Neural Computation, vol.16, no.5, pp.1077–1104, 2004.

Trading Variance Reduction with Unbiasedness — The Regularized Subspace Information Criterion for Robust Model Selection in Kernel Regression

Masashi Sugiyama (sugi@cs.titech.ac.jp)
Department of Computer Science, Tokyo Institute of Technology
2-12-1, O-okayama, Meguro-ku, Tokyo, 152-8552, Japan, and
Fraunhofer FIRST, IDA, Kekuléstr. 7, 12489 Berlin

Motoaki Kawanabe (nabe@first.fhg.de)
Fraunhofer FIRST, IDA, Kekuléstr. 7, 12489 Berlin

Klaus-Robert Müller (klaus@first.fhg.de)
Fraunhofer FIRST, IDA, Kekuléstr. 7, 12489 Berlin, and
Department of Computer Science, University of Potsdam
August-Bebel-Str.89, Haus 4, 14482 Potsdam, Germany

Abstract

A well-known result by Stein (1956) shows that in particular situations, *biased* estimators can yield better parameter estimates than their generally preferred *unbiased* counterparts. This paper follows the same spirit as we will stabilize the unbiased generalization error estimates by regularization and finally obtain more robust model selection criteria for learning. We trade a small bias against a larger variance reduction which has the beneficial effect of being more precise on a single training set. We focus on the subspace information criterion (SIC), which is an unbiased estimator of the expected generalization error measured by the reproducing kernel Hilbert space norm. SIC can be applied to the kernel regression and it was shown in earlier experiments that a small regularization of SIC has a stabilization effect. However, it remained open how to appropriately determine the degree of regularization in SIC. In this paper, we derive an unbiased estimator of the expected squared error between SIC and the expected generalization error, and propose determining the degree of regularization of SIC such that the estimator of the expected squared error is minimized. Computer simulations with artificial and real data sets illustrate that the proposed method works effectively for improving the precision of SIC, especially in the high noise level cases. We furthermore compare to the original SIC, the cross-validation, and an empirical Bayesian method in ridge parameter selection, with good results.

1 Introduction

Estimating the generalization capability of learning machines has been extensively studied so far because a good estimator of the generalization error can be used for model selection (e.g., Vapnik, 1982, 1995, 1998; Bishop, 1995; Devroye et al., 1996; Müller et al., 2001). Existing work for estimating the generalization error can be roughly classified into two approaches. One is to estimate the *expected* generalization error (e.g., Mallows, 1964, 1973; Akaike, 1974; Takeuchi, 1976; Sugiura, 1978; Craven and Wahba, 1979; Wahba, 1990; Murata et al., 1994; Konishi and Kitagawa, 1996; Murata, 1998; Sugiyama and Ogawa, 2001; Sugiyama and Müller, 2002), and the other is to estimate the *worst case* generalization error (e.g., Vapnik, 1995; Cherkassky et al., 1999; Cucker and Smale, 2002; Bousquet and Elisseeff, 2002). Both approaches have strong theoretical properties, e.g., the accuracy of the estimators of the expected generalization error is theoretically guaranteed in the sense of asymptotic or exact unbiasedness[1], or the validity of the estimators of the worst case generalization error (i.e., upper bounds on the generalization error) is theoretically guaranteed with certain probability. So far, these methods have been successfully applied to various practical learning tasks.

However, unbiased estimators of the expected generalization error can have large variance, or the probabilistic upper bounds on the generalization error can be loose. For this reason, it is very important (i) to reduce the variance of the unbiased estimators of the expected generalization error, or (ii) to tighten the probabilistic upper bounds on the generalization error. In this article, we focus on (i), and propose a method for improving the precision of unbiased estimators of the expected generalization error by regularization. Since we are trying to shrink unbiased estimators of the expected generalization error, this work can be regarded as an application of the idea of the Stein estimator (Stein, 1956) to model selection.

So far, the variance of the unbiased estimators of the expected generalization error has been investigated (e.g., Felsenstein, 1985; Linhart, 1988; Shimodaira, 1997, 1998), in a context where the small differences in the values of Akaike's information criterion (AIC)[2] (Akaike, 1974) is not statistically significant. These papers proposed using a set of 'good' models whose values of AIC are relatively small, rather than selecting the single best model that minimizes AIC. Although these studies instigated us of the need for investigating the variance of the unbiased estimators of the expected generalization error, they are not primarily intended to improve the precision of the estimators.

On the other hand, Tsuda et al. (2002) gave a method for reducing the variance of the subspace information criterion (SIC)[3] (Sugiyama and Ogawa, 2001; Sugiyama and Müller,

[1] Here, the term 'exact unbiasedness' is used for expressing ordinary 'unbiasedness' (i.e., the expectation agrees with the true value for finite samples) in order to emphasize the contrast with asymptotic unbiasedness (the expectation converges to the true value as the number of samples goes to infinity).

[2] AIC is an asymptotic unbiased estimator of the expected generalization error measured by the Kullback-Leibler divergence.

[3] SIC is an unbiased estimator of the expected generalization error measured by the reproducing kernel Hilbert space norm. As described in Sugiyama and Ogawa (2001), SIC can be regarded as an extension of Mallows's C_L (Mallows, 1973).

2002) by introducing a regularization parameter to SIC. It was experimentally shown that a small regularization of SIC highly contributes to stabilization. This work already alluded the possibility of obtaining more precise estimators of the expected generalization error. At the same time, it raised the – so far unresolved – question, how to appropriately determine the degree of regularization in regularized SIC (RSIC)?

In this article, we therefore propose a method for appropriately determining the degree of regularization in RSIC, such that the expected squared error between RSIC and the expected generalization error is minimized. However, we can not directly do so, since the expected squared error includes the unknown expected generalization error. To cope with this problem, we derive an unbiased estimator of the expected squared error that can be calculated from the given data, and propose determining the degree of regularization in RSIC such that this estimator of the expected squared error is minimized.

Finally, we apply the proposed method to the ridge parameter selection in ridge regression. There are several interesting works that theoretically investigate the asymptotic optimality of the choice of the ridge parameter (Craven and Wahba, 1979; Wahba, 1985; Li, 1986). Although we believe that showing the asymptotic optimality of the proposed method may be possible, we are especially interested in the performance with finite samples. For this reason, we shall experimentally investigate the model selection performance of the proposed method in finite sample situations. Simulations with artificial and benchmark data sets show that our regularization approach contributes to improving the precision of SIC, especially it has a stabilizing effect for high noise, and consequently the model selection performance is improved.

The rest of this paper is organized as follows. The regression problem is formulated in Section 2, and the derivation of SIC is briefly reviewed in Section 3. Section 4 introduces RSIC, and gives a method for determining the degree of regularization in RSIC. Computer simulations with artificial and real data sets are performed in Section 5, illustrating how RSIC works. Finally, Section 6 gives the conclusions and future prospects.

2 Problem Formulation

In this section, we formulate the regression problem of approximating a target function from training samples.

Let us denote the learning target function by $f(\boldsymbol{x})$, which is a real-valued function of d variables defined on a subset $\mathcal{D}$ of the d-dimensional Euclidean space $\mathbb{R}^d$. We are given a set of n samples called the *training examples*. A training example consists of a *sample point* $\boldsymbol{x}_i$ in $\mathcal{D}$ and a *sample value* y_i in $\mathbb{R}$. We consider the case that y_i is degraded by unknown additive noise ϵ_i, which is independently drawn from a normal distribution[4]

[4] The normality of the noise is not assumed in our previous works (Sugiyama and Ogawa, 2001; Sugiyama and Müller, 2002). We do assume the normality here because we are dealing with higher order statistics. The discussions in this paper may be generalized to any noise distributions where up to the fourth order moments of the noise are known or can be estimated. However, for simplicity, we focus on the normal noise.

with mean zero and variance σ^2. Then the training examples are expressed as

$$\{(\boldsymbol{x}_i, y_i) \mid y_i = f(\boldsymbol{x}_i) + \epsilon_i\}_{i=1}^n. \tag{1}$$

We assume that the unknown learning target function $f(\boldsymbol{x})$ belongs to a specified *reproducing kernel Hilbert space* (RKHS)[5] $\mathcal{H}$. The *reproducing kernel* of a functional Hilbert space $\mathcal{H}$, denoted by $K(\boldsymbol{x}, \boldsymbol{x}')$, is a bivariate function defined on $\mathcal{D} \times \mathcal{D}$ that satisfies the following conditions (see e.g., Aronszajn, 1950; Bergman, 1970; Saitoh, 1988, 1997; Wahba, 1990; Vapnik, 1998; Cristianini and Shawe-Taylor, 2000):

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

$$\langle f(\cdot), K(\cdot, \boldsymbol{x}')\rangle_{\mathcal{H}} = f(\boldsymbol{x}'), \tag{2}$$

where $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ stands for the inner product in $\mathcal{H}$.

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

$$\hat{f}(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i K(\boldsymbol{x}, \boldsymbol{x}_i), \tag{3}$$

where $\{\alpha_i\}_{i=1}^n$ are parameters to be estimated from training examples. Let us denote the estimated parameters by $\{\hat{\alpha}_i\}_{i=1}^n$. We consider the case that the estimated parameters $\{\hat{\alpha}_i\}_{i=1}^n$ are given by linear combinations of sample values $\{y_i\}_{i=1}^n$. More specifically, letting

$$\begin{aligned} \boldsymbol{y} &= (y_1, y_2, \ldots, y_n)^\top, & (4) \\ \hat{\boldsymbol{\alpha}} &= (\hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_n)^\top, & (5) \end{aligned}$$

where $^\top$ denotes the transpose of a vector (or a matrix), we consider the case that the estimated parameter vector $\hat{\boldsymbol{\alpha}}$ is given by

$$\hat{\boldsymbol{\alpha}} = \boldsymbol{X}\boldsymbol{y}, \tag{6}$$

where $\boldsymbol{X}$ is an n-dimensional matrix that does not depend on the noise $\{\epsilon_i\}_{i=1}^n$. The matrix $\boldsymbol{X}$, which we call the *learning matrix*, can be any matrix, but it is usually determined on the basis of a prespecified learning criterion. For example, in the case of ridge regression (Hoerl and Kennard, 1970), the learning matrix $\boldsymbol{X}$ is determined by minimizing the regularized training error

$$\min \left(\sum_{i=1}^{n} \left(\hat{f}(\boldsymbol{x}_i) - y_i \right)^2 + \lambda \sum_{j=1}^{n} \alpha_j^2 \right), \tag{7}$$

[5] In our early work (Sugiyama and Ogawa, 2001), only finite dimensional RKHSs could be dealt with. However, this restriction has been completely removed by Sugiyama and Müller (2002). The current paper is based on the latter work so we do not impose any restrictions on the choice of the RKHS, e.g., infinite dimensional RKHSs are also allowed.

where λ is a positive scalar called the *ridge parameter.* A minimizer of Eq.(7) is given by the following learning matrix:

$$\boldsymbol{X} = (\boldsymbol{K}^2 + \lambda \boldsymbol{I})^{-1}\boldsymbol{K}, \tag{8}$$

where $\boldsymbol{I}$ denotes the identity matrix and $\boldsymbol{K}$ is the so-called kernel matrix, i.e., the (i, j)-th element of $\boldsymbol{K}$ is given by

$$\boldsymbol{K}_{i,j} = K(\boldsymbol{x}_i, \boldsymbol{x}_j). \tag{9}$$

Note that Bayesian learning with a particular Gaussian process prior yields the same learning matrix (see e.g., Williams and Rasmussen, 1996; Williams, 1998; Cristianini and Shawe-Taylor, 2000). In the following sections, we focus on the above ridge regression for simplicity. However, all the discussions are valid for any learning matrix $\boldsymbol{X}$.

The purpose of regression is to obtain the optimal approximation $\hat{f}(\boldsymbol{x})$ to the unknown learning target function $f(\boldsymbol{x})$. For this purpose, we need a criterion that measures the *closeness* of two functions (i.e., the generalization measure). In this paper, we measure the generalization error by the squared norm in the RKHS $\mathcal{H}$.

$$\|\hat{f} - f\|_{\mathcal{H}}^2, \tag{10}$$

where $\|\cdot\|_{\mathcal{H}}$ denotes the norm in the RKHS $\mathcal{H}$. Using the function space norm as the error measure is rather common in the field of function approximation (e.g., Daubechies, 1992; Donoho and Johnstone, 1994; Donoho, 1995). The use of the RKHS norm is advantageous in the machine learning context since we can measure various different types of errors such as the interpolation error, the extrapolation error, the test error at points of interest, the error at training sample points (Mallows, 1973), the error measured by a weighted norm in the frequency domain (Smola et al., 1998; Girosi, 1998), or the error measured by the Sobolev norm (Wahba, 1990). When unlabeled samples ($\{\boldsymbol{x}_j\}$ without $\{y_j\}$) are available in addition to the usual training examples $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, another advantage of RKHS is that we can utilize those unlabeled samples beneficially and in straight forward manner (Sugiyama and Ogawa, 2002; Tsuda et al., 2002). (For further discussions on this generalization measure we refer to Sugiyama and Müller (2002)).

As stated in Section 1, we focus on estimating the expected generalization error.

$$J_0[\boldsymbol{X}] = \mathrm{E}_{\boldsymbol{\epsilon}}\|\hat{f} - f\|_{\mathcal{H}}^2, \tag{11}$$

where $\mathrm{E}_{\boldsymbol{\epsilon}}$ denotes the expectation over the *noise* $\{\epsilon_i\}_{i=1}^n$. Note that we do *not* take the expectation over the training sample points $\{\boldsymbol{x}_i\}_{i=1}^n$, which is often done in statistical learning frameworks (e.g., Akaike, 1974; Takeuchi, 1976; Murata et al., 1994; Konishi and Kitagawa, 1996; Murata, 1998). Thus, our framework is more data-dependent. We denote the expected generalization error J_0 as a functional of the learning matrix $\boldsymbol{X}$ since under the above setting, specifying $\hat{f}$ is equivalent to specifying the learning matrix $\boldsymbol{X}$. In the following, we often omit $\boldsymbol{X}$ if it is not relevant.

As can be seen from Eq.(11), J_0 includes the unknown learning target function $f(\boldsymbol{x})$, so it can not be directly calculated. The aim of this paper is to give an estimator of Eq.(11) that can be calculated from the given data.

3 Brief Review of the Subspace Information Criterion

The subspace information criterion (SIC) (Sugiyama and Ogawa, 2001; Sugiyama and Müller, 2002) is an unbiased estimator of an essential part of the expected generalization error J_0. In this section, we briefly review the derivation of SIC.

Let $\mathcal{S}$ be the subspace spanned by $\{K(\boldsymbol{x}, \boldsymbol{x}_i)\}_{i=1}^n$, and let $f_{\mathcal{S}}(\boldsymbol{x})$ be the orthogonal projection of $f(\boldsymbol{x})$ onto $\mathcal{S}$. Then the expected generalization error J_0 is expressed by

$$J_0 \;=\; \mathrm{E}_{\boldsymbol{\epsilon}} \|\hat{f} - f_{\mathcal{S}}\|_{\mathcal{H}}^2 + \|f_{\mathcal{S}} - f\|_{\mathcal{H}}^2, \tag{12}$$

where the second term $\|f_{\mathcal{S}} - f\|_{\mathcal{H}}^2$ does not depend on $\hat{f}$. For this reason, we will ignore it and let us denote the first term by J_1:

$$J_1[\boldsymbol{X}] = \mathrm{E}_{\boldsymbol{\epsilon}} \|\hat{f} - f_{\mathcal{S}}\|_{\mathcal{H}}^2. \tag{13}$$

Since the projection $f_{\mathcal{S}}(\boldsymbol{x})$ belongs to $\mathcal{S}$, it can be expressed by

$$f_{\mathcal{S}}(\boldsymbol{x}) = \sum_{i=1}^n \alpha_i^* K(\boldsymbol{x}, \boldsymbol{x}_i), \tag{14}$$

where the parameters $\boldsymbol{\alpha}^* = (\alpha_1^*, \alpha_2^*, \ldots, \alpha_n^*)^\top$ are unknown[6]. For convenience, let us define the weighted norm in $\mathbb{R}^n$:

$$\|\boldsymbol{\alpha}\|_{\boldsymbol{K}}^2 = \langle \boldsymbol{K}\boldsymbol{\alpha}, \boldsymbol{\alpha} \rangle, \tag{15}$$

where the inner product $\langle \cdot, \cdot \rangle$ in the right-hand side is the ordinary Euclidean inner product in $\mathbb{R}^n$. Then J_1 is expressed as

$$J_1 = \mathrm{E}_{\boldsymbol{\epsilon}} \|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_{\boldsymbol{K}}^2. \tag{16}$$

It is known that the above J_1 can be decomposed into the bias and variance terms (see e.g., Geman et al., 1992; Heskes, 1998):

$$J_1 = \|\mathrm{E}_{\boldsymbol{\epsilon}} \hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_{\boldsymbol{K}}^2 + \mathrm{E}_{\boldsymbol{\epsilon}} \|\hat{\boldsymbol{\alpha}} - \mathrm{E}_{\boldsymbol{\epsilon}} \hat{\boldsymbol{\alpha}}\|_{\boldsymbol{K}}^2. \tag{17}$$

The variance term $\mathrm{E}_{\boldsymbol{\epsilon}} \|\hat{\boldsymbol{\alpha}} - \mathrm{E}_{\boldsymbol{\epsilon}} \hat{\boldsymbol{\alpha}}\|_{\boldsymbol{K}}^2$ can be expressed as

$$\mathrm{E}_{\boldsymbol{\epsilon}} \|\hat{\boldsymbol{\alpha}} - \mathrm{E}_{\boldsymbol{\epsilon}} \hat{\boldsymbol{\alpha}}\|_{\boldsymbol{K}}^2 \;=\; \sigma^2 \mathrm{tr}\left(\boldsymbol{K}\boldsymbol{X}\boldsymbol{X}^\top\right), \tag{18}$$

where $\mathrm{tr}\,(\cdot)$ denotes the trace of a matrix, i.e., the sum of diagonal elements. Eq.(18) implies that the variance term $\mathrm{E}_{\boldsymbol{\epsilon}} \|\hat{\boldsymbol{\alpha}} - \mathrm{E}_{\boldsymbol{\epsilon}} \hat{\boldsymbol{\alpha}}\|_{\boldsymbol{K}}^2$ in Eq.(17) can be calculated if the noise

[6] When $\{K(\boldsymbol{x}, \boldsymbol{x}_i)\}_{i=1}^n$ are linearly dependent, $\boldsymbol{\alpha}^*$ is not determined uniquely. In this case, we adopt the minimum norm one.

variance σ^2 is available. When σ^2 is unknown, one of the practical estimates is given as follows (see e.g., Wahba, 1990; Gu et al., 1992):

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}\left(\hat{f}(\boldsymbol{x}_i) - y_i\right)^2}{n - \mathrm{tr}\left(\boldsymbol{K}\boldsymbol{X}\right)} = \frac{\|\boldsymbol{K}\boldsymbol{X}\boldsymbol{y} - \boldsymbol{y}\|^2}{n - \mathrm{tr}\left(\boldsymbol{K}\boldsymbol{X}\right)}. \tag{19}$$

Note that $\|\cdot\|$ in the numerator of the right-hand side of Eq.(19) denotes the ordinary Euclidean norm in $\mathbb{R}^n$.

On the other hand, the bias term $\|\mathrm{E}_{\boldsymbol{\epsilon}}\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_{\boldsymbol{K}}^2$ in Eq.(17) is totally inaccessible since both $\mathrm{E}_{\boldsymbol{\epsilon}}\hat{\boldsymbol{\alpha}}$ and $\boldsymbol{\alpha}^*$ are unknown. The key idea of SIC is to assume that a linear unbiased estimate $\hat{\boldsymbol{\alpha}}_u$ of the unknown true parameter vector $\boldsymbol{\alpha}^*$ is available:

$$\mathrm{E}_{\boldsymbol{\epsilon}}\hat{\boldsymbol{\alpha}}_u = \boldsymbol{\alpha}^*, \tag{20}$$

where $\hat{\boldsymbol{\alpha}}_u$ is given by

$$\hat{\boldsymbol{\alpha}}_u = \boldsymbol{X}_u\boldsymbol{y}. \tag{21}$$

Sugiyama and Müller (2002) proved that such $\boldsymbol{X}_u$ is given by

$$\boldsymbol{X}_u = \boldsymbol{K}^{\dagger}, \tag{22}$$

where † denotes the Moore-Penrose generalized inverse. Using the unbiased estimate $\hat{\boldsymbol{\alpha}}_u$, the bias term $\|\mathrm{E}_{\boldsymbol{\epsilon}}\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_{\boldsymbol{K}}^2$ in Eq.(17) is expressed by

$$\begin{aligned} \|\mathrm{E}_{\boldsymbol{\epsilon}}\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_{\boldsymbol{K}}^2 &= \|\hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}_u\|_{\boldsymbol{K}}^2 + 2\langle \boldsymbol{K}\mathrm{E}_{\boldsymbol{\epsilon}}(\hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}_u), \mathrm{E}_{\boldsymbol{\epsilon}}(\hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}_u) - (\hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}_u)\rangle \\ &\quad - \|\mathrm{E}_{\boldsymbol{\epsilon}}(\hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}_u) - (\hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}_u)\|_{\boldsymbol{K}}^2. \end{aligned} \tag{23}$$

However, the second and third terms in the right-hand side of Eq.(23) are still inaccessible since $\mathrm{E}_{\boldsymbol{\epsilon}}(\hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}_u)$ is unknown, so we replace them by their expectations over the noise.

Then we have the subspace information criterion (SIC)[7] (Sugiyama and Ogawa, 2001; Sugiyama and Müller, 2002):

$$\begin{aligned} \mathrm{SIC}_1[\boldsymbol{X}] &= \|(\boldsymbol{X} - \boldsymbol{X}_u)\boldsymbol{y}\|_{\boldsymbol{K}}^2 - \sigma^2\mathrm{tr}\left(\boldsymbol{K}(\boldsymbol{X} - \boldsymbol{X}_u)(\boldsymbol{X} - \boldsymbol{X}_u)^{\top}\right) \\ &\quad + \sigma^2\mathrm{tr}\left(\boldsymbol{K}\boldsymbol{X}\boldsymbol{X}^{\top}\right). \end{aligned} \tag{24}$$

Note that the subscript 1 is added to 'SIC' in order to emphasize that it is an estimator of J_1 (cf. Section 4.1). It was shown that, for any learning matrix $\boldsymbol{X}$, SIC_1 is an unbiased estimator of J_1:

$$\mathrm{E}_{\boldsymbol{\epsilon}}\mathrm{SIC}_1[\boldsymbol{X}] = J_1[\boldsymbol{X}]. \tag{25}$$

[7] The name subspace information criterion (SIC) came from the fact that it was first introduced for selecting subspace models (Sugiyama and Ogawa, 2001). However, nowadays SIC is not only used for choosing the subspace (i.e., the range of $\boldsymbol{X}$), but also used for choosing the learning matrix $\boldsymbol{X}$ itself (Sugiyama and Müller, 2002). Therefore, in Eq.(24), we described SIC as a functional of the learning matrix $\boldsymbol{X}$. For example, in the case of ridge regression (see Eq.(7)), SIC is regarded as a function of the ridge parameter λ and can be used for choosing the best ridge parameter.

4 Regularization Approach to Stabilizing SIC

As shown in the previous section, SIC is an unbiased estimator of the essential generalization error J_1, and this good property still holds even in finite sample cases (i.e., non-asymptotic cases). Sugiyama and Müller (2002) demonstrated that SIC can be successfully applied to the ridge parameter selection when the noise level is low or medium. However, when the noise level is very high, the performance of SIC sometimes becomes unstable because the variance of SIC can be large. In this section, we propose a method for stabilizing SIC.

4.1 Extracting Essential Part of SIC

SIC_1 defined by Eq.(24) includes terms that do not depend on $\boldsymbol{X}$. Indeed, SIC_1 can be expressed as

$$\begin{aligned}\mathrm{SIC}_1[\boldsymbol{X}] \;=\;& \langle \boldsymbol{KXy}, \boldsymbol{Xy}\rangle - 2\langle \boldsymbol{KXy}, \boldsymbol{X}_u\boldsymbol{y}\rangle + \langle \boldsymbol{KX}_u\boldsymbol{y}, \boldsymbol{X}_u\boldsymbol{y}\rangle \\ & +2\sigma^2 \mathrm{tr}\left(\boldsymbol{X}_u^\top \boldsymbol{KX}\right) - \sigma^2 \mathrm{tr}\left(\boldsymbol{X}_u^\top \boldsymbol{KX}_u\right). \end{aligned} \tag{26}$$

Since SIC_1 is used for choosing the learning matrix $\boldsymbol{X}$, the third and fifth terms in Eq.(26) can be ignored for this purpose. From here on, we use the term 'SIC' for referring to Eq.(26) without the third and fifth terms, i.e., we define

$$\mathrm{SIC}[\boldsymbol{X}] = \langle \boldsymbol{KXy}, \boldsymbol{Xy}\rangle - 2\langle \boldsymbol{KXy}, \boldsymbol{X}_u\boldsymbol{y}\rangle + 2\sigma^2 \mathrm{tr}\left(\boldsymbol{X}_u^\top \boldsymbol{KX}\right). \tag{27}$$

Similarly, J_1 defined by Eq.(13) can be expressed as

$$\begin{aligned} J_1[\boldsymbol{X}] &= \mathrm{E}_{\boldsymbol{\epsilon}}\|\hat{f}\|_{\mathcal{H}}^2 - 2\mathrm{E}_{\boldsymbol{\epsilon}}\langle \hat{f}, f_{\mathcal{S}}\rangle_{\mathcal{H}} + \|f_{\mathcal{S}}\|_{\mathcal{H}}^2 \\ &= \mathrm{E}_{\boldsymbol{\epsilon}}\langle \boldsymbol{KXy}, \boldsymbol{Xy}\rangle - 2\mathrm{E}_{\boldsymbol{\epsilon}}\langle \boldsymbol{KXy}, \boldsymbol{X}_u\boldsymbol{z}\rangle + \langle \boldsymbol{KX}_u\boldsymbol{z}, \boldsymbol{X}_u\boldsymbol{z}\rangle, \end{aligned} \tag{28}$$

where $\boldsymbol{z}$ is the noiseless sample value vector defined by

$$\boldsymbol{z} = (f(\boldsymbol{x}_1), f(\boldsymbol{x}_2), \ldots, f(\boldsymbol{x}_n))^\top. \tag{29}$$

Let us denote the first two terms in Eq.(28) by J:

$$\begin{aligned} J[\boldsymbol{X}] &= \mathrm{E}_{\boldsymbol{\epsilon}}\|\hat{f}\|_{\mathcal{H}}^2 - 2\mathrm{E}_{\boldsymbol{\epsilon}}\langle \hat{f}, f_{\mathcal{S}}\rangle_{\mathcal{H}} \\ &= \mathrm{E}_{\boldsymbol{\epsilon}}\langle \boldsymbol{KXy}, \boldsymbol{Xy}\rangle - 2\mathrm{E}_{\boldsymbol{\epsilon}}\langle \boldsymbol{KXy}, \boldsymbol{X}_u\boldsymbol{z}\rangle. \end{aligned} \tag{30}$$

Then it can be confirmed that, for any learning matrix $\boldsymbol{X}$, SIC given by Eq.(27) is an unbiased estimator of J:

$$\mathrm{E}_{\boldsymbol{\epsilon}}\mathrm{SIC}[\boldsymbol{X}] = J[\boldsymbol{X}]. \tag{31}$$

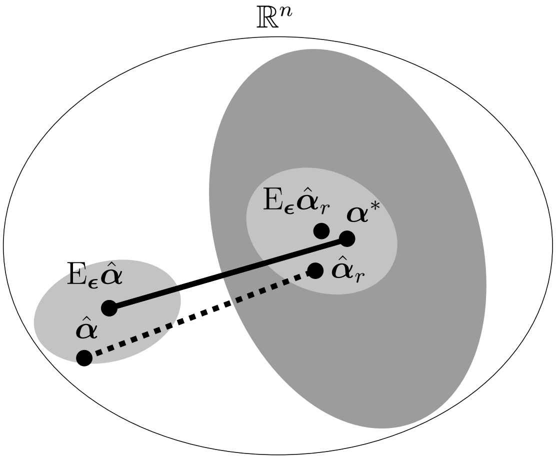
$\mathbb{R}^n$
$\mathrm{E}_{\boldsymbol{\epsilon}}\hat{\boldsymbol{\alpha}}_r$
$\boldsymbol{\alpha}^*$
$\mathrm{E}_{\boldsymbol{\epsilon}}\hat{\boldsymbol{\alpha}}$
$\hat{\boldsymbol{\alpha}}_r$
$\hat{\boldsymbol{\alpha}}$

Figure 1: Basic idea of the regularized SIC (RSIC). The bias term $\|\mathrm{E}_{\boldsymbol{\epsilon}}\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}^*\|_{\boldsymbol{K}}^2$ (depicted by the solid line) is roughly estimated by $\|\hat{\boldsymbol{\alpha}}-\hat{\boldsymbol{\alpha}}_r\|_{\boldsymbol{K}}^2$ (depicted by the dotted line), where $\hat{\boldsymbol{\alpha}}_r$ is a regularized estimate. The regularized estimate $\hat{\boldsymbol{\alpha}}_r$ is slightly biased, so its expectation $\mathrm{E}_{\boldsymbol{\epsilon}}\hat{\boldsymbol{\alpha}}_r$ no longer agrees with the true parameter $\boldsymbol{\alpha}^*$. On the other hand, the 'scatter' of $\hat{\boldsymbol{\alpha}}_r$ (denoted by the thin-colored circle) may be far smaller than that of the unbiased estimate $\hat{\boldsymbol{\alpha}}_u$ (denoted by the dark-colored circle).

4.2 The Regularized SIC

According to Tsuda et al. (2002), the instability of SIC is mainly caused by the large variance of the unbiased estimate $\hat{\boldsymbol{\alpha}}_u$, which plays an essential role in the derivation of SIC (see Section 3). In order to reduce the variance of SIC, Tsuda et al. (2002) proposed replacing the linear unbiased estimate $\hat{\boldsymbol{\alpha}}_u$ by a linear regularized estimate $\hat{\boldsymbol{\alpha}}_r$:

$$\hat{\boldsymbol{\alpha}}_r = \boldsymbol{X}_r\boldsymbol{y}. \tag{32}$$

Namely, the bias term $\|\mathrm{E}_{\boldsymbol{\epsilon}}\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}^*\|_{\boldsymbol{K}}^2$ in Eq.(17) is roughly estimated by $\|\hat{\boldsymbol{\alpha}}-\hat{\boldsymbol{\alpha}}_u\|_{\boldsymbol{K}}^2$ in the original SIC, while Tsuda et al. (2002) proposed estimating it by $\|\hat{\boldsymbol{\alpha}}-\hat{\boldsymbol{\alpha}}_r\|_{\boldsymbol{K}}^2$ (see Figure 1). The regularized estimate $\hat{\boldsymbol{\alpha}}_r$ is slightly biased, so its expectation $\mathrm{E}_{\boldsymbol{\epsilon}}\hat{\boldsymbol{\alpha}}_r$ no longer agrees with the true parameter $\boldsymbol{\alpha}^*$. On the other hand, the 'scatter' of $\hat{\boldsymbol{\alpha}}_r$ may be far smaller than that of the unbiased estimate $\hat{\boldsymbol{\alpha}}_u$. The learning matrix $\boldsymbol{X}_r$ that provides the linear regularized estimate $\hat{\boldsymbol{\alpha}}_r$ is given, e.g., by

$$\boldsymbol{X}_r = (\boldsymbol{K}^2+\gamma\boldsymbol{I})^{-1}\boldsymbol{K}, \tag{33}$$

where γ is the regularization parameter that controls the degree of regularization in SIC. Note that the following discussions are valid for any learning matrix $\boldsymbol{X}_r$, but we mainly focus on Eq.(33) for simplicity. We refer to SIC defined by Eq.(27) with $\boldsymbol{X}_u$ replaced by $\boldsymbol{X}_r$ as the regularized SIC (RSIC):

$$\mathrm{RSIC}[\boldsymbol{X};\boldsymbol{X}_r] = \langle\boldsymbol{KXy},\boldsymbol{Xy}\rangle - 2\langle\boldsymbol{KXy},\boldsymbol{X}_r\boldsymbol{y}\rangle + 2\sigma^2\mathrm{tr}\left(\boldsymbol{X}_r^\top\boldsymbol{KX}\right), \tag{34}$$

where the notation $\mathrm{RSIC}[\boldsymbol{X};\boldsymbol{X}_r]$ means that RSIC is a functional of a learning matrix $\boldsymbol{X}$ with a 'parameter' matrix $\boldsymbol{X}_r$. It was experimentally shown that this regularization approach works effectively for stabilizing SIC (Tsuda et al., 2002). However, the degree of regularization (e.g., the regularization parameter γ in Eq.(33)) should be appropriately determined, which is still a open problem. In the following, we propose a method to determine the degree of regularization of RSIC.

4.3 Expected Squared Error of RSIC

Let us define the expected squared error (ESE) between RSIC and J by

$$\mathrm{ESE}_{\mathrm{RSIC}}[\boldsymbol{X}_r;\boldsymbol{X}] = \mathrm{E}_{\boldsymbol{\epsilon}}(\mathrm{RSIC}[\boldsymbol{X};\boldsymbol{X}_r] - J[\boldsymbol{X}])^2, \tag{35}$$

where the notation $\mathrm{ESE}_{\mathrm{RSIC}}[\boldsymbol{X}_r;\boldsymbol{X}]$ means that we treat $\mathrm{ESE}_{\mathrm{RSIC}}$ as a functional of the matrix $\boldsymbol{X}_r$ with a 'parameter' matrix $\boldsymbol{X}$. In the following, we often omit $[\boldsymbol{X}_r;\boldsymbol{X}]$. Our aim is to determine $\boldsymbol{X}_r$ in RSIC so that the above $\mathrm{ESE}_{\mathrm{RSIC}}$ is minimized.

Similar to Eq.(17), $\mathrm{ESE}_{\mathrm{RSIC}}$ can be decomposed into the bias and variance terms:

$$\mathrm{ESE}_{\mathrm{RSIC}}[\boldsymbol{X}_r;\boldsymbol{X}] = \mathrm{Bias}^2_{\mathrm{RSIC}}[\boldsymbol{X}_r;\boldsymbol{X}] + \mathrm{Var}_{\mathrm{RSIC}}[\boldsymbol{X}_r;\boldsymbol{X}], \tag{36}$$

where

$$\begin{aligned}\mathrm{Bias}_{\mathrm{RSIC}}[\boldsymbol{X}_r;\boldsymbol{X}] &= \mathrm{E}_{\boldsymbol{\epsilon}}\mathrm{RSIC}[\boldsymbol{X};\boldsymbol{X}_r] - J[\boldsymbol{X}], & (37)\\ \mathrm{Var}_{\mathrm{RSIC}}[\boldsymbol{X}_r;\boldsymbol{X}] &= \mathrm{E}_{\boldsymbol{\epsilon}}(\mathrm{RSIC}[\boldsymbol{X};\boldsymbol{X}_r] - \mathrm{E}_{\boldsymbol{\epsilon}}\mathrm{RSIC}[\boldsymbol{X};\boldsymbol{X}_r])^2. & (38)\end{aligned}$$

Note that the bias of SIC is zero (see Eq.(31)), but there is no guarantee that ESE of SIC is small since the variance of SIC can be large.

Let $\boldsymbol{B}$ and $\boldsymbol{C}$ be n-dimensional matrices defined by

$$\begin{aligned}\boldsymbol{B} &= 2\boldsymbol{X}_u^\top\boldsymbol{K}\boldsymbol{X} - 2\boldsymbol{X}_r^\top\boldsymbol{K}\boldsymbol{X}, & (39)\\ \boldsymbol{C} &= \boldsymbol{X}^\top\boldsymbol{K}\boldsymbol{X} - 2\boldsymbol{X}_r^\top\boldsymbol{K}\boldsymbol{X}. & (40)\end{aligned}$$

Then we have the following lemmas.

Lemma 1 $\mathrm{Bias}_{\mathrm{RSIC}}$ *is expressed by*

$$\mathrm{Bias}_{\mathrm{RSIC}} = \langle \boldsymbol{B}\boldsymbol{z}, \boldsymbol{z}\rangle, \tag{41}$$

where $\boldsymbol{z}$ is defined by Eq.(29).

Lemma 2 *Under the assumption that $\{\epsilon_i\}_{i=1}^n$ are independently drawn from the normal distribution with mean zero and variance σ^2,* $\mathrm{Var}_{\mathrm{RSIC}}$ *is expressed by*

$$\mathrm{Var}_{\mathrm{RSIC}} = \sigma^2\|(\boldsymbol{C}+\boldsymbol{C}^\top)\boldsymbol{z}\|^2 + \sigma^4\mathrm{tr}\left(\boldsymbol{C}^2 + \boldsymbol{C}^\top\boldsymbol{C}\right). \tag{42}$$

Sketches of the proofs of all lemmas and theorems are given in Appendix. See the separate technical report (Sugiyama et al., 2003) for the complete proofs. Note that the normality of the noise is used only in Lemma 2, not in Lemma 1.

4.4 Estimating the Expected Squared Error of RSIC

In Eqs.(41) and (42), the noiseless sample value vector $\boldsymbol{z}$ defined by Eq.(29) is unknown. Therefore, $\mathrm{Bias}_{\mathrm{RSIC}}$ and $\mathrm{Var}_{\mathrm{RSIC}}$ can not be directly calculated in practice. Now let us define

$$\begin{aligned}
\widehat{\mathrm{Bias}^2}_{\mathrm{RSIC}}[\boldsymbol{X}_r;\boldsymbol{X}] &= \langle \boldsymbol{B}\boldsymbol{y},\boldsymbol{y}\rangle^2 - \sigma^2\|(\boldsymbol{B}+\boldsymbol{B}^\top)\boldsymbol{y}\|^2 - 2\sigma^2\mathrm{tr}\left(\boldsymbol{B}\right)\langle \boldsymbol{B}\boldsymbol{y},\boldsymbol{y}\rangle \\
&\quad +\sigma^4\mathrm{tr}\left(\boldsymbol{B}^2+\boldsymbol{B}^\top\boldsymbol{B}\right) + \sigma^4\mathrm{tr}\left(\boldsymbol{B}\right)^2, \qquad (43) \\
\widehat{\mathrm{Var}}_{\mathrm{RSIC}}[\boldsymbol{X}_r;\boldsymbol{X}] &= \sigma^2\|(\boldsymbol{C}+\boldsymbol{C}^\top)\boldsymbol{y}\|^2 - \sigma^4\mathrm{tr}\left(\boldsymbol{C}^2+\boldsymbol{C}^\top\boldsymbol{C}\right). \qquad (44)
\end{aligned}$$

Then the following theorem holds.

Theorem 3 *Under the assumption that $\{\epsilon_i\}_{i=1}^n$ are independently drawn from the normal distribution with mean zero and variance σ^2, the following relations hold for any $\boldsymbol{X}_r$ and $\boldsymbol{X}$.*

$$\begin{aligned}
\mathrm{E}_{\boldsymbol{\epsilon}}\widehat{\mathrm{Bias}^2}_{\mathrm{RSIC}}[\boldsymbol{X}_r;\boldsymbol{X}] &= \mathrm{Bias}^2_{\mathrm{RSIC}}[\boldsymbol{X}_r;\boldsymbol{X}], \qquad (45) \\
\mathrm{E}_{\boldsymbol{\epsilon}}\widehat{\mathrm{Var}}_{\mathrm{RSIC}}[\boldsymbol{X}_r;\boldsymbol{X}] &= \mathrm{Var}_{\mathrm{RSIC}}[\boldsymbol{X}_r;\boldsymbol{X}]. \qquad (46)
\end{aligned}$$

The above theorem shows that $\widehat{\mathrm{Bias}^2}_{\mathrm{RSIC}}$ and $\widehat{\mathrm{Var}}_{\mathrm{RSIC}}$ are unbiased estimators of $\mathrm{Bias}^2_{\mathrm{RSIC}}$ and $\mathrm{Var}_{\mathrm{RSIC}}$, respectively.

Let us define

$$\widehat{\mathrm{ESE}}_{\mathrm{RSIC}}[\boldsymbol{X}_r;\boldsymbol{X}] = \widehat{\mathrm{Bias}^2}_{\mathrm{RSIC}}[\boldsymbol{X}_r;\boldsymbol{X}] + \widehat{\mathrm{Var}}_{\mathrm{RSIC}}[\boldsymbol{X}_r;\boldsymbol{X}]. \qquad (47)$$

Then, from Theorem 3, we immediately have the following corollary.

Corollary 4 *Under the assumption that $\{\epsilon_i\}_{i=1}^n$ are independently drawn from the normal distribution with mean zero and variance σ^2, the following relation holds for any $\boldsymbol{X}_r$ and $\boldsymbol{X}$.*

$$\mathrm{E}_{\boldsymbol{\epsilon}}\widehat{\mathrm{ESE}}_{\mathrm{RSIC}}[\boldsymbol{X}_r;\boldsymbol{X}] = \mathrm{ESE}_{\mathrm{RSIC}}[\boldsymbol{X}_r;\boldsymbol{X}]. \qquad (48)$$

Corollary 4 shows that the $\widehat{\mathrm{ESE}}_{\mathrm{RSIC}}$ defined by Eq.(47) is an unbiased estimator of $\mathrm{ESE}_{\mathrm{RSIC}}$. Based on this corollary, we propose using $\widehat{\mathrm{ESE}}_{\mathrm{RSIC}}[\boldsymbol{X}_r;\boldsymbol{X}]$ for determining the degree of regularization of RSIC, i.e., $\boldsymbol{X}_r$ is determined such that $\widehat{\mathrm{ESE}}_{\mathrm{RSIC}}[\boldsymbol{X}_r;\boldsymbol{X}]$ is minimized. Note that $\widehat{\mathrm{ESE}}_{\mathrm{RSIC}}[\boldsymbol{X}_r;\boldsymbol{X}]$ depends on the learning matrix $\boldsymbol{X}$, so $\boldsymbol{X}_r$ is individually optimized for each $\boldsymbol{X}$.

For example, when $\boldsymbol{X}$ and $\boldsymbol{X}_r$ are both ridge regression[8], RSIC is treated as a function of λ with a tuning parameter γ and $\widehat{\mathrm{ESE}}_{\mathrm{RSIC}}$ is treated as a function of γ that depends on λ. The regularization parameter γ in RSIC is determined for each ridge parameter λ such that $\widehat{\mathrm{ESE}}_{\mathrm{RSIC}}$ is minimized, and then λ is determined such that RSIC is minimized:

$$\hat{\lambda}_{\mathrm{RSIC}} = \underset{\lambda}{\mathrm{argmin}}\,\mathrm{RSIC}(\lambda;\hat{\gamma}_\lambda), \qquad (49)$$

[8] Namely, $\boldsymbol{X}$ is given by Eq.(8) and $\boldsymbol{X}_r$ is given by Eq.(33).

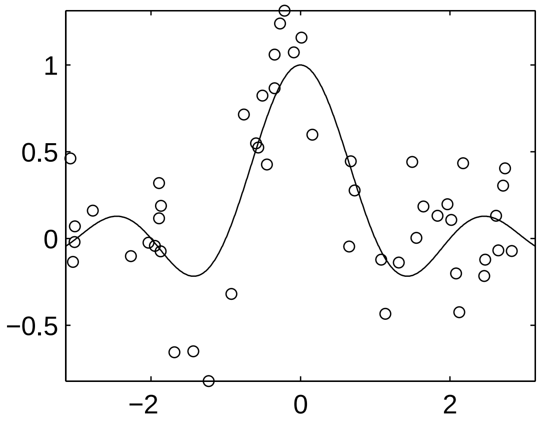

Figure 2: Learning target function and 50 training examples with noise variance $\sigma^2 = 0.09$.

where

$$\hat{\gamma}_\lambda = \underset{\gamma}{\mathrm{argmin}}\, \widehat{\mathrm{ESE}}_{\mathrm{RSIC}}(\gamma; \lambda). \tag{50}$$

When the noise variance σ^2 is unknown, it can be estimated, e.g., by Eq.(19).

5 Computer Simulations

In this section, the effectiveness of the proposed generalization error estimation method is investigated through computer simulations.

5.1 Illustrative Examples

First, a simple artificial simulation shows how the proposed method works[9].

5.1.1 Setting

For illustration purpose, let the dimension d of the input vector be 1. We use the Gaussian RKHS with width $c = 1$, which may be one of the standard RKHSs (see e.g., Vapnik, 1998; Schölkopf et al., 2000):

$$K(x, x') = \exp\left(-\frac{(x - x')^2}{2c^2}\right). \tag{51}$$

We use $f(x) = \mathrm{sinc}(x)$ as the learning target function (see Figure 2), which is often used as an illustrative regression example (e.g., Vapnik, 1998; Schölkopf et al., 2000). Note

[9] Because of the space limitation, we describe the results only briefly here. For extensive discussions, see Sugiyama et al. (2003)

that the above sinc function is included in the Gaussian RKHS[10].

The sample points $\{x_i\}_{i=1}^n$ are independently drawn from the uniform distribution on $(-\pi, \pi)$. The sample values $\{y_i\}_{i=1}^n$ are created as $y_i = f(x_i) + \epsilon_i$, where the noise $\{\epsilon_i\}_{i=1}^n$ are independently drawn from the normal distribution with mean zero and variance σ^2. We consider the following four cases as the number n of training examples and the noise variance σ^2:

$$\begin{aligned}(n, \sigma^2) &= (100, 0.01), (100, 0.09), \\ &\quad (50, 0.01), (50, 0.09),\end{aligned} \tag{52}$$

i.e., we investigate the cases with small/large noise levels and small/large samples. An example of the training set is also illustrated in Figure 2. The simulations are repeated 100 times for each (n, σ^2) in Eq.(52), randomly drawing the sample points $\{x_i\}_{i=1}^n$ and noise $\{\epsilon_i\}_{i=1}^n$ from scratch in each trial. Note that in theory, we fix the training sample points $\{x_i\}_{i=1}^n$ and only change the noise $\{\epsilon_i\}_{i=1}^n$ (see Section 2). However, in this experiment, we change both the training sample points $\{x_i\}_{i=1}^n$ and noise $\{\epsilon_i\}_{i=1}^n$ because we would like to investigate whether the proposed method works irrespective of the choice of the training set.

We use the kernel regression model (3), and the parameters $\{\alpha_i\}_{i=1}^n$ in the model are learned by ridge regression, i.e., the learning matrix is given by Eq.(8).

5.1.2 Investigating Generalization Error Estimation Performance

First, we illustrate how SIC and RSIC work in generalization error estimation. The precision of SIC and RSIC is investigated as a function of the ridge parameter λ, using the following values:

$$\lambda \in \{10^{-3}, 10^{-2.5}, 10^{-2}, \ldots, 10^3\}. \tag{53}$$

When the ridge regression (8) is used, it holds that $\boldsymbol{K}^\top = \boldsymbol{K}$, $\boldsymbol{X}^\top = \boldsymbol{X}$, and $\boldsymbol{K}^\dagger \boldsymbol{K} \boldsymbol{X} = \boldsymbol{X}$. Therefore, SIC given by Eq.(27) can be expressed in the following simpler form.

$$\mathrm{SIC}(\lambda) = \langle \boldsymbol{X}_\lambda \boldsymbol{K} \boldsymbol{X}_\lambda \boldsymbol{y}, \boldsymbol{y} \rangle - 2\langle \boldsymbol{X}_\lambda \boldsymbol{y}, \boldsymbol{y} \rangle + 2\sigma^2 \mathrm{tr}\left(\boldsymbol{X}_\lambda\right), \tag{54}$$

where $\boldsymbol{X}_\lambda$ denotes the learning matrix (8) with a ridge parameter λ.

We calculate SIC by the above simpler form, where the noise variance σ^2 is estimated by

$$\hat{\sigma}_\lambda^2 = \frac{\|\boldsymbol{K}\boldsymbol{X}_\lambda \boldsymbol{y} - \boldsymbol{y}\|^2}{n - \mathrm{tr}\left(\boldsymbol{K}\boldsymbol{X}_\lambda\right)}. \tag{55}$$

[10] As described in Smola et al. (1998) and Girosi (1998), the Gaussian RKHS is spanned by the function $f(x)$ that belongs to $L_2(\mathbb{R})$ and satisfies

$$\int_{-\infty}^{\infty} \frac{|\tilde{f}(\omega)|^2}{\tilde{k}(\omega)} d\omega < \infty,$$

where $\tilde{f}(\omega)$ is the Fourier transform of the function $f(x)$ and $\tilde{k}(\omega)$ is the Fourier transform of $\exp\left(-\frac{x^2}{2c^2}\right)$. The sinc function belongs to $L_2(\mathbb{R})$, and its Fourier transform is zero for $|\omega| > \pi$. Therefore, the above conditions are fulfilled so the sinc function is included in the Gaussian RKHS.

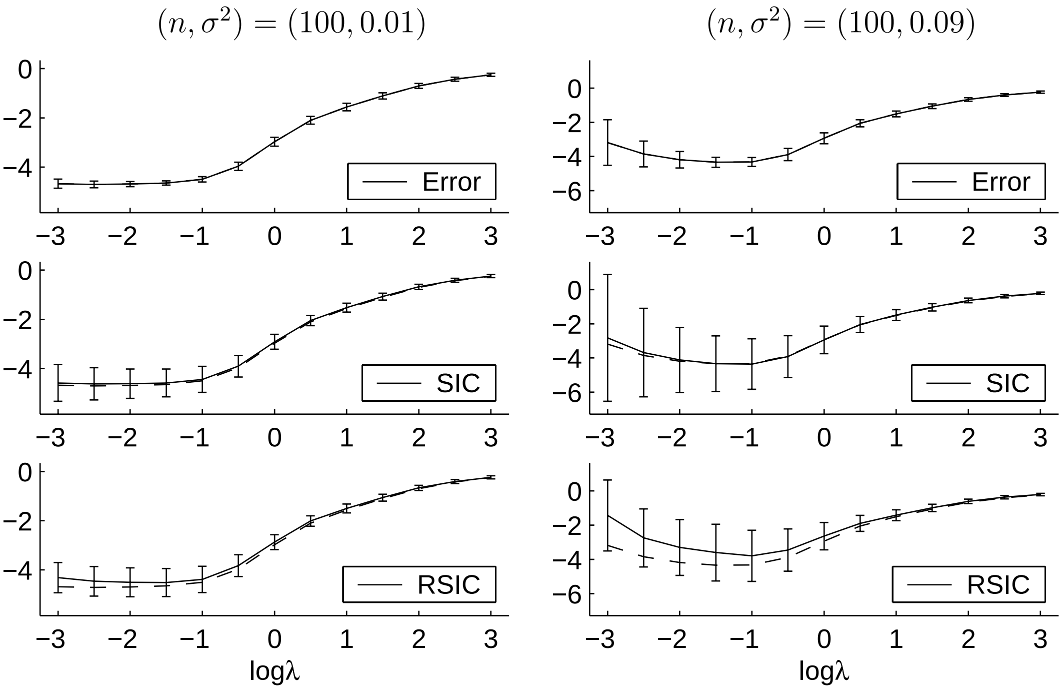

Figure 3: Values of Error(λ), SIC(λ), and RSIC(λ). The horizontal axis denotes the value of λ in log-scale. From top, the graphs denote the mean Error with error bar, the mean SIC with error bar, and the mean RSIC with error bar. Dashed curves in the bottom two graphs are the mean Error (same as the curve in the top graph).

RSIC is calculated by Eq.(34), where the ridge regression (33) is used for obtaining the regularized estimator $\hat{\boldsymbol{\alpha}}_r$. The regularization parameter γ in RSIC is determined so that $\widehat{\mathrm{ESE}}_{\mathrm{RSIC}}(\gamma; \lambda)$ is minimized (see Eq.(47)). Note that the optimization of γ is individually carried out for each λ in Eq.(53). The regularization parameter γ is selected from $\{10^{-3}, 10^{-2.5}, 10^{-2}, \ldots, 10^{3}\}$. The noise variance σ^2 in RSIC and $\widehat{\mathrm{ESE}}_{\mathrm{RSIC}}$ is estimated by Eq.(55).

In this experiment, we measure the generalization error by the following criterion, which is equivalent to J without the expectation $\mathrm{E}_{\boldsymbol{\epsilon}}$ (see Eq.(30)):

$$\begin{aligned} \mathrm{Error}(\lambda) &= \|\hat{f}_\lambda\|_{\mathcal{H}}^2 - 2\langle \hat{f}_\lambda, f_{\mathcal{S}}\rangle_{\mathcal{H}} \\ &= \langle \boldsymbol{K}\boldsymbol{X}_\lambda \boldsymbol{y}, \boldsymbol{X}_\lambda \boldsymbol{y}\rangle - 2\langle \boldsymbol{X}_\lambda \boldsymbol{y}, \boldsymbol{z}\rangle, \end{aligned} \tag{56}$$

where $\hat{f}_\lambda$ denotes the learned function with a ridge parameter λ.

Figure 3 displays the values of Error(λ), SIC(λ), and RSIC(λ) as a function of the ridge parameter λ for each (n, σ^2) in Eq.(52). The horizontal axis denotes the values of λ in log-scale. From top, the graphs denote the mean Error with error bar, the mean SIC with error bar, and the mean RSIC with error bar. The mean is taken over 100 trials, and the error bar denotes the standard deviation over 100 trials. In order to clearly compare the mean curves, the mean Error is also drawn by the dashed line in the bottom two graphs.

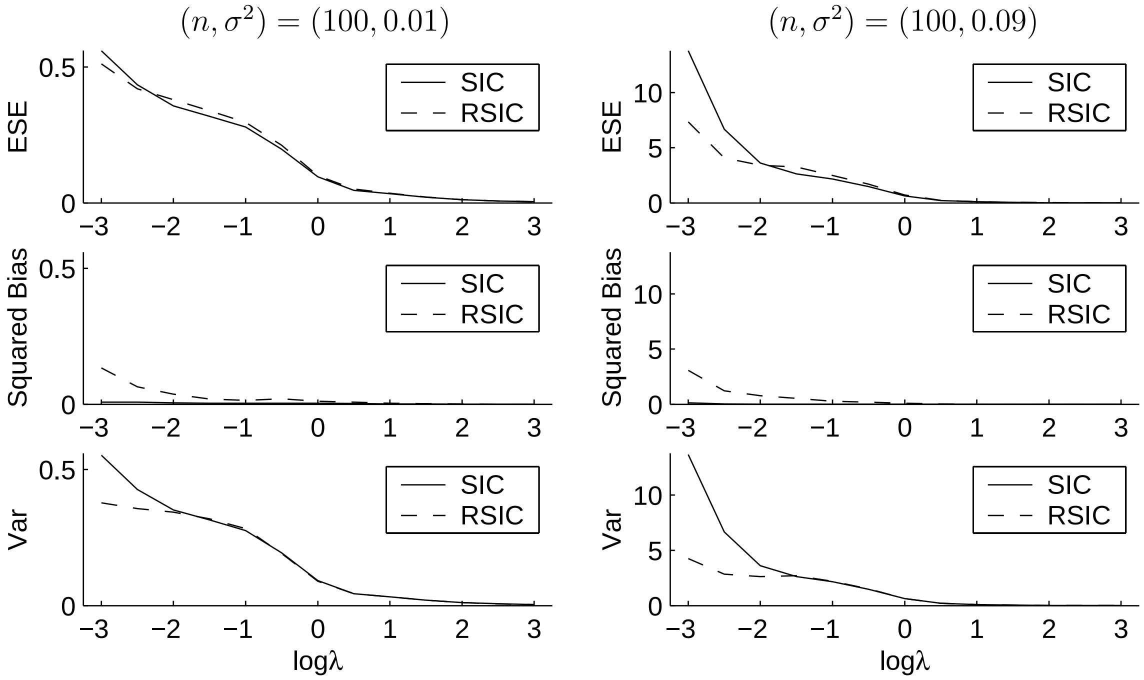

Figure 4: Values of ESE (35), Bias^2 (37), and Var (38) for SIC and RSIC. The horizontal axis denotes the value of λ in log-scale.

Figure 4 depicts the values of ESE (35), Bias^2 (37), and Var (38) of SIC and RSIC as a function of the ridge parameter λ. Note that, in this simulation, the expectation over the noise included in the definitions of ESE, Bias, and Var is replaced by the mean over 100 trials, where both the training sample points $\{x_i\}_{i=1}^n$ and noise $\{\epsilon_i\}_{i=1}^n$ are changed.

When $(n, \sigma^2) = (100, 0.01)$, the left graphs in Figure 3 show that the mean SIC seems to capture the mean Error very well and the size of the error bar looks reasonable. The mean RSIC looks almost the same as the mean SIC for medium/large λ, but the mean RSIC is slightly over-estimated for small λ. In exchange, the error bar of RSIC is slightly smaller than that of SIC for small λ. Indeed, the left graphs in Figure 4 show that for small λ, $\mathrm{Bias}^2_{\mathrm{RSIC}}$ is slightly larger than $\mathrm{Bias}^2_{\mathrm{SIC}}$ but $\mathrm{Var}_{\mathrm{RSIC}}$ is slightly smaller than $\mathrm{Var}_{\mathrm{SIC}}$. Consequently, $\mathrm{ESE}_{\mathrm{RSIC}}$ and $\mathrm{ESE}_{\mathrm{SIC}}$ are comparable. When $(n, \sigma^2) = (100, 0.09)$, the right graphs in Figure 3 show that the mean SIC still captures the mean Error very well. However, the size of the error bar is rather large for small λ. In contrast, the size of the error bars of RSIC is compressed for small λ, in exchange for the slight over-estimation of the mean RSIC for small λ. Indeed, the right graphs in Figure 4 show that while the variance is largely suppressed for small λ, the increase in the squared bias is relatively small. As a result, ESE is much improved for small λ and it stays almost the same for medium/large λ. When the number n of training examples is 50, all the results are almost identical to the case with $n = 100$. For this reason, we omit the graphs.

The above simulation results show that RSIC with $\widehat{\mathrm{ESE}}_{\mathrm{RSIC}}$ maintains the good performance of SIC when the noise level is low, and it highly improves the precision over SIC

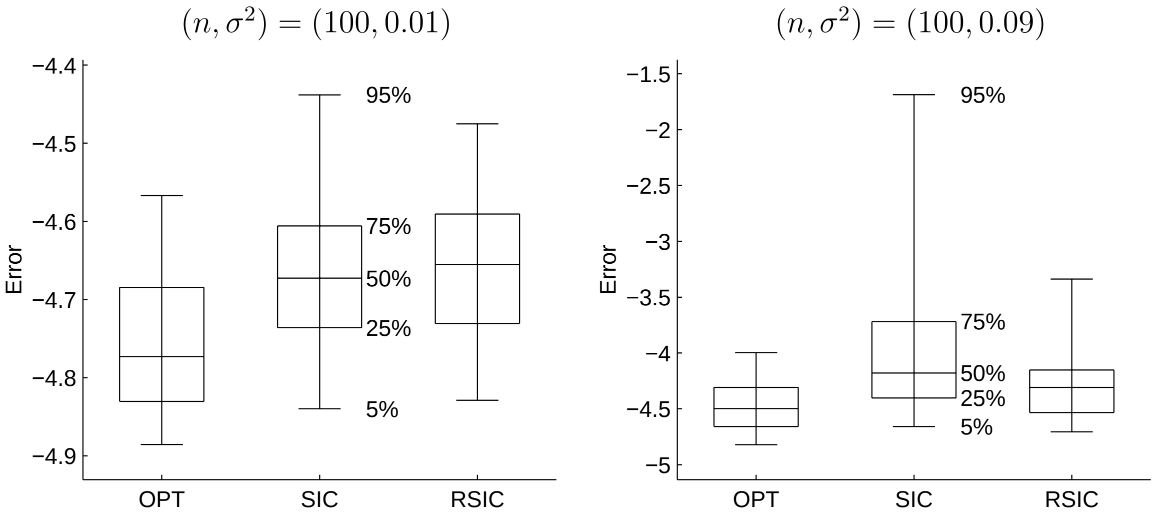

Figure 5: Box plot of Error obtained by the ridge parameter selected based on SIC or RSIC. The box plot notation specifies marks at 5, 25, 50, 75, and 95 percentiles of values from bottom. ‘OPT’ indicates the optimal choice of the ridge parameter. Note that the values of Error can be negative since a positive constant is ignored.

when the noise level is high. Furthermore, it is notable that the simulation results are almost unchanged even when the number of training examples is decreased. This may be a useful property in practice.

5.1.3 Investigating Model Selection Performance

Now we illustrate how SIC and RSIC work in model selection. We choose the ridge parameter λ from Eq.(53) so that SIC or RSIC is minimized. The goodness of the selected ridge parameter is again evaluated by the Error from Eq.(56).

Figure 5 depicts the values of Error obtained by the ridge parameter selected based on SIC or RSIC. The box plot notation specifies marks at 5, 25, 50, 75, and 95 percentiles from bottom. ‘OPT’ indicates the optimal choice of the ridge parameter, i.e., we actually calculate Error for each λ in Eq.(53) and selected the one that minimizes Error. Note that the values of Error can be negative since a positive constant is ignored in the definition of Error (56) (cf. Eq.(10)).

When $(n, \sigma^2) = (100, 0.01)$, the error obtained by RSIC is comparable to that of SIC (see the left plot in Figure 5), this fact is also confirmed by the 95% *t-test* (see e.g., Henkel, 1979). When $(n, \sigma^2) = (100, 0.09)$, the distributions of the error obtained by SIC and RSIC are comparable for 5, 25, and 50 percentiles, but RSIC improves 75 and 95 percentiles over SIC (see the right plot in Figure 5). The t-test says that RSIC surely improves over SIC. When the number n of training examples is 50, all the results are again similar to the case with $n = 100$ (although the the improvement of RSIC over SIC is not statistically significant when $(n, \sigma^2) = (50, 0.09)$). For this reason, we omit the plots.

The above model selection simulation results show that RSIC and SIC perform similarly when the noise level is low, and RSIC works better than SIC when the noise level is high. Especially, RSIC mostly improves higher percentiles of the obtained error (see Figure 5), from which we conjecture that RSIC is a robust model selection criterion against 'wicked' training sets.

5.2 Real Data Sets

In Section 5.1, we found that RSIC works well for a very simple artificial data set. Here we apply RSIC to real data sets, and evaluate whether this good property can be carried over to practical problems. We will use 10 practical data sets provided by DELVE (Rasmussen et al., 1996): *Abalone, Boston, Bank-8fm, Bank-8nm, Bank-8fh, Bank-8nh, Kin-8fm, Kin-8nm, Kin-8fh*, and *Kin-8nh*.

The *Abalone* data set includes 4177 samples, each of which consists of 9 physical measurements. The task is to estimate the last attribute (the age of abalones) from the rest. The first attribute is qualitative (male/female/infant) so it is ignored, i.e., 7-dimensional input and 1-dimensional output data is used. The *Boston* data set includes 506 samples with 13-dimensional input and 1-dimensional output data. The *'Bank'* data family consists of four different data sets. They are labeled as 'fm', 'nm', 'fh', and 'nh', 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 of the 4 data sets includes 8192 samples, consisting of 8-dimensional input and 1-dimensional output data. The 'Kin' data family also consists of four different data sets labeled as 'fm', 'nm', 'fh', and 'nh'. Each of the 4 data sets includes 8192 samples, consisting of 8-dimensional input and 1-dimensional output data.

For convenience, every attribute is normalized to $[0, 1]$. 100 randomly selected samples $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{100}$ are used for training. In the real data set, we can not measure the generalization error by Eq.(56) since neither the true function f nor its projection $f_{\mathcal{S}}$ is known. Instead, we evaluate the performance by the mean squared test error defined by

$$\text{Test Error} = \frac{1}{n'} \sum_{i=1}^{n'} \left(\hat{f}(\boldsymbol{x}'_i) - y'_i \right)^2, \tag{57}$$

where $\{(\boldsymbol{x}'_i, y'_i)\}_{i=1}^{n'}$ denote the test samples which are not used for training. A Gaussian kernel with width $c = 1$ is again employed (see Eq.(51)), and the kernel regression model (3) with ridge regression (8) is used for learning. The ridge parameter λ is selected from

$$\lambda \in \{10^{-3}, 10^{-2}, 10^{-1}, \ldots, 10^{3}\}. \tag{58}$$

As ridge parameter selection strategies, we compare SIC, RSIC, leave-one-out cross-validation (CV)[11] , and an empirical Bayesian method (EB) (Akaike, 1980). SIC is calculated by Eq.(54), where the noise variance σ^2 is estimated by Eq.(55). For each λ in

[11] For the kernel regression model (3), there are two possibilities of calculating the leave-one-out error. One is to use the full kernel regression model with n kernels all through the leave-one-out procedure, i.e., when one sample is left, the corresponding kernel function is kept. The other is to use the reduced

Table 1: Normalized mean test errors and their standard deviations. The results of the best method and all other methods with no significant difference (95% t-test) are described in italic face.

Data	SIC	RSIC	Cross Validation	Empirical Bayes
Abalone	*1.005 ± 0.050*	*1.015 ± 0.045*	*1.015 ± 0.043*	1.044 ± 0.083
Boston	*1.000 ± 0.218*	*1.000 ± 0.218*	1.113 ± 0.199	1.138 ± 0.178
Bank-8fm	*1.001 ± 0.066*	1.034 ± 0.100	1.040 ± 0.095	1.029 ± 0.092
Bank-8nm	*1.002 ± 0.063*	*1.013 ± 0.071*	1.023 ± 0.077	1.054 ± 0.090
Bank-8fh	1.081 ± 0.088	*1.037 ± 0.097*	1.063 ± 0.082	1.066 ± 0.104
Bank-8nh	1.062 ± 0.079	*1.008 ± 0.056*	*1.004 ± 0.050*	1.344 ± 0.113
Kin-8fm	*1.000 ± 0.077*	*1.000 ± 0.077*	*1.005 ± 0.093*	1.526 ± 0.253
Kin-8nm	*1.009 ± 0.060*	*1.006 ± 0.056*	1.078 ± 0.063	1.135 ± 0.025
Kin-8fh	1.046 ± 0.080	*1.022 ± 0.061*	*1.029 ± 0.067*	1.086 ± 0.045
Kin-8nh	1.160 ± 0.094	1.077 ± 0.091	*1.020 ± 0.031*	1.031 ± 0.047

Eq.(58), the regularization parameter γ in RSIC is chosen from $\{10^{-3}, 10^{-2}, 10^{-1}, \ldots, 10^{3}\}$ so that $\widehat{\mathrm{ESE}}_{\mathrm{RSIC}}$ is minimized. The noise variance σ^2 in RSIC and $\widehat{\mathrm{ESE}}_{\mathrm{RSIC}}$ is estimated by Eq.(55).

The simulation is repeated 100 times, randomly selecting the training set $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{100}$ from scratch in each trial (i.e., sampling without replacement). Note that the test set $\{(\boldsymbol{x}'_i, y'_i)\}_{i=1}^{n'}$ also varies in each trial.

Simulation results are summarized in Table 1. The table describes the normalized mean test errors and their standard deviations, where the values of the test error are normalized so that the mean test error obtained by the optimal ridge parameter is 1. The results of the best method and all other methods with no significant difference (95% t-test) are described in italic face.

The result shows that RSIC gives the best or comparable results for 8 out of 10 data sets. It is interesting to note that RSIC outperforms SIC for data sets with high noise (*Bank-8fh, Bank-8nh, Kin-8fh*, and *Kin-8nh* data sets), while RSIC gives fairly comparable results to SIC for data sets with medium noise (*Bank-8nm, Kin-8fm*, and *Kin-8nm* data sets). Therefore, RSIC can improve the degraded performance of SIC in the high noise cases, and it tends to maintain the good performance of SIC in the medium noise cases. In theory, we assumed that the noise $\{\epsilon_i\}_{i=1}^{n}$ are independently drawn from the normal distribution with mean zero and common variance. On the other hand, this assumption may not be fulfilled in the DELVE data sets. This implies that when using RSIC in practice, the above assumption on the noise does not have to be rigorously

kernel regression model with $n-1$ kernels in the leave-one-out procedure, i.e., when one sample is left, the corresponding kernel function is also left. We took the former standpoint and used the closed-formula for calculating the leave-one-out error (see e.g., Wahba, 1990; Orr, 1996).

Table 2: Normalized mean test errors and their standard deviations for the ridge regression with RSIC and the support vector regression with 10-fold cross-validation. The results of the significantly better method (95% t-test) are described in bold face.

Data	Ridge+RSIC	SVR+10CV
Abalone	$\mathbf{1.015 \pm 0.045}$	1.096 ± 0.118
Boston	1.000 ± 0.218	0.955 ± 0.198
Bank-8fm	$\mathbf{1.034 \pm 0.100}$	1.157 ± 0.148
Bank-8nm	$\mathbf{1.013 \pm 0.071}$	1.217 ± 0.168
Bank-8fh	$\mathbf{1.037 \pm 0.097}$	1.095 ± 0.127
Bank-8nh	1.008 ± 0.056	0.988 ± 0.104
Kin-8fm	$\mathbf{1.000 \pm 0.077}$	1.059 ± 0.143
Kin-8nm	$\mathbf{1.006 \pm 0.056}$	1.056 ± 0.103
Kin-8fh	1.022 ± 0.061	1.020 ± 0.091
Kin-8nh	1.077 ± 0.091	1.078 ± 0.101

satisfied. Compared with CV and EB, RSIC is comparable or better for most of the data sets.

From the above experimental results, we conjecture that RSIC should be regarded as a practical model selection criterion for choosing the ridge parameter.

Finally, we compare our results also with ε-support vector regression (ε-SVR) (Vapnik, 1998; Schölkopf and Smola, 2002), which became recently one of the most popular regression algorithms. In SVR, we used the same Gaussian kernel with width $c = 1$ (see Eq.(51)). The regularization parameter C and the tube width ε in SVR are chosen from a wide range of values using 10-fold cross-validation. We obtained the solutions of SVR by the SVM^{light} package (Joachims, 1999).

The simulation results are described in Table 2, where the results of the significantly better method (95% t-test) are described in bold face. The table shows that SVR works well for the *Boston, Bank-8nh, Kin-8fh*, and *Kin-8nh* data sets (although the 95% t-test does not say that they are significantly different from the results of the ridge regression with RSIC), and it tends to give larger errors for other data sets. Given the fact that the *Boston, Bank-8nh*, and *Kin-8fh* data sets may include large noise, the ε-insensitive loss seems to be more robust for such large noise cases (cf. Müller et al., 1998). However, SVR tends to give large errors for the given data sets that include small noise (*Bank-8fm, Bank-8nm, Kin-8fm*, and *Kin-8nm* data sets). Therefore, the ε-insensitive loss is not as effective as the squared loss on the medium/small noise cases considered in the table (see also Müller et al., 1998; Schölkopf and Smola, 2002). Note that the main difference between the ridge regression and SVR is the loss function: The ridge regression uses a squared loss (see Eq.(7)) while SVR uses the ε-insensitive loss. Which one will be advantageous, certainly depends on what noise type is inherent to the data generating process.

Note that the computation time for the ridge regression with RSIC is faster than that for SVR with cross-validation because latter requires retraining[12]. For this reason, we consider using ridge regression with RSIC to be advantageous in practice.

6 Conclusions and Outlook

In this paper, we proposed using Stein's idea in the context of model selection, i.e., we suggested that the use of a biased estimator, e.g., by means of regularization, can yield more stable and robust and thus better estimators of the generalization error than its unbiased counterpart. Thus we sacrificed the unbiasedness for the sake of variance reduction in a model selection criterion by actively optimizing and balancing out this bias/variance trade-off.

This general idea was applied for a particular criterion where we regularized the unbiased estimator of the expected generalization error called the subspace information criterion (SIC). Our approach was to directly estimate the expected squared error between the generalization error estimator and the expected generalization error, and determine the degree of regularization in the regularized SIC (RSIC) such that the estimator of the expected squared error is minimized. Computer simulations with artificial and real data sets showed that our approach surely contributes to obtaining a more precise estimator of the expected generalization error, and it can be successfully applied to the ridge parameter selection.

In this paper, we focused on the case that SIC is regularized by $\boldsymbol{X}_r$ given by Eq.(33). However, the proposed method for determining the degree of RSIC is valid for any type of regularization, i.e., the estimator of the expected squared error given by Eq.(47) does not depend on the *form* of $\boldsymbol{X}_r$. Finding improved ways of regularization in particular using domain knowledge, is left to future exploration. Furthermore, it would be interesting to extend the current framework such that efficient non-linear estimators such as the LASSO (Tibshirani, 1996) can be dealt with.

In Eq.(47), we gave an unbiased estimator of the expected squared error between RSIC and the expected generalization error. The simulation results reported in Section 5 showed that the unbiased estimator of the expected squared error contributes beneficially to stabilizing SIC. However, the unbiased estimator of the expected squared error can again have large variance because of its unbiasedness (see the experimental results reported in Sugiyama et al., 2003, for details). One of the promising future directions is to improve the unbiased estimator of the expected squared error, to further enhance the precision of RSIC.

The theoretical discussions given in this paper (Section 4) do not include the analysis of estimating the noise variance σ^2. From the simulation with artificial data sets (Section 5.1), the influence of estimating the noise variance σ^2 appears unproblematic because

[12] Note that retraining is not needed also for the ridge regression with leave-one-out cross-validation because the leave-one-out error can be calculated analytically (see e.g., Wahba, 1990; Orr, 1996).

the unbiasedness of SIC is almost satisfied, and therefore RSIC can improve the precision over SIC. It still remains open to see whether this property can be shown to always hold or not. Therefore, it is a further important step to investigate the influence of the noise variance estimation more formally.

Our previous work (Sugiyama and Müller, 2002) showed that a linear unbiased estimate of the projection $f_{\mathcal{S}}$ exists if and only if the regression model is included in the span of $\{K(\boldsymbol{x}, \boldsymbol{x}_i)\}_{i=1}^{n}$. For this reason, we chose to use the kernel regression model given by Eq.(3). However, due to this fact, SIC given in Sugiyama and Müller (2002) can not be used for selecting the kernel parameters (e.g., kernel width). In RSIC, on the other hand, the linear unbiased estimate of the projection $f_{\mathcal{S}}$ has not appeared explicitly anymore in the definition (see Eq.(34)). Therefore, in principle, RSIC could be applied to regression models which are not included in the span of $\{K(\boldsymbol{x}, \boldsymbol{x}_i)\}_{i=1}^{n}$, e.g., models with different kernel width. However, currently we are still utilizing the linear unbiased estimate of the projection $f_{\mathcal{S}}$ for determining the degree of regularization in RSIC (see Section 4.3). It is therefore interesting to devise other methods for determining the degree of regularization in RSIC that do not use the linear unbiased estimate of the projection $f_{\mathcal{S}}$, to enable an optimization of even the kernel parameters by RSIC.

In this paper, we pursued a better estimator of the generalization error. Another important issue in model selection research is to investigate the model selection performance. For several model selection criteria such as Mallows's C_L (Mallows, 1964, 1973) and the generalized cross-validation (Craven and Wahba, 1979; Wahba, 1990), asymptotic optimality of the choice of the model has been investigated throughly (Craven and Wahba, 1979; Wahba, 1985; Li, 1986). It will be instructive to see whether similar discussions can be made for SIC and RSIC.

Finally, another future direction is to apply our general idea of stabilizing model selection criteria to other existing criteria. For example, the leave-one-out error is shown to be an almost unbiased estimate of the expected generalization error (Luntz and Brailovsky, 1969, see also Schölkopf and Smola, 2002), but it can have a large variance. For this reason, it is often recommended to use 5- or 10-fold cross-validation (i.e., divide the training set into 5 or 10 disjoint sets). However, the number of folds in cross-validation actually controls the trade-off between the bias and variance of the cross-validation estimates of the expected generalization error. For this reason, it is highly important to determine the number of folds in cross-validation so that the expected squared error between the cross-validation estimate and the expected generalization error is minimized. We conjecture that the approach taken in this paper can also play an important role in this challenging problem.

Acknowledgements

We would like to thank Stefan Harmeling, Pavel Laskov, Koji Tsuda, Hidemitsu Ogawa, Hidetoshi Shimodaira, and Takafumi Kanamori for their valuable comments. Special thanks also go to Satoru Tanaka for finding an error in the proof. We are also grateful to the anonymous reviewers. Their comments greatly helped us to improve the readability of this manuscript. M. S. thanks Fraunhofer FIRST for warm hospitality during his research stay in Berlin. M. S. also acknowledges partial financial support from MEXT, Grants-in-Aid for Scientific Research, 14380158 and 14780262, and the Alexander von Humboldt Foundation. K.-R. M. acknowledges partial the BMBF under contract FKZ 01IBB02A.

A Sketch of Proof of Lemma 1

It follows from Eq.(34) that $\mathrm{E}_{\boldsymbol{\epsilon}}\mathrm{RSIC}$ is expressed as

$$\mathrm{E}_{\boldsymbol{\epsilon}}\mathrm{RSIC} \;=\; \langle \boldsymbol{X}^\top \boldsymbol{K}\boldsymbol{X}\boldsymbol{z}, \boldsymbol{z}\rangle + \sigma^2 \mathrm{tr}\left(\boldsymbol{X}^\top \boldsymbol{K}\boldsymbol{X}\right) - 2\langle \boldsymbol{X}_r^\top \boldsymbol{K}\boldsymbol{X}\boldsymbol{z}, \boldsymbol{z}\rangle. \tag{59}$$

Similarly, it follows from Eq.(30) that J is expressed as

$$J \;=\; \langle \boldsymbol{X}^\top \boldsymbol{K}\boldsymbol{X}\boldsymbol{z}, \boldsymbol{z}\rangle + \sigma^2 \mathrm{tr}\left(\boldsymbol{X}^\top \boldsymbol{K}\boldsymbol{X}\right) - 2\langle \boldsymbol{X}_u^\top \boldsymbol{K}\boldsymbol{X}\boldsymbol{z}, \boldsymbol{z}\rangle, \tag{60}$$

where only the third term is different from Eq.(59). Then $\mathrm{Bias}_{\mathrm{RSIC}}$ is expressed as

$$\mathrm{Bias}_{\mathrm{RSIC}} = \langle (2\boldsymbol{X}_u^\top \boldsymbol{K}\boldsymbol{X} - 2\boldsymbol{X}_r^\top \boldsymbol{K}\boldsymbol{X})\boldsymbol{z}, \boldsymbol{z}\rangle. \tag{61}$$

Eqs.(61) and (39) yield Eq.(41). ■

B Sketch of Proof of Lemma 2

Let $\boldsymbol{\epsilon}$ be the noise vector defined by

$$\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)^\top. \tag{62}$$

Then $\mathrm{Var}_{\mathrm{RSIC}}$ is expressed as

$$\begin{aligned}\mathrm{Var}_{\mathrm{RSIC}} \;&=\; \sigma^2\|(\boldsymbol{C}+\boldsymbol{C}^\top)\boldsymbol{z}\|^2 + \mathrm{E}_{\boldsymbol{\epsilon}}\langle \boldsymbol{C}\boldsymbol{\epsilon}, \boldsymbol{\epsilon}\rangle^2 + \sigma^4 \mathrm{tr}\left(\boldsymbol{C}\right)^2 \\ &\quad +2\mathrm{E}_{\boldsymbol{\epsilon}}\langle (\boldsymbol{C}+\boldsymbol{C}^\top)\boldsymbol{z}, \boldsymbol{\epsilon}\rangle\langle \boldsymbol{C}\boldsymbol{\epsilon}, \boldsymbol{\epsilon}\rangle - 2\sigma^4 \mathrm{tr}\left(\boldsymbol{C}\right)^2.\end{aligned} \tag{63}$$

On the other hand, it holds that

$$\mathrm{E}_{\boldsymbol{\epsilon}}\langle \boldsymbol{C}\boldsymbol{\epsilon}, \boldsymbol{\epsilon}\rangle^2 \;=\; \mathrm{E}_{\boldsymbol{\epsilon}} \sum_{i,j,k,l=1}^{n} \boldsymbol{C}_{i,j}\boldsymbol{C}_{k,l}\epsilon_i\epsilon_j\epsilon_k\epsilon_l, \tag{64}$$

$$\mathrm{E}_{\boldsymbol{\epsilon}}\langle (\boldsymbol{C}+\boldsymbol{C}^\top)\boldsymbol{z}, \boldsymbol{\epsilon}\rangle\langle \boldsymbol{C}\boldsymbol{\epsilon}, \boldsymbol{\epsilon}\rangle \;=\; \mathrm{E}_{\boldsymbol{\epsilon}} \sum_{i,j,k,l=1}^{n} (\boldsymbol{C}_{i,j}+\boldsymbol{C}_{j,i})\boldsymbol{C}_{k,l}z_i\epsilon_j\epsilon_k\epsilon_l, \tag{65}$$

where $\boldsymbol{C}_{i,j}$ denotes the (i,j)-th element of $\boldsymbol{C}$. It is known that when the random variable ϵ_i is drawn from the normal distribution with mean zero and variance σ^2, it holds that $\mathrm{E}_{\boldsymbol{\epsilon}}\epsilon_i^3 = 0$ and $\mathrm{E}_{\boldsymbol{\epsilon}}\epsilon_i^4 = 3\sigma^4$ (e.g., Lehmann, 1983). They imply that all terms in

$\mathrm{E}_{\boldsymbol{\epsilon}} \sum_{i,j,k,l=1}^{n} \boldsymbol{C}_{i,j}\boldsymbol{C}_{k,l}\epsilon_i\epsilon_j\epsilon_k\epsilon_l$ vanish except four cases: $i = j = k = l$, $i = j \neq k = l$, $i = k \neq j = l$, and $i = l \neq j = k$. Therefore, we have

$$\mathrm{E}_{\boldsymbol{\epsilon}}\langle \boldsymbol{C}\boldsymbol{\epsilon}, \boldsymbol{\epsilon}\rangle^2 = \sigma^4 \mathrm{tr}\,(\boldsymbol{C})^2 + \sigma^4 \mathrm{tr}\left(\boldsymbol{C}^\top \boldsymbol{C}\right) + \sigma^4 \mathrm{tr}\left(\boldsymbol{C}^2\right). \tag{66}$$

Similarly, all terms in $\sum_{i,j,k,l=1}^{n}(\boldsymbol{C}_{i,j} + \boldsymbol{C}_{j,i})\boldsymbol{C}_{k,l} z_i\epsilon_j\epsilon_k\epsilon_l$ vanish, i.e.,

$$\mathrm{E}_{\boldsymbol{\epsilon}}\langle (\boldsymbol{C} + \boldsymbol{C}^\top)\boldsymbol{z}, \boldsymbol{\epsilon}\rangle\langle \boldsymbol{C}\boldsymbol{\epsilon}, \boldsymbol{\epsilon}\rangle = 0. \tag{67}$$

Substituting Eqs.(66) and (67) into Eq.(63), we obtain Eq.(42). ■

C Sketch of Proof of Theorem 3

It holds that

$$\begin{aligned}\mathrm{E}_{\boldsymbol{\epsilon}}\langle \boldsymbol{B}\boldsymbol{y}, \boldsymbol{y}\rangle^2 = {} & \mathrm{Bias}_{\mathrm{RSIC}}^2 + \sigma^2 \mathrm{E}_{\boldsymbol{\epsilon}}\|(\boldsymbol{B} + \boldsymbol{B}^\top)\boldsymbol{y}\|^2 + 2\sigma^2 \mathrm{tr}\,(\boldsymbol{B})\,\mathrm{E}_{\boldsymbol{\epsilon}}\langle \boldsymbol{B}\boldsymbol{y}, \boldsymbol{y}\rangle \\ & -\sigma^4 \mathrm{tr}\left(\boldsymbol{B}^2 + \boldsymbol{B}^\top \boldsymbol{B}\right) - \sigma^4 \mathrm{tr}\,(\boldsymbol{B})^2\end{aligned} \tag{68}$$

from which we have Eq.(45). Similarly, it holds that

$$\mathrm{Var}_{\mathrm{RSIC}} = \mathrm{E}_{\boldsymbol{\epsilon}}\left(\sigma^2 \|(\boldsymbol{C} + \boldsymbol{C}^\top)\boldsymbol{y}\|^2 - \sigma^4 \mathrm{tr}\left(\boldsymbol{C}^2 + \boldsymbol{C}^\top \boldsymbol{C}\right)\right) \tag{69}$$

from which we have Eq.(46). ■

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, and A. F. M. Smith (Eds.), *Bayesian Statistics* (pp. 141–166). Valencia: University Press.

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

Bergman, S. (1970). *The Kernel Function and Conformal Mapping.* Providence, Rhode Island: The American Mathematical Society.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition.* Oxford: Clarendon Press.

Bousquet, O., and Elisseeff, A. (2002). Stability and generalization. *Journal of Machine Learning Research*, 2(Mar), 499–526.

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

Craven, P., and 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.

Cristianini, N., and Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge: Cambridge University Press.

Cucker, F., and Smale, S. (2002). On the mathematical foundation of learning. *Bulletin of the American Mathematical Society*, 39(1), 1–49.

Daubechies, I. (1992). *Ten Lectures on Wavelets*. Philadelphia and Pennsylvania: Society for Industrial and Applied Mathematics.

Devroye, L., Györfi, L., and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. New York: Springer.

Donoho, D. L. (1995). De-noising by soft thresholding. *IEEE Transactions on Information Theory*, 41(3), 613–627.

Donoho D. L., and Johnstone, I. M. (1994). Ideal spatial adaptation via wavelet shrinkage. *Biometrika*, 81, 425–455.

Felsenstein, J. (1985). Confidence limits on phylogenies: An approach using the bootstrap. *Evolution*, 39, 783–791.

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

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

Gu, C., Heckman, N., and Wahba, G. (1992). A note on generalized cross-validation with replicates. *Statistics & Probability Letters*, 14, 283–287.

Henkel, R. E. (1979). *Tests of Significance*. Beverly Hills: SAGE Publication.

Heskes, T. (1998). Bias/variance decompositions for likelihood-based estimators. *Neural Computation*, 10(6), 1425–1433.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(3), 55–67.

Joachims, T. (1999). Making large-scale SVM learning practical. In B. Schölkopf, C. J. C. Burges, and A. J. Smola (Eds.), *Advances in Kernel Methods — Support Vector Learning* (pp. 169–184). Cambridge, MA: The MIT Press.

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

Lehmann, E. L. (1983). *Theory of Point Estimation.* New York: Wiley.

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.

Linhart, H. (1988). A test whether two AIC's differ significantly. *South Africa Statistical Journal*, 22, 153–161.

Luntz, A., and Brailovsky, V. (1969). On estimation of characters obtained in statistical procedure of recognition. *Technicheskaya Kibernetica*, 3. (In Russian)

Mallows, C. L. (1964). Choosing a subset regression. *Presented at the Central Regional Meeting of the Institute of Mathematical Statistics.*

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

Müller, K.-R., Mika, S., Rätsch, G., Tsuda, K., and Schölkopf, B. (2001). An introduction to kernel-based learning algorithms. *IEEE Transactions on Neural Networks*, 12, 181–201.

Müller, K.-R., Smola, A. J. , Rätsch, G., Schölkopf, B., Kohlmorgen, J., and Vapnik, V. (1998). Using support vector machines for time series prediction. In B. Schölkopf, C. J. C. Burges, and A. J. Smola (Eds.), *Advances in Kernel Methods — Support Vector Learning*, 243–254. Cambridge, MA: The MIT Press.

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

Murata, N., Yoshizawa, S., and 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. URL `http://www.anc.ed.ac.uk/~mjo/papers/intro.ps.gz`.

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

Saitoh, S. (1988). *Theory of Reproducing Kernels and Its Applications.* UK: Longman Scientific & Technical.

Saitoh, S. (1997). *Integral Transforms, Reproducing Kernels and Their Applications.* UK: Longman.

Schölkopf, B., Smola, A. J., Williamson, R., and Bartlett, P. (2000). New support vector algorithms. *Neural Computation*, 12(5), 1207–1245.

Schölkopf, B., and Smola, A. J. *Learning with Kernels.* Cambridge, MA: MIT Press.

Shimodaira, H. (1997). Assessing the error probability of the model selection test. *Annals of Institute of Statistical Mathematics*, 49(3), 395–410.

Shimodaira, H. (1998). An application of multiple comparison techniques to model selection. *Annals of Institute of Statistical Mathematics*, 50(1), 1–13.

Smola, A. J., Schölkopf, B., and Müller, K.-R. (1998). The connection between regularization operators and support vector kernels. *Neural Networks*, 11(4), 637–649.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the 3rd Berkeley Symposium on Mathematical Statistics and Probability*, 1 (pp. 197–206), Berkeley, CA: University of California Press.

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., Kawanabe, M., and Müller, K.-R. (2003). Trading variance reduction with unbiasedness — The regularized subspace information criterion for robust model selection in kernel regression. Technical Report TR03-0003, Department of Computer Science, Tokyo Institute of Technology. URL `http://www.cs.titech.ac.jp/`.

Sugiyama, M., and Müller, K.-R. (2002). The subspace information criterion for infinite dimensional hypothesis spaces. *Journal of Machine Learning Research*, 3(Nov), 323–359.

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

Sugiyama, M., and Ogawa, H. (2002). Optimal design of regularization term and regularization parameter by subspace information criterion. *Neural Networks*, 15(3), 349–361.

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

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58(1), 267–288.

Tsuda, K., Sugiyama, M., and Müller, K.-R. Subspace information criterion for non-quadratic regularizers — Model selection for sparse regressors. *IEEE Transactions on Neural Networks*, 13(1), 70–80.

Vapnik, V. N. (1982). *Estimation of Dependencies Based on Empirical Data.* New York: Springer-Verlag.

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory.* Berlin: Springer-Verlag.

Vapnik, V. N. (1998). *Statistical Learning Theory.* New York: John Wiley & Sons, Inc.

Wahba, G. (1985). A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. *The Annals of Statistics*, 13(4), 1378–1402.

Wahba, G. (1990). *Spline Model for Observational Data.* Philadelphia and Pennsylvania: Society for Industrial and Applied Mathematics.

Williams, C. K. I. (1998). Prediction with Gaussian processes: From linear regression to linear prediction and beyond. In M. I. Jordan (Ed.), *Learning in Graphical Models* (pp. 599–621). Cambridge: The MIT Press.

Williams, C. K. I., and Rasmussen, C. E. (1996). Gaussian processes for regression. In D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo (Eds.), *Advances in Neural Information Processing Systems*, 8 (pp. 514–520). Cambridge, MA: The MIT Press.