Selecting Ridge Parameters in Infinite Dimensional Hypothesis Spaces

Masashi Sugiyama[1] and Klaus-Robert Müller[2,3]

[1]Dept. of CS, Tokyo Inst. of Tech., 2-12-1, Ookayama, Meguro, Tokyo 152-8552, Japan
[2] Fraunhofer FIRST, IDA, Kekulestr. 7, 12489 Berlin, Germany
[3] Dept. of CS, U Potsdam, August-Bebel-Str. 89, 14482 Potsdam, Germany
sugi@og.cs.titech.ac.jp, klaus@first.fhg.de

Abstract. Previously, an unbiased estimator of the generalization error called the subspace information criterion (SIC) was proposed for a finite dimensional reproducing kernel Hilbert space (RKHS). In this paper, we extend SIC so that it can be applied to any RKHSs including *infinite* dimensional ones. Computer simulations show that the extended SIC works well in ridge parameter selection.

1 Introduction

Estimating the generalization capability is one of the central issues in supervised learning. So far, a large number of generalization error estimation methods have been proposed (e.g. [1, 9, 4, 8, 5]).

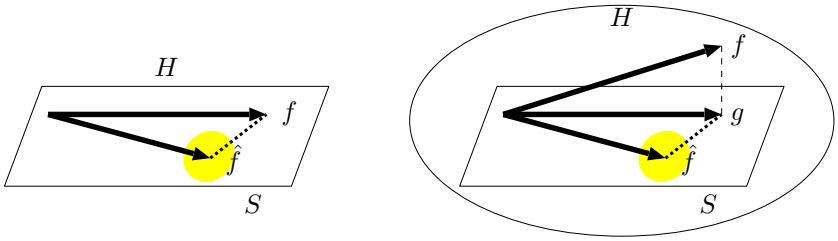
Typically an asymptotic limit in the number of training samples is considered [1, 9, 4]. However, in supervised learning, the small sample case is of high practical importance. Hence methods that work in the finite sample case as e.g. the VC-bound [8], which gives a probabilistic upper bound of the generalization error, are becoming increasingly popular.

Another generalization error estimation method that works effectively with finite samples is the subspace information criterion (SIC) [5]. Among several interesting theoretical properties, SIC is proved to be an unbiased estimator of the generalization error. The original SIC has been successfully applied to the selection of subspace models in linear regression. However, its range of applicability is limited to the case where the learning target function belongs to a specified *finite* dimensional reproducing kernel Hilbert space (RKHS).

In this paper, we therefore extend SIC so that it can be applied to any RKHSs including *infinite* dimensional ones. We further show that when the kernel matrix is invertible, SIC can be expressed in a much simpler form, making its computation highly efficient. Computer simulations underline that the extended SIC works well in ridge parameter selection.

2 Supervised learning and kernel ridge regression

Let us discuss the regression problem of approximating a target function from a set of M *training examples*. Let $f(\boldsymbol{x})$ be a *learning target function* of L variables



(a) $S = H$ (Original SIC) (b) $S \subset H$ (This paper)

Fig. 1. Original SIC and extension carried out in this paper. H is a reproducing kernel Hilbert space that includes the learning target function f. S is the subspace spanned by $\{k(\boldsymbol{x}, \boldsymbol{x}_m)\}_{m=1}^M$. g is the orthogonal projection of f onto S. (a) Setting of the original SIC [5]. It was shown that when $S = H$, SIC is an unbiased estimator of the generalization error between $\hat{f}$ and f with finite samples. $S = H$ implies that the RKHS H whose dimension is at most M ($< \infty$) is considered. (b) Setting of this paper. We consider the case that $S \subset H$, which allows any RKHS H including infinite dimensional ones. We show that the extended SIC is an unbiased estimator of the generalization error between $\hat{f}$ and g.

defined on a subset $\mathcal{D}$ of the L-dimensional Euclidean space $\mathbb{R}^L$. The training examples consist of *sample points* $\boldsymbol{x}_m$ in $\mathcal{D}$ and corresponding *sample values* y_m in $\mathbb{R}$: $\{(\boldsymbol{x}_m, y_m) \mid y_m = f(\boldsymbol{x}_m) + \epsilon_m\}_{m=1}^M$, 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. The purpose of regression is to obtain the optimal approximation $\hat{f}(\boldsymbol{x})$ to the learning target function $f(\boldsymbol{x})$ that minimizes a *generalization error*.

In this paper, we assume that the unknown learning target function $f(\boldsymbol{x})$ belongs to a specified *reproducing kernel Hilbert space* (RKHS) H [9, 8]. We denote the reproducing kernel of H by $k(\boldsymbol{x}, \boldsymbol{x}')$. In previous work [5], it was assumed that $\{k(\boldsymbol{x}, \boldsymbol{x}_m)\}_{m=1}^M$ span the whole RKHS H (Figure 1 (a)). This holds only if $\dim H \leq M$ ($< \infty$). In contrast, we do not impose any restriction on the dimension of the RKHS H in this work. Possibly the dimension is infinity, so we can treat a rich class of function spaces such as e.g. a Gaussian RKHS (Figure 1 (b)). We measure the generalization error of $\hat{f}(\boldsymbol{x})$ by

$$J_G = \mathrm{E}_{\boldsymbol{\epsilon}} \|\hat{f} - f\|^2, \tag{1}$$

where $\mathrm{E}_{\boldsymbol{\epsilon}}$ denotes the expectation over the noise and $\|\cdot\|$ is the norm in the RKHS H. This generalization measure is commonly used in the field of function approximation (e.g. [3]). Since Eq.(1) includes the unknown learning target function $f(\boldsymbol{x})$, it cannot be directly calculated. The aim of this paper is to give an estimator of Eq.(1) that can be calculated without using $f(\boldsymbol{x})$.

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

$$\hat{f}(\boldsymbol{x}) = \textstyle\sum_{p=1}^M \theta_p k(\boldsymbol{x}, \boldsymbol{x}_p), \tag{2}$$

where $\{\theta_p\}_{p=1}^M$ are parameters to be estimated from training examples. We consider the case that the parameter vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_M)^\top$ is determined so

that the regularized training error is minimized[4]:

$$\hat{\boldsymbol{\theta}}_\alpha = \mathrm{argmin}_{\boldsymbol{\theta}}(\sum_{m=1}^{M}(\sum_{p=1}^{M}\theta_p k(\boldsymbol{x}_m,\boldsymbol{x}_p) - y_m)^2 + \alpha\sum_{p=1}^{M}\theta_p^2), \tag{3}$$

where α is a positive scalar called the *ridge parameter*. Let $\boldsymbol{y} = (y_1, y_2, \ldots, y_M)^\top$, I denote the identity matrix, and K be the M-dimensional matrix with the (m,p)-th element $k(\boldsymbol{x}_m,\boldsymbol{x}_p)$. Then $\hat{\boldsymbol{\theta}}_\alpha$ is given by

$$\hat{\boldsymbol{\theta}}_\alpha = (\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_M)^\top = X_\alpha \boldsymbol{y}, \quad \text{where} \quad X_\alpha = (K^2 + \alpha I)^{-1}K. \tag{4}$$

3 SIC for infinite dimensional RKHSs

First, we will briefly review the original SIC [5] that is applicable when $\{k(\boldsymbol{x},\boldsymbol{x}_m)\}_{m=1}^M$ span the whole RKHS H.

When the functions $\{k(\boldsymbol{x},\boldsymbol{x}_m)\}_{m=1}^M$ span the whole space H, the learning target function $f(\boldsymbol{x})$ is expressed as $f(\boldsymbol{x}) = \sum_{p=1}^M \theta_p^* k(\boldsymbol{x},\boldsymbol{x}_p)$, where the true parameter vector $\boldsymbol{\theta}^* = (\theta_1^*, \theta_2^*, \ldots, \theta_M^*)^\top$ is unknown[5]. Letting $\|\boldsymbol{\theta}\|_K^2 = \langle K\boldsymbol{\theta}, \boldsymbol{\theta}\rangle$, the generalization error J_G is expressed as $J_G = \mathrm{E}_{\boldsymbol{\epsilon}}\|\hat{\boldsymbol{\theta}}_\alpha - \boldsymbol{\theta}^*\|_K^2$.

The key idea of SIC is to assume that a learning matrix X_u that gives an unbiased estimator $\hat{\boldsymbol{\theta}}_u$ of the unknown $\boldsymbol{\theta}^*$ is available:

$$\mathrm{E}_{\boldsymbol{\epsilon}}\hat{\boldsymbol{\theta}}_u = \boldsymbol{\theta}^*, \quad \text{where} \quad \hat{\boldsymbol{\theta}}_u = X_u\boldsymbol{y}. \tag{5}$$

Using $\hat{\boldsymbol{\theta}}_u$, the generalization error J_G is roughly estimated by $\|\hat{\boldsymbol{\theta}}_\alpha - \hat{\boldsymbol{\theta}}_u\|_K^2$. Performing some approximations based on this idea, the subspace information criterion (SIC) is given as follows [5]:

$$SIC = \|\hat{\boldsymbol{\theta}}_\alpha - \hat{\boldsymbol{\theta}}_u\|_K^2 - \sigma^2\mathrm{tr}\left(K(X_\alpha - X_u)(X_\alpha - X_u)^\top\right) + \sigma^2\mathrm{tr}\left(KX_\alpha X_\alpha^\top\right), \tag{6}$$

where $\mathrm{tr}\,(\cdot)$ denotes the trace of a matrix. The name *subspace information criterion* (SIC) was first introduced for selecting subspace models. It was shown in [5] that Eq.(6) is an unbiased estimator of the generalization error J_G, i.e., $\mathrm{E}_{\boldsymbol{\epsilon}}SIC = J_G$. When the noise variance σ^2 in Eq.(6) is unknown, one of the practical methods for estimating σ^2 is given by $\hat{\sigma}^2 = \|KX_\alpha\boldsymbol{y} - \boldsymbol{y}\|^2/(M - \mathrm{tr}\,(KX_\alpha))$ [9].

SIC requires a learning matrix X_u that gives an unbiased estimate $\hat{\boldsymbol{\theta}}_u$ of the true parameter $\boldsymbol{\theta}^*$. When $\{k(\boldsymbol{x},\boldsymbol{x}_m)\}_{m=1}^M$ span the whole RKHS H, such X_u surely exists and is given by

$$X_u = K^\dagger, \tag{7}$$

where † denotes the Moore-Penrose generalized inverse. However, obtaining X_u when $\{k(\boldsymbol{x},\boldsymbol{x}_m)\}_{m=1}^M$ do not span the whole RKHS H is an open problem that we aim to solve in the following.

[4] Note that the discussion in this article is valid for any linear estimators.

[5] When $\{k(\boldsymbol{x},\boldsymbol{x}_m)\}_{m=1}^M$ are over-complete, $\{\theta_p^*\}_{p=1}^M$ are not determined uniquely. In this case, we assume that $\boldsymbol{\theta}^*$ is given by $K^\dagger(f(x_1), f(x_2), \ldots, f(x_M))^\top$, where † denotes the Moore-Penrose generalized inverse.

Now, we consider the case when $\{k(\boldsymbol{x}, \boldsymbol{x}_m)\}_{m=1}^M$ do not span the whole RKHS H, possibly $\dim H$ is infinity. Let S be a subspace spanned by $\{k(\boldsymbol{x}, \boldsymbol{x}_m)\}_{m=1}^M$. Since the learning target function $f(\boldsymbol{x})$ does not generally lie in the subspace S, $f(\boldsymbol{x})$ can be decomposed as $f(\boldsymbol{x}) = g(\boldsymbol{x}) + h(\boldsymbol{x})$, where $g(\boldsymbol{x})$ belongs to the subspace S and $h(\boldsymbol{x})$ is orthogonal to S. Then the generalization error can be expressed as $\mathrm{E}_{\boldsymbol{\epsilon}}\|\hat{f} - f\|^2 = \mathrm{E}_{\boldsymbol{\epsilon}}\|\hat{f} - g\|^2 + \|h\|^2$. Since the second term $\|h\|^2$ is irrelevant to $\hat{f}$, we ignore it and focus on the first term $\mathrm{E}_{\boldsymbol{\epsilon}}\|\hat{f} - g\|^2$ (see Figure 1(b)). Let us denote the first term by J'_G:

$$J'_G = \mathrm{E}_{\boldsymbol{\epsilon}}\|\hat{f} - g\|^2. \tag{8}$$

If we regard $g(\boldsymbol{x})$ as the learning target function, then the setting is exactly the same as the original SIC. Therefore, we can apply the idea of SIC and obtain an unbiased estimator of J'_G. However, the problem is that we need a learning matrix X_u that gives an unbiased estimate $\hat{\boldsymbol{\theta}}_u$ of the true parameter[6] $\boldsymbol{\theta}^*$. The following theorem solves this problem.

Theorem 1 [7] *For an arbitrarily chosen RKHS H and the kernel regression model given by Eq.(2), a learning matrix X_u that gives an unbiased estimate $\hat{\boldsymbol{\theta}}_u$ of the true parameter $\boldsymbol{\theta}^*$ is given by*

$$X_u = K^\dagger. \tag{9}$$

Eq.(9) is equivalent to Eq.(7). Therefore, the above theorem shows that SIC is applicable irrespective of the choice of the RKHS H. If $\{k(\boldsymbol{x}, \boldsymbol{x}_m)\}_{m=1}^M$ span the whole RKHS H, SIC is an unbiased estimator of the generalization error J_G. Otherwise SIC is an unbiased estimator of J'_G, which is an essential part of the generalization error J_G (see Figure 1 again):

$$\mathrm{E}_{\boldsymbol{\epsilon}} SIC = J'_G. \tag{10}$$

Now we show that when the kernel matrix K is invertible, SIC can be computed much simpler. Substituting Eq.(9) into Eq.(6), SIC is expressed as

$$SIC = \|\hat{\boldsymbol{\theta}}_\alpha\|_K^2 - 2\langle K\hat{\boldsymbol{\theta}}_\alpha, K^\dagger \boldsymbol{y}\rangle + \|K^\dagger \boldsymbol{y}\|_K^2 + 2\sigma^2 \mathrm{tr}\left(KX_\alpha K^\dagger\right) - \sigma^2 \mathrm{tr}\left(K^\dagger\right). \tag{11}$$

Since the third and fifth terms are irrelevant to α, they can be safely ignored. When K^{-1} exists, a practical expression of SIC for kernel regression is given by

$$\begin{aligned} SIC_{practical} &= \|\hat{\boldsymbol{\theta}}_\alpha\|_K^2 - 2\langle K\hat{\boldsymbol{\theta}}_\alpha, K^{-1}\boldsymbol{y}\rangle + 2\sigma^2 \mathrm{tr}\left(KX_\alpha K^{-1}\right) \\ &= \boldsymbol{y}^\top X_\alpha^\top K X_\alpha \boldsymbol{y} - 2\boldsymbol{y}^\top X_\alpha \boldsymbol{y} + 2\sigma^2 \mathrm{tr}\left(X_\alpha\right). \end{aligned} \tag{12}$$

It is easy to confirm that $\mathrm{E}_{\boldsymbol{\epsilon}} SIC_{practical} + (\text{constant}) = J'_G$, which implies that $SIC_{practical}$ is essentially equivalent to SIC. However, Eq.(12) has the excellent property that K^{-1} is no longer needed. This will highly contribute to the numerical stability of the computations since the matrix inversion can become unstable if the matrix K is ill-conditioned.

[6] When the true function $f(\boldsymbol{x})$ is not included in the model, the term 'true parameter' is used for indicating the parameter in $g(\boldsymbol{x})$, i.e., $g(\boldsymbol{x}) = \sum_{p=1}^M \theta_p^* k(\boldsymbol{x}, \boldsymbol{x}_p)$.

[7] Proof is available from '`ftp://ftp.cs.titech.ac.jp/pub/TR/01/TR01-0016.pdf`'.

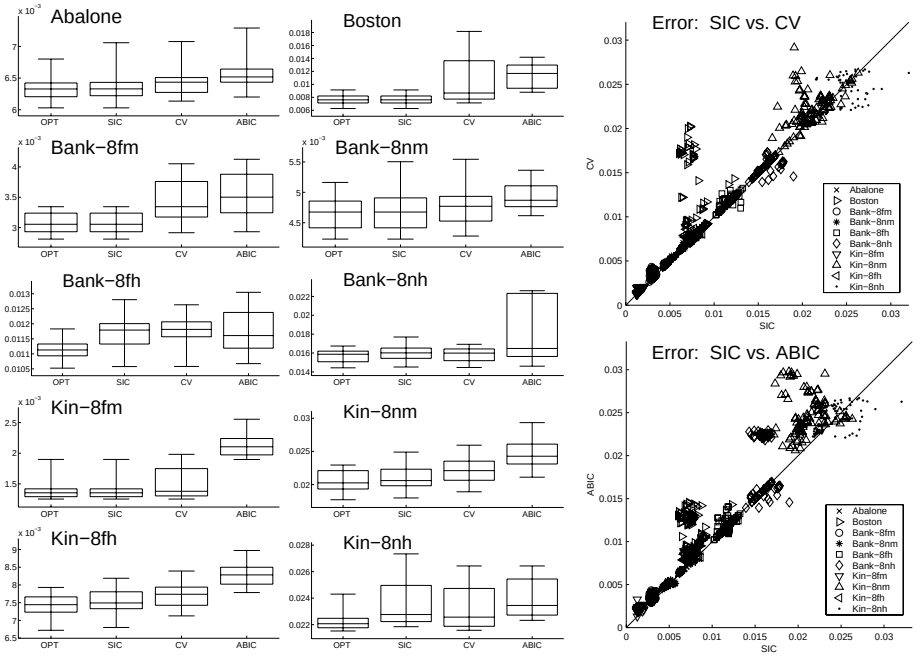


Fig. 2. Simulation results with DELVE data sets (Gaussian kernel).

4 Computer simulations

We use 10 data sets provided by DELVE (`http://www.cs.toronto.edu/~delve/`): *Abalone, Boston, Bank-8fm, Bank-8nm, Bank-8fh, Bank-8nh, Kin-8fm, Kin-8nm, Kin-8fh*, and *Kin-8nh*, where 'f' or 'n' signifies 'fairly linear' or 'non-linear', respectively, and 'm' or 'h' signifies 'medium unpredictability/noise' or 'high unpredictability/noise', respectively. Each dataset consists of several attributes, and the task is to estimate the last one from the rest. For convenience, every attribute is normalized in $[0, 1]$. 100 randomly selected samples are used for training, and the rest is used for testing. We use the Gaussian kernel with variance 1, i.e., $k(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\|\boldsymbol{x} - \boldsymbol{x}'\|^2/2\right)$. The ridge parameter α is selected from $\{10^{-3}, 10^{-2}, 10^{-1}, \ldots, 10^{3}\}$. We compare SIC with Leave-one-out cross-validation (CV) [9] and Akaike's Bayesian information criterion (ABIC) [2]. ABIC is a so-called empirical Bayesian method. The simulation is repeated 100 times for each dataset, changing the training set in each trial.

The left 10 graphs in Figure 2 depict the test error by each method with standard box plot, which specifies marks at 95, 75, 50, 25, and 5 percentiles of values. 'OPT' indicates the test error obtained by the optimal ridge parameter. The right 2 scatter plots show the test error in every trial. These graphs show that SIC works well on the whole. Especially, for *Boston*, *Bank-8fm*, *Kin-8fm*, and *Kin-8nm*, SIC exceedingly outperforms CV and ABIC. However, SIC is rather unstable for *Bank-8nh* and *Kin-8nh*, which are the datasets with high unpredictability/noise. This may be caused by the fact that SIC is derived so that it becomes an *exact* unbiased estimator of the generalization error, but the

variance of the estimator is not taken into account. Therefore, in very high noise cases, the variance of SIC can be large and the SIC estimates may be unstable.

5 Conclusions and discussion

This paper studied an extension of SIC, which allows an efficient model selection even in infinite dimensional RKHSs. In a series of simulations, SIC is shown to work well for most of the data sets.

We found that this good performance can degrade for very high noise levels. This may occur since SIC is an *exact* unbiased estimator of (an essential part of) the generalization error, however, without taking the variance of SIC into account. A future line of research is therefore to investigate the role of the variance of SIC, a path that we have partially explored for the original SIC by adding a small stabilizing bias to SIC (cf. [7]). It remains to be seen whether this or some alternative strategy will also be successful for infinite dimensional RKHSs.

Throughout this paper, we assumed that the target function belongs to a specified RKHS. In extensive experiments that are omitted in this paper, we observed that SIC works properly even when the target function does not exactly lie in the specified RKHS (i.e., unrealizable cases). Although this is surely a useful property in practice, it still remains open how to find an appropriate RKHS.

Acknowledgement: We thank Dr. Tanaka for showing us a preprint of his paper [6]. It greatly motivated us to conduct the current research. Special thanks also go to Dr. Droge, Dr. Laskov, Dr. Rätsch, Dr. Kawanabe, Dr. Tsuda, and Dr. Ogawa for valuable discussions. K.-R. M. acknowledges partial financial support from EU-Neurocolt 2, DFG-MU 987/1-1 and BMBF-BCI.

References

1. H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, AC-19(6):716–723, 1974.
2. H. Akaike. Likelihood and the Bayes procedure. In N. J. Bernardo, et al., editors, *Bayesian Statistics*, pages 141–166, Valencia, 1980. University Press.
3. D. L. Donoho and I. M. Johnstone. Ideal spatial adaptation via wavelet shrinkage. *Biometrika*, 81:425–455, 1994.
4. N. Murata, S. Yoshizawa, and S. Amari. Network information criterion — Determining the number of hidden units for an artificial neural network model. *IEEE Transactions on Neural Networks*, 5(6):865–872, 1994.
5. M. Sugiyama and H. Ogawa. Subspace information criterion for model selection. *Neural Computation*, 13(8):1863–1889, 2001.
6. A. Tanaka, H. Imai, and M. Miyakoshi. Choosing the parameter of image restoration filters by modified subspace information criterion. *IEICE Transactions on Fundamentals*, 2002. to appear.
7. K. Tsuda, M. Sugiyama, and K.-R. Müller. Subspace information criterion for non-quadratic regularizers — Model selection for sparse regressors. *IEEE Transactions on Neural Networks*, 13(1), 70–80, 2002.
8. V. N. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, New York, 1998.
9. H. Wahba. *Spline Model for Observational Data*. Society for Industrial and Applied Mathematics, Philadelphia and Pennsylvania, 1990.