they are clear. Lemmas 4 and 5 imply that $\alpha_{\widehat{SIC}}$ given by Eq.(48) can be regarded as an estimate of $\alpha_{\widehat{OPT}}$ with the denominator in Eq.(51) estimated by its unbiased estimate[1]. This assures the validity of $\alpha_{\widehat{SIC}}$. Note that the unbiasedness of the denominator does not imply the unbiasedness of $\alpha_{\widehat{SIC}}$.

[1]This property is still maintained if the noise variance σ^2 in $\alpha_{\widehat{SIC}}$ is estimated by Eq.(26). In this case, the numerator of $\alpha_{\widehat{SIC}}$ is also an unbiased estimate of that of $\alpha_{\widehat{OPT}}$.

Remark. In the above discussion, the learning matrix X_α given by Eq.(44) is used for calculating the learning result function, while higher order terms caused by X_α are ignored when the regularization parameter is optimized.

In contrast, if an approximated learning matrix X'_α is used for both calculating learning result functions and optimizing the regularization parameter, all the discussions can be exact. Indeed, let X'_α be the first two terms in Eq.(45):

$$X'_\alpha = B^{-1}A^\top - \alpha B^{-2}A^\top. \tag{53}$$

Then α'_{SIC} and α'_{OPT} that rigorously minimize SIC and J_G with X_α replaced by X'_α are given by

$$\alpha'_{SIC} = \frac{\sigma^2 \mathrm{tr}\left(\alpha U B^{-2}\right)}{\|B^{-2}A^\top \boldsymbol{y}\|_U^2}, \tag{54}$$

$$\alpha'_{OPT} = \frac{\sigma^2 \mathrm{tr}\left(U B^{-2}\right)}{\|B^{-1}\boldsymbol{\theta}\|_U^2 + \sigma^2 \mathrm{tr}\left(U B^{-3}\right)}. \tag{55}$$

5.2 Rigorous solution when $T = A$

We gave the closed form of the optimal regularization parameter when SIC is approximated up to the second order terms. Here, we derive the rigorous solution when the regularization matrix T is given by[2]

$$T = A. \tag{56}$$

In this case, the regularization learning matrix is given by

$$X_{A,\alpha} = (B + \alpha B)^{-1}A^\top = \frac{1}{\alpha + 1}B^{-1}A^\top. \tag{57}$$

Then the following theorem holds.

Theorem 2 *Let α_{SIC} be defined by*

$$\alpha_{SIC} = \frac{\sigma^2 \mathrm{tr}\left(U B^{-1}\right)}{\|B^{-1}A^\top \boldsymbol{y}\|_U^2 - \sigma^2 \mathrm{tr}\left(U B^{-1}\right)}. \tag{58}$$

Under the condition

$$\|B^{-1}A^\top \boldsymbol{y}\|_U^2 > \sigma^2 \mathrm{tr}\left(U B^{-1}\right), \tag{59}$$

SIC is minimized with respect to α if and only if $\alpha = \alpha_{SIC}$.

If Eq.(59) does not hold, α_{SIC} tends to be infinity. The following lemmas play important roles for assessing the validity of α_{SIC}.

[2]This setting is borrowed from Hagiwara and Kuno (2000).

Lemma 6 *Let α_{OPT} be defined by*

$$\alpha_{OPT} = \frac{\sigma^2 \mathrm{tr}\,(UB^{-1})}{\|\boldsymbol{\theta}\|_U^2}. \tag{60}$$

J_G is minimized with respect to α if and only if $\alpha = \alpha_{OPT}$.

Lemma 7 *It holds that*

$$\mathrm{E}_{\boldsymbol{\epsilon}} \|B^{-1}A^\top \boldsymbol{y}\|_U^2 = \|\boldsymbol{\theta}\|_U^2 + \sigma^2 \mathrm{tr}\left(UB^{-1}\right). \tag{61}$$

A proof of Lemma 7 is omitted since it is clear. Lemmas 6 and 7 imply that α_{SIC} given by Eq.(58) can be regarded as an estimate of α_{OPT} with the denominator in Eq.(60) estimated by its unbiased estimate[3]. This assures the validity of α_{SIC}.

6 Computer simulations

In this section, the effectiveness of SIC is experimentally investigated through computer simulations. We assumed the availability of the following items in the derivation of SIC (Section 3.1):

- Basis functions $\{\varphi_p(\boldsymbol{x})\}_{p=1}^{\mu}$ whose span includes the learning target function $f(\boldsymbol{x})$,
- Unbiased learning matrix X_u,
- Noise variance σ^2,
- Correlation matrix U.

Here, we consider practical situations where none of them are available, and experimentally evaluate the robustness of SIC.

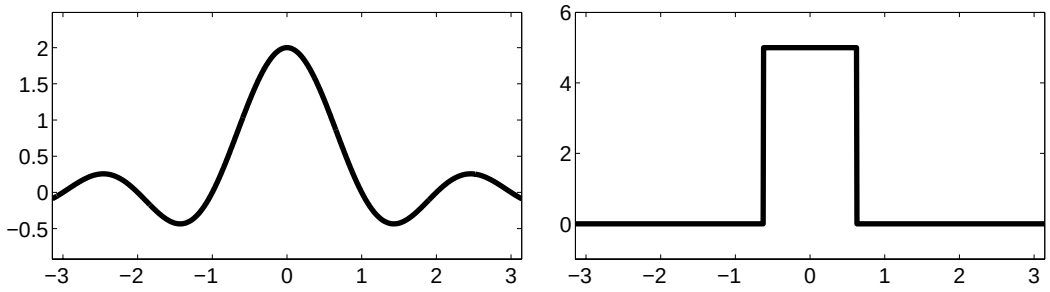
6.1 One-dimensional artificial data

Let the dimension L of the input vector $\boldsymbol{x}$ be 1, and the number μ of basis functions be 21. Let the basis functions $\{\varphi_p(x)\}_{p=1}^{21}$ be

$$\left\{1, \sqrt{2}\sin px, \sqrt{2}\cos px\right\}_{p=1}^{10} \tag{62}$$

defined on $\mathcal{D} = [-\pi, \pi]$. Let us consider the following functions as the learning target function $f(x)$, which are not included in the span of Eq.(62).

[3]Similar to the discussion in Section 5.1, this property is still maintained if σ^2 is estimated by Eq.(26).



(i) Continuous function (ii) Discontinuous function

Figure 2: Profiles of learning target functions.

(i) Continuous function:

$$f(x) = 2\,\mathrm{sinc}\,x. \tag{63}$$

(ii) Discontinuous function:

$$f(x) = \begin{cases} 5 & : \quad x \in (-\frac{\pi}{10}, \frac{\pi}{10}), \\ 0 & : \quad \text{otherwise.} \end{cases} \tag{64}$$

Their profiles are illustrated in Fig.2.

The identity matrix I is adopted as the regularization matrix T. In this case, the regularization learning matrix X_α is given by Eq.(44), and the learning result function $\hat{f}_\alpha(x)$ is given by $X_\alpha \boldsymbol{y}$. We use the following values as candidates of the regularization parameter α:

$$\alpha = 10^{-2}, 10^{-1.5}, 10^{-1}, \ldots, 10^{2}. \tag{65}$$

M sample points $\{x_m\}_{m=1}^M$ are randomly created in the domain $\mathcal{D}$. The noises $\{\epsilon_m\}_{m=1}^M$ are independently drawn from the normal distribution with mean 0 and variance σ^2. We attempt $(M, \sigma^2) = (200, 0.05)$ and $(50, 0.2)$. Simulations are repeated 100 times for each condition, changing the noises $\{\epsilon_m\}_{m=1}^M$ in each trial.

We compare SIC with the *leave-one-out cross-validation* (CV) (see e.g. Orr, 1996), the *generalized cross-validation* (GCV) (Craven & Wahba, 1979), the *network information criterion* (NIC) (Murata *et al.*, 1994; Murata, 1998), *a Bayesian information criterion* (ABIC) (Akaike, 1980), and *Vapnik's measure* (VM) (Cherkassky *et al.*, 1999).

In SIC, the noise variance σ^2 is estimated by Eq.(27). For estimating the correlation matrix U, we use 1000 randomly created unlabeled sample points. Using such unlabeled sample points, U is estimated by Eq.(28).

We also investigate the performance of $\alpha_{\widehat{SIC}}$ given by Eq.(48). In this case, σ^2 is estimated by Eq.(26) since Eq.(27) can not be used for $\alpha_{\widehat{SIC}}$.

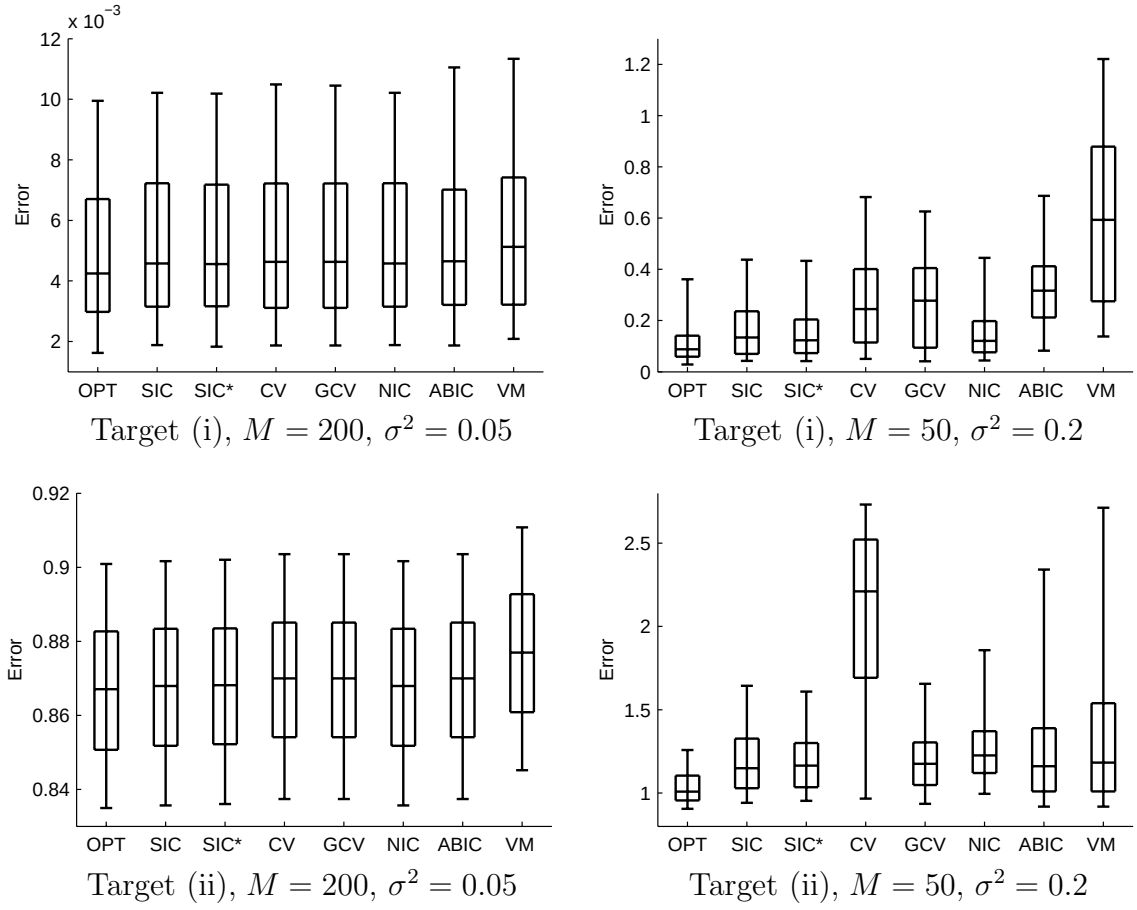


Figure 3: Distributions of error measured by Eq.(66). 'OPT' denotes the optimal selection that minimizes the error. 'SIC*' denotes $\alpha_{\widehat{SIC}}$. Note that the vertical scale is different in each graph.

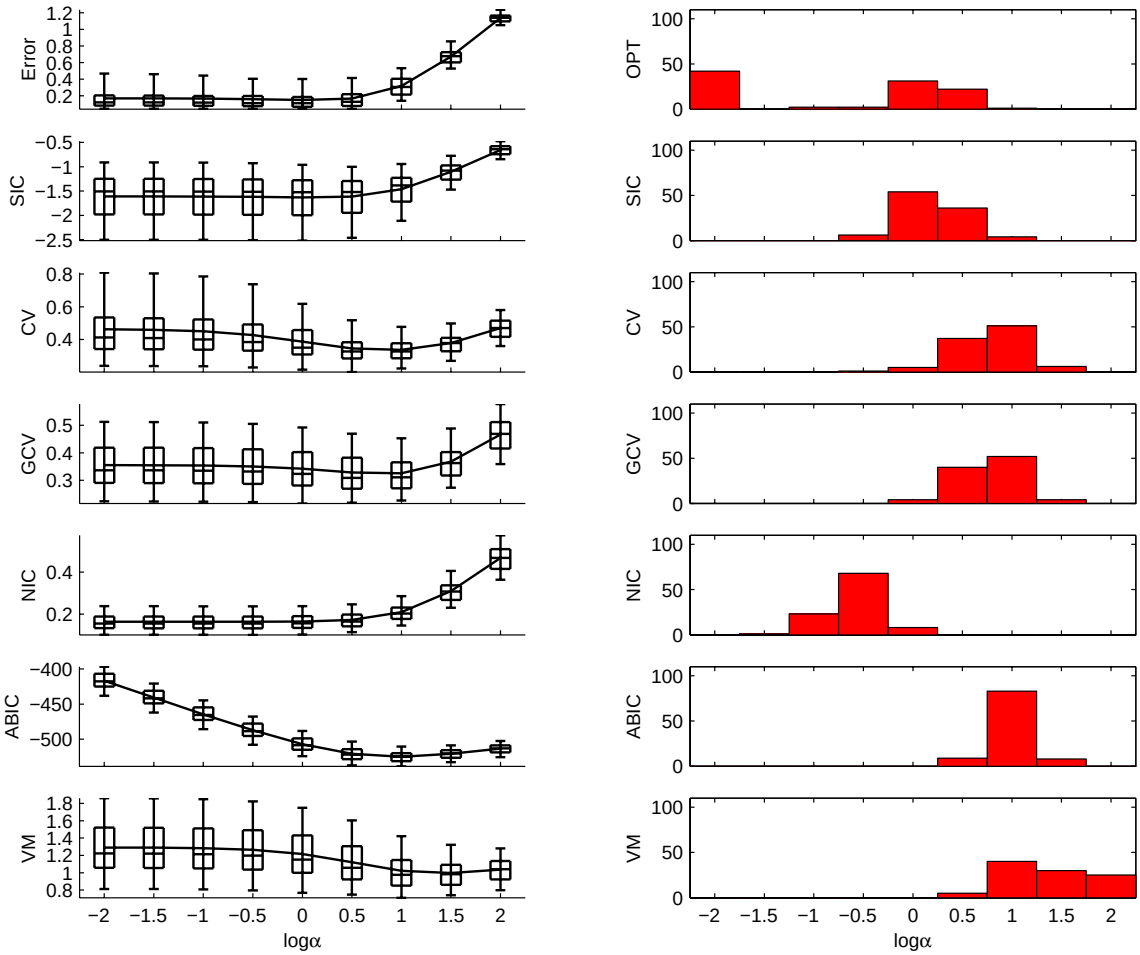
We shall measure the error of a learning result function $\hat{f}_\alpha(x)$ at 1000 randomly created test points $\{x''_m\}_{m=1}^{1000}$:

$$\mathrm{Error}[\alpha] = \frac{1}{1000}\sum_{m=1}^{1000}\left(\hat{f}_\alpha(x''_m) - f(x''_m)\right)^2. \tag{66}$$

Note that the test points are different from the unlabeled sampled points used in SIC.

Fig.3 depicts the simulation results, where the vertical axis denotes the error measured by Eq.(66). The box plot notation specifies marks at 95, 75, 50, 25, and 5 percentiles of values. 'OPT' denotes the optimal selection from Eq.(65) that minimizes the error. 'SIC*' denotes $\alpha_{\widehat{SIC}}$.

When $(M, \sigma^2) = (200, 0.05)$, all methods work well. When $(M, \sigma^2) = (50, 0.2)$, SIC, GCV, and NIC work better than other methods. As discussed in Section 3.2.2, SIC essentially agrees with NIC when training sample points are used instead of unlabeled

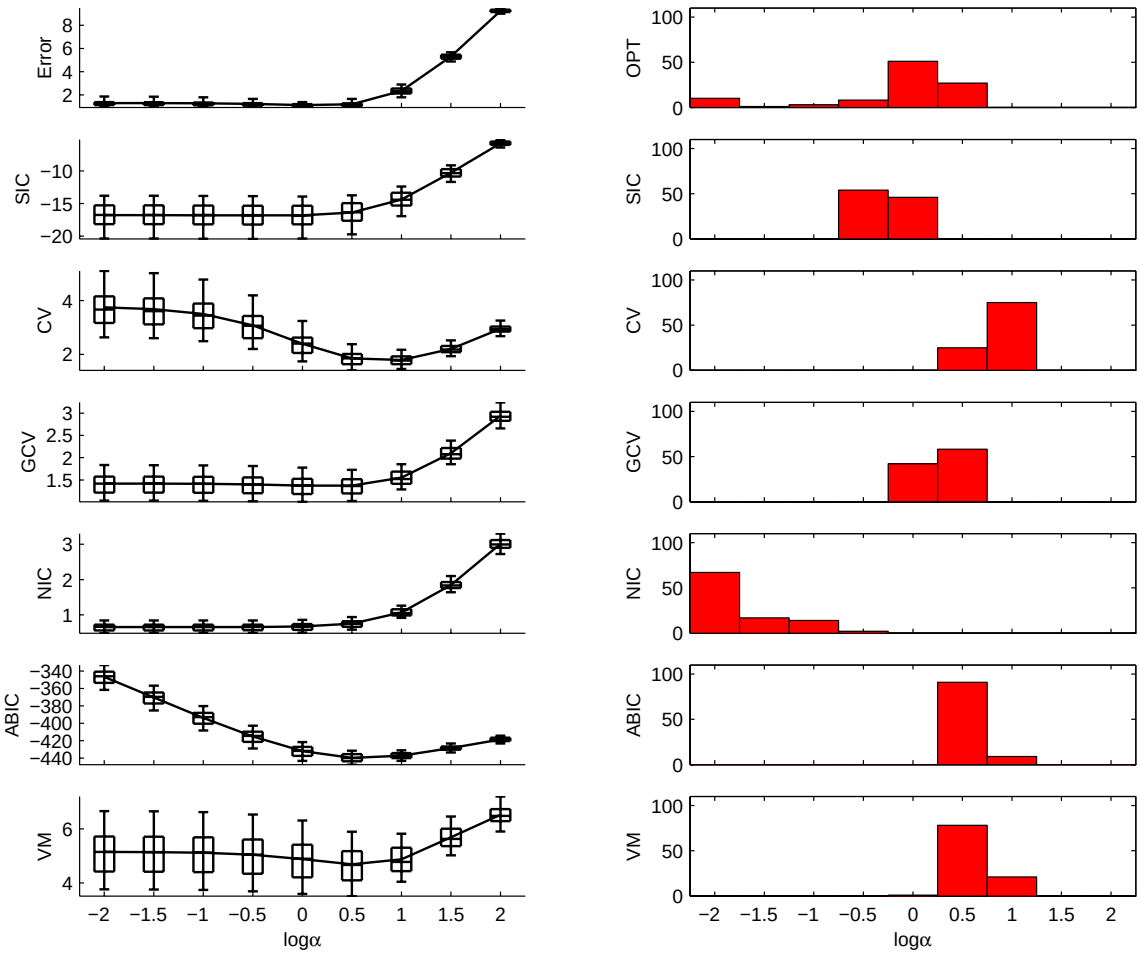


Target (i), $M = 50$, $\sigma^2 = 0.2$

Figure 4: Simulation results in detail. Value of each criterion (left) and selected values of the regularization parameter (right).

sample points for estimating the matrix U. We expected that SIC outperforms NIC since we used additional unlabeled sample points. However, NIC also works very well for the target function (i) so no clear difference can be observed. For the target function (ii), SIC works slightly better than NIC.

In order to further analyze the simulation result, we investigate the value of each criterion and selected values of the regularization parameter when $(M, \sigma^2) = (50, 0.2)$ (see Figs.4 and 5). The left graphs show the values of the error and model selection criteria corresponding to each α. The solid line denotes the mean values over 100 trials. The right graphs show the distributions of selected α. 'OPT' denotes the optimal selection that minimizes the error. Note that in this simulation, SIC is calculated by Eq.(39) where constant terms are ignored, so SIC is shifted in the graph. These graphs show that SIC, GCV, and NIC give good estimates of the error. CV, GCV, ABIC, and VM tend to



Target (ii), $M = 50$, $\sigma^2 = 0.2$

Figure 5: Simulation results in detail.

select larger regularization parameter while NIC is inclined to select smaller regularization parameter, if they are compared with SIC.

6.2 Multi-dimensional real data

To further investigate the performance of SIC, we perform a simulation with a multi-dimensional dataset called the *abalone dataset* (Rasmussen *et al*, 1996).

The abalone dataset has 4177 samples with 9 variables. We use the second to eighth variables as the input vector $\boldsymbol{x}$ and the ninth variable as the output value y. The first variable is ignored because it is qualitative. 120 randomly chosen samples $\{(\boldsymbol{x}_m, y_m)\}_{m=1}^{120}$ are used for training, and the remaining 4057 samples $\{(\boldsymbol{x}''_m, y''_m)\}_{m=1}^{4057}$ are used for measuring

the generalization error:

$$\mathrm{Error}[\alpha] = \frac{1}{4057}\sum_{m=1}^{4057}\left(\hat{f}_\alpha(\boldsymbol{x}''_m) - y''_m\right)^2. \tag{67}$$

Let the number μ of basis functions be 50, and the basis functions $\{\varphi_p(\boldsymbol{x})\}_{p=1}^{50}$ be the Gaussian functions with variance 10 centered on the first 50 training sample points $\{\boldsymbol{x}_m\}_{m=1}^{50}$:

$$\varphi_p(\boldsymbol{x}) = \exp\left(\frac{\|\boldsymbol{x}-\boldsymbol{x}_p\|^2}{10}\right) \quad \text{for } p = 1, 2, \ldots, 50. \tag{68}$$

The identity matrix is adopted as the regularization matrix T, and the following values are used as candidates of the regularization parameter α:

$$\alpha = 10^{-8}, 10^{-7}, 10^{-6}, \ldots, 10^{1}. \tag{69}$$

Simulations are repeated 100 times, changing the training sample points $\{\boldsymbol{x}_m\}_{m=1}^{120}$ in each trial. We again compare 6 criteria described in Section 6.1. In this simulation, we do not have unlabeled sample points, so the correlation matrix U should be estimated only from training examples. Here we will estimate U by Eq.(31) with $\phi(\boldsymbol{x};\boldsymbol{x}_m)$ being the normalized Gaussian function with standard deviation 0.01.

The simulation results are depicted in Fig.6. The top graph shows the obtained error. This result implies that SIC outperforms CV, GCV and NIC, and it is comparable to ABIC and VM. The bottom-left graphs show the values of the error and model selection criteria corresponding to each α. The solid line denotes the mean values of 100 trials. The mean value of ABIC when $\alpha = 10^{-8}$ is 298.9. The bottom-right graphs show the distributions of selected α. 'OPT' denotes the optimal selection from Eq.(69) that minimizes the error. In this simulation, CV, GCV, and NIC tend to select smaller regularization parameter, while SIC, ABIC, and VM seem to specify reasonable regularization parameter.

7 Conclusion

The problem of designing the regularization term and regularization parameter for optimal generalization was discussed. Based on the subspace information criterion (SIC), we gave a method for choosing the optimal regularization term and regularization parameter from given candidates, and derived the closed form of the optimal regularization parameter for a fixed regularization term. The simulation studies showed that SIC can be considered as one of the good model selection criteria.

SIC is an unbiased estimator of the generalization error if the probability density function of future (test) input points is available. When unlabeled sample points are available, the probability density function can be accurately estimated. However, when such additional information is not available, the probability density function should be estimated only from training sample points. If the probability density function is replaced

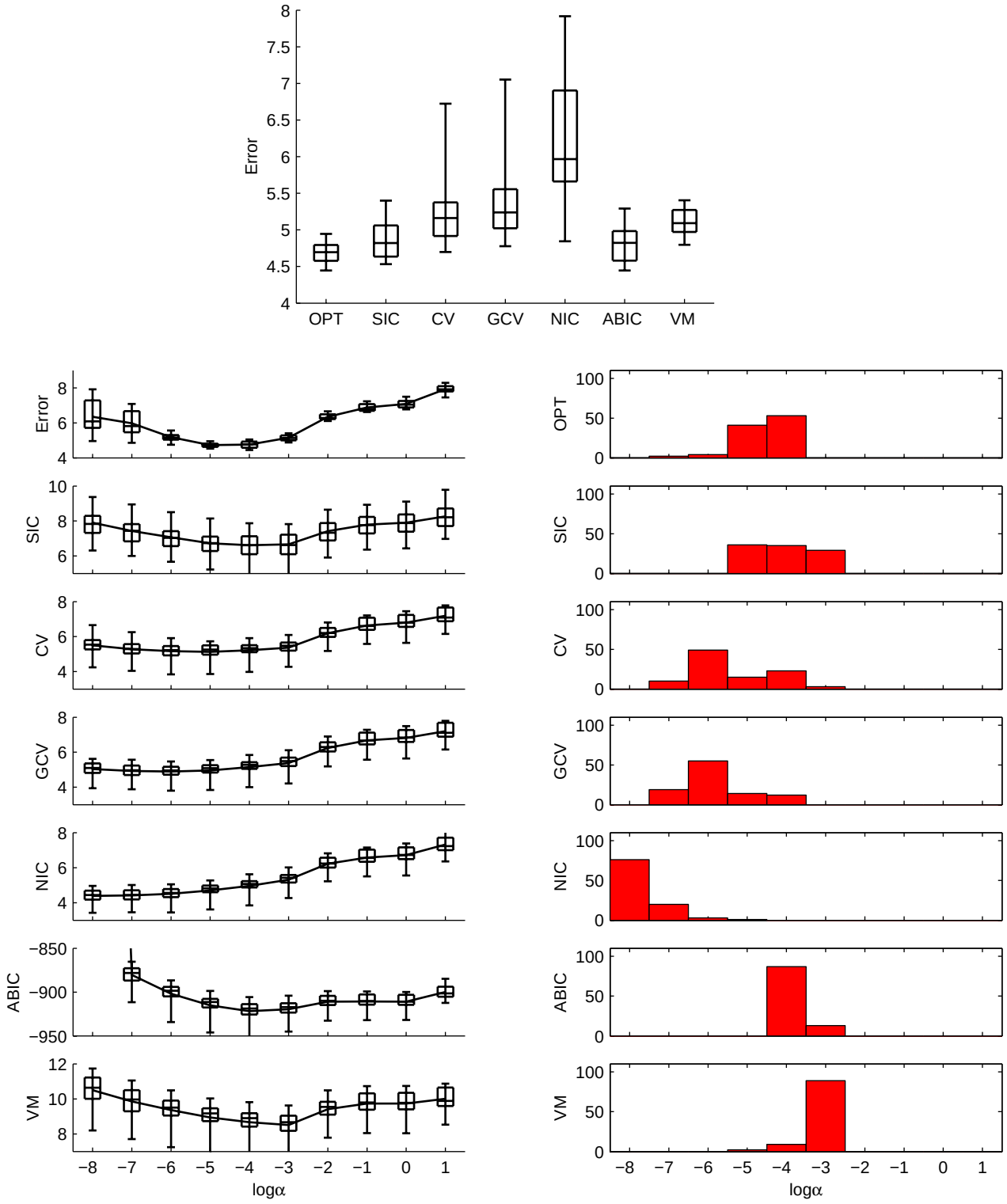


Figure 6: Simulation results with abalone dataset. The mean value of ABIC when $\alpha = 10^{-8}$ is 298.9.

by the empirical density function, SIC is reduced to the traditional Mallows's C_L. Devising a better estimation method of the probability density function is a promising future work.

We proved in Section 3.2.3 that even if the learning target function is not included in the model, SIC is an asymptotic unbiased estimator of the generalization error. However, as described there, if an unbiased estimate of the best approximation in the model is available, SIC is an exact unbiased estimator of the generalization error even in unrealizable scenarios. Our important future work is to investigate the condition when an unbiased estimate of the best approximation can be obtained in unrealizable cases.

Finally, it is important to further investigate the properties of SIC, especially its relation to other model selection criteria.

Acknowledgement

The authors would like to thank Dr. Klaus-Robert Müller, Dr. Koji Tsuda, and Dr. Motoaki Kawanabe for their valuable comments.

Appendix

A Proof of Lemma 1

It holds for any symmetric positive matrix Z that

$$(I+Z)^{-1} = Z^{-1}(I+Z^{-1})^{-1}, \tag{70}$$

$$(I+Z^{-1})^{-1} = I - Z^{-1}(I+Z^{-1})^{-1}. \tag{71}$$

Eq.(70) can be confirmed by multiplying $(I+Z)$ from the left-hand side and $(I+Z^{-1})$ from the right-hand side. Eq.(71) can be confirmed by multiplying $(I+Z^{-1})$ from the right-hand side. If Eq.(71) is repeatedly applied to Eq.(70), we have

$$\begin{aligned}
(I+Z)^{-1} &= Z^{-1}\left[I - Z^{-1}(I+Z^{-1})^{-1}\right] \\
&= Z^{-1} - Z^{-2}(I+Z^{-1})^{-1} \\
&= Z^{-1} - Z^{-2}\left[I - Z^{-1}(I+Z^{-1})^{-1}\right] \\
&= Z^{-1} - Z^{-2} + Z^{-3}(I+Z^{-1})^{-1} \\
&\vdots \\
&= -\sum_{j=1}^{n}(-Z)^{-j} - (-Z)^{-(n+1)}(I+Z^{-1})^{-1},
\end{aligned} \tag{72}$$

where n is an arbitrary fixed positive integer. Then it follows from Eqs.(44) and (72) with $Z = \alpha^{-1}B$ that

$$X_\alpha = \alpha^{-1}(I+\alpha^{-1}B)^{-1}A^\top$$

$$= \alpha^{-1}\Bigg(-\sum_{j=1}^{n}(-\alpha)^j B^{-j} - (-\alpha)^{n+1}B^{-(n+1)}(I+\alpha B^{-1})^{-1}\Bigg)A^\top$$
$$= \sum_{j=1}^{n}(-\alpha)^{j-1}B^{-j}A^\top + (-\alpha)^n B^{-(n+1)}(I+\alpha B^{-1})^{-1}A^\top, \tag{73}$$

which concludes the proof. ■

B Proof of Lemma 2

In the following discussion, terms dominated by $\mathcal{O}\left(\left(\frac{\alpha}{M}\right)^3\right)$ are ignored. It follows from Eq.(45) that

$$X_\alpha \approx B^{-1}A^\top - \alpha B^{-2}A^\top + \alpha^2 B^{-3}A^\top. \tag{74}$$

Then it holds that

$$\begin{aligned} X_\alpha X_\alpha^\top &\approx \left(B^{-1}A^\top - \alpha B^{-2}A^\top + \alpha^2 B^{-3}A^\top\right)\left(AB^{-1} - \alpha AB^{-2} + \alpha^2 AB^{-3}\right) \\ &\approx B^{-1} - 2\alpha B^{-2} + 3\alpha^2 B^{-3}. \end{aligned} \tag{75}$$

It follows from Eqs.(74) and (22) that

$$X_\alpha - X_u \approx -\alpha B^{-2}A^\top + \alpha^2 B^{-3}A^\top, \tag{76}$$

which yields

$$\begin{aligned} (X_\alpha - X_u)^\top U(X_\alpha - X_u) &\approx \left(-\alpha AB^{-2} + \alpha^2 AB^{-3}\right)U\left(-\alpha B^{-2}A^\top + \alpha^2 B^{-3}A^\top\right) \\ &\approx \alpha^2 AB^{-2}UB^{-2}A^\top. \end{aligned} \tag{77}$$

Then it follows from Eqs.(24), (77), and (75) that

$$\begin{aligned} SIC[\alpha] &= \boldsymbol{y}^\top(X_\alpha - X_u)^\top U(X_\alpha - X_u)\boldsymbol{y} \\ &\quad - \sigma^2 \mathrm{tr}\left((X_\alpha - X_u)^\top U(X_\alpha - X_u)\right) + \sigma^2 \mathrm{tr}\left(UX_\alpha X_\alpha^\top\right) \\ &\approx \alpha^2 \boldsymbol{y}^\top AB^{-2}UB^{-2}A^\top\boldsymbol{y} - \alpha^2\sigma^2 \mathrm{tr}\left(AB^{-2}UB^{-2}A^\top\right) \\ &\quad + \sigma^2 \mathrm{tr}\left(UB^{-1} - 2\alpha UB^{-2} + 3\alpha^2 UB^{-3}\right) \\ &= \widehat{SIC}[\alpha], \end{aligned} \tag{78}$$

which concludes the proof. ■

C Proof of Lemma 3

In the following discussion, terms dominated by $\mathcal{O}\left(\left(\frac{\alpha}{M}\right)^3\right)$ are ignored. It follows from Eq.(74) that

$$X_\alpha A - I \approx -\alpha B^{-1} + \alpha^2 B^{-2}, \tag{79}$$

which yields

$$\begin{aligned}(X_\alpha A - I)^\top U (X_\alpha A - I) &\approx (-\alpha B^{-1} + \alpha^2 B^{-2}) U (-\alpha B^{-1} + \alpha^2 B^{-2}) \\ &\approx \alpha^2 B^{-1} U B^{-1}. \end{aligned} \tag{80}$$

It follows from Eqs.(19), (80), and (75) that

$$\begin{aligned} J_G[\alpha] &= \boldsymbol{\theta}^\top (X_\alpha A - I)^\top U (X_\alpha A - I)\boldsymbol{\theta} + \sigma^2 \mathrm{tr}\left(U X_\alpha X_\alpha^\top\right) \\ &\approx \alpha^2 \boldsymbol{\theta}^\top B^{-1} U B^{-1} \boldsymbol{\theta} + \sigma^2 \mathrm{tr}\left(U B^{-1} - 2\alpha U B^{-2} + 3\alpha^2 U B^{-3}\right) \\ &= \widehat{J_G}[\alpha], \end{aligned} \tag{81}$$

which concludes the proof. ■

D Proof of Theorem 2

It follows from Eqs.(57) and (22) that

$$X_{A,\alpha} - X_u = -\frac{\alpha}{\alpha+1} B^{-1} A^\top. \tag{82}$$

Then it follows from Eqs.(24), (82), and (57) that

$$\begin{aligned} SIC[\alpha] &= \frac{\alpha^2 \|B^{-1}A^\top \boldsymbol{y}\|_U^2}{(\alpha+1)^2} - \frac{\alpha^2 \sigma^2 \mathrm{tr}\left(U B^{-1}\right)}{(\alpha+1)^2} + \frac{\sigma^2 \mathrm{tr}\left(U B^{-1}\right)}{(\alpha+1)^2} \\ &= \frac{\alpha^2 \left(\|B^{-1}A^\top \boldsymbol{y}\|_U^2 - \sigma^2 \mathrm{tr}\left(U B^{-1}\right)\right) + \sigma^2 \mathrm{tr}\left(U B^{-1}\right)}{(\alpha+1)^2}. \end{aligned} \tag{83}$$

It is straightforward to show that $(a\alpha^2 + b)/(\alpha+1)^2$ is minimized with respect to α if and only if $\alpha = b/a$. Therefore, SIC is minimized with respect to α if and only if $\alpha = \alpha_{SIC}$. ■

E Proof of Lemma 6

It follows from Eq.(57) that

$$X_{A,\alpha} A - I = -\frac{\alpha}{\alpha+1} I. \tag{84}$$

Then it follows from Eqs.(19), (84), and (57) that

$$\begin{aligned} J_G[\alpha] &= \boldsymbol{\theta}^\top (X_{A,\alpha} A - I)^\top U (X_{A,\alpha} A - I)\boldsymbol{\theta} + \sigma^2 \mathrm{tr}\left(U X_{A,\alpha} X_{A,\alpha}^\top\right) \\ &= \frac{\alpha^2 \|\boldsymbol{\theta}\|_U^2}{(\alpha+1)^2} + \frac{\sigma^2 \mathrm{tr}\left(U B^{-1}\right)}{(\alpha+1)^2}, \end{aligned} \tag{85}$$

which is minimized with respect to α if and only if $\alpha = \alpha_{OPT}$ (see the proof of Theorem 2 for detail). ■

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.

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.

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

Groetsch, C. W. (1984). *The theory of Tikhonov regularization for Fredholm equations of the first kind*. Research Notes in Mathematics Series, 105. Boston: Pitman Advanced Publishing Program.

Hagiwara, K., & Kuno, K. (2000). Regularization learning and early stopping in linear networks. In *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks, 4* (pp. 511–516). Como, Italy.

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

Kunisch, K., & Zou, J. (1998). Iterative choices of regularization parameters in linear inverse problem. *Inverse Problem, 14*, 1247–1264.

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.

Morozov, V. A. (1993). *Regularization methods for ill-posed problems*. Boca Raton: CRC Press, Inc.

Murata, N. (1998). Bias of estimators and regularization terms. In *Proceedings of 1998 Workshop on Information-Based Induction Sciences (IBIS'98)* (pp. 87–94). Izu, Japan.

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 (`http://www.anc.ed.ac.uk/~mjo/papers/intro.ps.gz`).

C. E. Rasmussen, R. M. Neal, G. E. Hinton, D. van Camp, M. Revow, Z. Ghahramani, R. Kustra, and R. Tibshirani. (1996). *The DELVE manual.* (`http://www.cs.toronto.edu/~delve/`).

Schuurmans, D. & Southey, F. (2000). An adaptive regularization criterion for supervised learning. In *Proceedings of International Conference on Machine Learning* (pp. 847–854).

Shibata, R. (1989). Statistical aspects of model selection. In J. C. Willems (Ed.), *From Data to Model* (pp. 215–240). New York: Springer-Verlag.

Sugiyama, M., & Ogawa, H. (2001). Subspace information criterion for model selection. *Neural Computation, 13(8)*, 1863–1889.

Sugiyama, M., & Ogawa, H. (2002). Theoretical and experimental evaluation of subspace information criterion. *Machine Learning, Special Issue on New Methods for Model Selection and Model Combination* (to appear).

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.

Watanabe, S. (2001). Algebraic analysis for non-identifiable learning machines. *Neural Computation, 13(4)*, 899–933.