Regularizing Generalization Error Estimators: A Novel Approach to Robust Model Selection

Masashi Sugiyama[1,2]*, Motoaki Kawanabe[1], Klaus-Robert Müller[1,3]

[1] Fraunhofer FIRST, IDA, Kekuléstr. 7, 12489 Berlin, Germany
[2] Department of Computer Science, Tokyo Institute of Technology
2-12-1, O-okayama, Meguro-ku, Tokyo, 152-8552, Japan
[3] Department of Computer Science, University of Potsdam
August-Bebel-Str.89, Haus 4, 14482 Potsdam, Germany

Abstract. A well-known result by Stein shows that regularized estimators with small bias often yield better estimates than unbiased estimators. In this paper, we adapt this spirit to model selection, and propose regularizing unbiased generalization error estimators for stabilization. We trade a small bias in a model selection criterion against a larger variance reduction which has the beneficial effect of being more precise on a single training set.

1 Introduction

Almost all learning algorithms proposed so far include some tuning (or hyper) parameters, and appropriately determining the values of such tuning parameters (i.e., model selection) is crucial for better generalization [8]. Usually, the values of the tuning parameters are set so that an estimator of the generalization error is minimized. So far, several unbiased estimators of the generalization error have been proposed, and they have been successfully used as model selection criteria in various practical learning tasks.

However, unbiased estimators of the generalization error can have large variance under some severe conditions, making model selection unstable. For this reason, it is very important to reduce the variance of the unbiased generalization error estimators for robust model selection. In this paper, we therefore propose a method for improving the precision of unbiased generalization error estimators by regularization. Since we are trying to shrink unbiased estimators

We would like to thank Stefan Harmeling, Pavel Laskov, Koji Tsuda, Hidemitsu Ogawa, Hidetoshi Shimodaira, and Takafumi Kanamori for their valuable comments. M. S. thanks the Alexander von Humboldt Foundation for partial financial support. K.-R. M. acknowledges partial financial support from DFG under contracts JA 379/92 and MU 987/11 and the BMBF under contract FKZ 01IBB02A.

*E-mail: sugi@cs.titech.ac.jp

of the generalization error, this work can be regarded as an application of the idea of the Stein estimator [3] to model selection.

We focus on the subspace information criterion (SIC) [6, 5], which is an unbiased estimator of the generalization error measured by the reproducing kernel Hilbert space norm. It was shown in earlier experiments [7] 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 estimator of the expected squared error between SIC and the generalization error, and propose determining the degree of regularization of SIC so that the estimator of the expected squared error is minimized.

2 Regression Problem and Model Selection

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 $\mathcal{D} \subset \mathbb{R}^d$. We are given a set of n *training examples*, each of which 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 with mean zero and variance σ^2. Namely the training examples are expressed as $\{(\boldsymbol{x}_i, y_i) \mid y_i = f(\boldsymbol{x}_i) + \epsilon_i\}_{i=1}^n$. In theory, we assume that σ^2 is known, although it is estimated from the training examples in experiments.

We assume that the unknown learning target function $f(\boldsymbol{x})$ belongs to a specified *reproducing kernel Hilbert space* (RKHS) $\mathcal{H}$. Let us denote the reproducing kernel of a functional Hilbert space $\mathcal{H}$ by $K(\boldsymbol{x}, \boldsymbol{x}')$. We will employ the following kernel regression model $\hat{f}(\boldsymbol{x})$:

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

where $\{\alpha_i\}_{i=1}^n$ are parameters to be estimated from training examples. We estimate the parameters in a linear fashion, i.e., estimated parameters $\{\hat{\alpha}_i\}_{i=1}^n$ are given by

$$\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_n)^\top = \boldsymbol{X}\boldsymbol{y}, \tag{2}$$

where $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^\top$ and $\boldsymbol{X}$ is an n-dimensional matrix that does not depend on the noise $\{\epsilon_i\}_{i=1}^n$. All the discussions in this paper are valid for any matrix $\boldsymbol{X}$, but for simplicity we focus on ridge estimation.

$$\hat{\boldsymbol{\alpha}}_\lambda = \boldsymbol{X}_\lambda \boldsymbol{y}, \quad \text{where} \quad \boldsymbol{X}_\lambda = (\boldsymbol{K}^2 + \lambda \boldsymbol{I})^{-1} \boldsymbol{K}. \tag{3}$$

$\boldsymbol{I}$ denotes the identity matrix and $\boldsymbol{K}$ is the so-called kernel matrix whose (i, j)-th element is $\boldsymbol{K}_{i,j} = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$.

We would like to determine the value of the ridge parameter λ so that $\hat{f}(\boldsymbol{x})$ well approximates 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 expected squared norm in the RKHS $\mathcal{H}$.

$$J_0(\lambda) \equiv \mathrm{E}_{\boldsymbol{\epsilon}} \|\hat{f}_\lambda - f\|_{\mathcal{H}}^2, \tag{4}$$

where $\mathrm{E}_{\boldsymbol{\epsilon}}$ denotes the expectation over the noise $\{\epsilon_i\}_{i=1}^n$ and $\|\cdot\|_{\mathcal{H}}$ 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. For further discussions on this generalization measure, readers may refer to [5].

The generalization error J_0 includes the unknown learning target function $f(\boldsymbol{x})$ so it can not be directly calculated. The subspace information criterion (SIC) [6, 5] is an estimator of it:

$$\mathrm{SIC}(\lambda) \equiv \langle \boldsymbol{K}\boldsymbol{X}_\lambda \boldsymbol{y}, \boldsymbol{X}_\lambda \boldsymbol{y} \rangle - 2\langle \boldsymbol{K}\boldsymbol{X}_\lambda \boldsymbol{y}, \boldsymbol{X}_u \boldsymbol{y} \rangle + 2\sigma^2 \mathrm{trace}\left(\boldsymbol{K}\boldsymbol{X}_\lambda \boldsymbol{X}_u^\top\right), \tag{5}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product in $\mathbb{R}^n$ and $\boldsymbol{X}_u = \boldsymbol{K}^\dagger$. SIC satisfies

$$\mathrm{E}_{\boldsymbol{\epsilon}}\mathrm{SIC}(\lambda) = J_0(\lambda) - \|f\|_{\mathcal{H}}^2 \equiv J(\lambda). \tag{6}$$

Since $\|f\|_{\mathcal{H}}^2$ is constant, SIC is an unbiased estimator of an essential part J of the generalization error J_0. The purpose of this paper is to improve the precision of SIC. To this end, we briefly review the derivation of SIC in the rest of this section.

Let $\boldsymbol{\alpha}^*$ be the parameter vector corresponding to the orthogonal projection $f_{\mathcal{S}}(\boldsymbol{x})$ of the unknown learning target function $f(\boldsymbol{x})$ onto the subspace $\mathcal{S}$ spanned by $\{K(\boldsymbol{x}, \boldsymbol{x}_i)\}_{i=1}^n$:

$$f_{\mathcal{S}}(\boldsymbol{x}) = \sum_{i=1}^n \alpha_i^* K(\boldsymbol{x}, \boldsymbol{x}_i). \tag{7}$$

Note that $f_{\mathcal{S}}(\boldsymbol{x})$ is the optimal approximation to the target function $f(\boldsymbol{x})$ in the kernel regression model (1). Letting $\|\boldsymbol{\alpha}\|_{\boldsymbol{K}}^2 = \langle \boldsymbol{K}\boldsymbol{\alpha}, \boldsymbol{\alpha} \rangle$, we can express the true generalization error J_0 by

$$J_0(\lambda) = \|\mathrm{E}_{\boldsymbol{\epsilon}}\hat{\boldsymbol{\alpha}}_\lambda - \boldsymbol{\alpha}^*\|_{\boldsymbol{K}}^2 + \sigma^2 \mathrm{trace}\left(\boldsymbol{K}\boldsymbol{X}_\lambda \boldsymbol{X}_\lambda^\top\right) + \|f_{\mathcal{S}} - f\|_{\mathcal{H}}^2, \tag{8}$$

where the first and second terms are the squared bias and variance of $\hat{\boldsymbol{\alpha}}_\lambda$, and the last one is constant. A key idea of SIC is that the bias term $\|\mathrm{E}_{\boldsymbol{\epsilon}}\hat{\boldsymbol{\alpha}}_\lambda - \boldsymbol{\alpha}^*\|_{\boldsymbol{K}}^2$ is roughly estimated by $\|\hat{\boldsymbol{\alpha}}_\lambda - \hat{\boldsymbol{\alpha}}_u\|_{\boldsymbol{K}}^2$, where $\hat{\boldsymbol{\alpha}}_u$ is a linear unbiased estimate of the optimal parameter $\boldsymbol{\alpha}^*$:

$$\mathrm{E}_{\boldsymbol{\epsilon}}\hat{\boldsymbol{\alpha}}_u = \boldsymbol{\alpha}^*, \quad \text{where} \quad \hat{\boldsymbol{\alpha}}_u = \boldsymbol{X}_u \boldsymbol{y} \quad \text{and} \quad \boldsymbol{X}_u = \boldsymbol{K}^\dagger. \tag{9}$$

Then an unbiased estimator of the bias term $\|\mathrm{E}_{\boldsymbol{\epsilon}}\hat{\boldsymbol{\alpha}}_\lambda - \boldsymbol{\alpha}^*\|_{\boldsymbol{K}}^2$ is given by

$$\|\hat{\boldsymbol{\alpha}}_\lambda - \hat{\boldsymbol{\alpha}}_u\|_{\boldsymbol{K}}^2 - \sigma^2 \mathrm{trace}\left(\boldsymbol{K}(\boldsymbol{X}_\lambda - \boldsymbol{X}_u)(\boldsymbol{X}_\lambda - \boldsymbol{X}_u)^\top\right). \tag{10}$$

Replacing the bias term $\|\mathrm{E}_{\boldsymbol{\epsilon}}\hat{\boldsymbol{\alpha}}_\lambda - \boldsymbol{\alpha}^*\|_{\boldsymbol{K}}^2$ in Eq.(8) by this unbiased estimator and ignoring some constant terms, we have SIC given by Eq.(5).

3 Regularization Approach to Stabilizing SIC

As shown in Eq.(6), SIC is an unbiased estimator of an essential part J of the generalization error J_0, and this good property still holds even in finite sample cases (i.e., non-asymptotic cases). It is demonstrated that SIC can be successfully applied to the ridge parameter selection when the noise level is low or medium [5]. 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.

Our previous work [7] showed that the instability of SIC is mainly caused by the large variance of the unbiased estimate $\hat{\boldsymbol{\alpha}}_u$, which played an essential role in the derivation of SIC. In order to reduce the variance of SIC, the paper [7] proposed replacing the linear unbiased estimate $\hat{\boldsymbol{\alpha}}_u$ by a linear regularized estimate $\hat{\boldsymbol{\alpha}}_\gamma$. Namely, the bias term $\|\mathrm{E}_{\boldsymbol{\epsilon}}\hat{\boldsymbol{\alpha}}_\lambda - \boldsymbol{\alpha}^*\|_{\boldsymbol{K}}^2$ in Eq.(8) is estimated by using $\|\hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}_\gamma\|_{\boldsymbol{K}}^2$ instead of $\|\hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}_u\|_{\boldsymbol{K}}^2$. The regularized estimate $\hat{\boldsymbol{\alpha}}_\gamma$ is slightly biased, so its expectation $\mathrm{E}_{\boldsymbol{\epsilon}}\hat{\boldsymbol{\alpha}}_\gamma$ no longer agrees with the true parameter $\boldsymbol{\alpha}^*$. However, the 'scatter' of $\hat{\boldsymbol{\alpha}}_\gamma$ may be far smaller than that of the unbiased estimate $\hat{\boldsymbol{\alpha}}_u$. The following discussion is valid for any linear estimator $\hat{\boldsymbol{\alpha}}_\gamma$, but here we focus on the ridge estimator for simplicity.

$$\hat{\boldsymbol{\alpha}}_\gamma = \boldsymbol{X}_\gamma \boldsymbol{y}, \quad \text{where} \quad \boldsymbol{X}_\gamma = (\boldsymbol{K}^2 + \gamma \boldsymbol{I})^{-1}\boldsymbol{K}. \tag{11}$$

γ (≥ 0) is the regularization parameter that controls the degree of regularization in SIC. We refer to SIC with $\boldsymbol{X}_u$ replaced by $\boldsymbol{X}_\gamma$ as the regularized SIC (RSIC):

$$\mathrm{RSIC}(\lambda;\gamma) \equiv \langle \boldsymbol{K}\boldsymbol{X}_\lambda \boldsymbol{y}, \boldsymbol{X}_\lambda \boldsymbol{y}\rangle - 2\langle \boldsymbol{K}\boldsymbol{X}_\lambda \boldsymbol{y}, \boldsymbol{X}_\gamma \boldsymbol{y}\rangle + 2\sigma^2 \mathrm{trace}\left(\boldsymbol{K}\boldsymbol{X}_\lambda \boldsymbol{X}_\gamma^\top\right). \tag{12}$$

The notation $\mathrm{RSIC}(\lambda;\gamma)$ means that RSIC is a function of the ridge parameter λ with a tuning parameter γ. Note that when $\gamma = 0$, RSIC agrees with the original SIC. It was experimentally shown that this regularization approach works effectively for stabilizing SIC [7]. However, the value of the regularization parameter γ should be appropriately determined, which is still a open problem. This is the problem that we tackle in this paper.

Let us consider the expected squared error (ESE) between RSIC and J, where J is an essential part of the generalization error J_0 (see Eq.(6)):

$$ESE_{\mathrm{RSIC}}(\gamma;\lambda) \equiv \mathrm{E}_{\boldsymbol{\epsilon}}(\mathrm{RSIC}(\lambda;\gamma) - J(\lambda))^2. \tag{13}$$

The notation $ESE_{\mathrm{RSIC}}(\gamma;\lambda)$ means that we treat ESE_{RSIC} as a function of the regularization parameter γ and ESE_{RSIC} depends on the ridge parameter λ. Our aim is to determine γ in RSIC so that the above ESE_{RSIC} is minimized.

An unbiased estimator of ESE_{RSIC} is given by

$$\begin{aligned}\widehat{ESE}_{\mathrm{RSIC}}(\gamma;\lambda) \;=\; & \langle \boldsymbol{B}\boldsymbol{y}, \boldsymbol{y}\rangle^2 - \sigma^2\|(\boldsymbol{B} + \boldsymbol{B}^\top)\boldsymbol{y}\|^2 - 2\sigma^2 \mathrm{trace}\,(\boldsymbol{B})\,\langle \boldsymbol{B}\boldsymbol{y}, \boldsymbol{y}\rangle \\ & + \sigma^4 \mathrm{trace}\left(\boldsymbol{B}^2 + \boldsymbol{B}^\top\boldsymbol{B}\right) + \sigma^4 \mathrm{trace}\,(\boldsymbol{B})^2 \\ & + \sigma^2\|(\boldsymbol{C} + \boldsymbol{C}^\top)\boldsymbol{y}\|^2 - \sigma^4 \mathrm{trace}\left(\boldsymbol{C}^2 + \boldsymbol{C}^\top\boldsymbol{C}\right),\end{aligned} \tag{14}$$

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 bold face.

Data	SIC	RSIC	Cross Validation	Empirical Bayes
Abalone	$\mathbf{1.0131 \pm 0.0002}$	$\mathbf{1.0144 \pm 0.0002}$	$\mathbf{1.0146 \pm 0.0002}$	1.0204 ± 0.0003
Boston	$\mathbf{1.0001 \pm 0.0007}$	$\mathbf{1.0016 \pm 0.0007}$	$\mathbf{1.0071 \pm 0.0007}$	1.1406 ± 0.0008
Bank-8fm	$\mathbf{1.0001 \pm 0.0001}$	1.0703 ± 0.0001	1.0708 ± 0.0001	1.0030 ± 0.0001
Bank-8nm	$\mathbf{1.0001 \pm 0.0004}$	$\mathbf{1.0002 \pm 0.0004}$	1.0461 ± 0.0005	1.0477 ± 0.0005
Bank-8fh	1.0604 ± 0.0004	$\mathbf{1.0025 \pm 0.0003}$	$\mathbf{1.0026 \pm 0.0003}$	$\mathbf{1.0003 \pm 0.0003}$
Bank-8nh	1.0987 ± 0.0004	$\mathbf{1.0028 \pm 0.0005}$	1.2177 ± 0.0008	1.4200 ± 0.0008
Kin-8fm	$\mathbf{1.0000 \pm 0.0001}$	$\mathbf{1.0000 \pm 0.0001}$	$\mathbf{1.0010 \pm 0.0001}$	1.4548 ± 0.0004
Kin-8nm	$\mathbf{1.0104 \pm 0.0011}$	$\mathbf{1.0097 \pm 0.0010}$	1.0241 ± 0.0007	1.0371 ± 0.0006
Kin-8fh	1.1103 ± 0.0002	$\mathbf{1.0021 \pm 0.0003}$	$\mathbf{1.0057 \pm 0.0003}$	1.2025 ± 0.0001
Kin-8nh	1.1015 ± 0.0008	1.0451 ± 0.0009	$\mathbf{1.0017 \pm 0.0004}$	1.0361 ± 0.0004

where $\boldsymbol{B}$ and $\boldsymbol{C}$ are n-dimensional matrices defined by

$$\boldsymbol{B} = 2\boldsymbol{X}_u^\top \boldsymbol{K}\boldsymbol{X}_\lambda - 2\boldsymbol{X}_\gamma^\top \boldsymbol{K}\boldsymbol{X}_\lambda \quad \text{and} \quad \boldsymbol{C} = \boldsymbol{X}_\lambda^\top \boldsymbol{K}\boldsymbol{X}_\lambda - 2\boldsymbol{X}_\gamma^\top \boldsymbol{K}\boldsymbol{X}_\lambda. \quad (15)$$

We propose determining the value of γ (for each ridge parameter λ) so that $\widehat{ESE}_{\mathrm{RSIC}}$ is minimized. Consequently, the ridge parameter λ is determined as follows.

$$\hat{\lambda}_{\mathrm{RSIC}} = \operatorname*{argmin}_{\lambda} \mathrm{RSIC}(\lambda; \hat{\gamma}_\lambda), \quad \text{where} \quad \hat{\gamma}_\lambda = \operatorname*{argmin}_{\gamma} \widehat{ESE}_{\mathrm{RSIC}}(\gamma; \lambda). \quad (16)$$

4 Simulations

In this section, the effectiveness of the proposed method is experimentally investigated by using 10 standard benchmark data sets provided by DELVE [2]. 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. For the DELVE data sets, we can not calculate the true generalization error since the true function f is unknown. Instead, we evaluate the performance by the mean squared test error using the samples not used for training. The Gaussian kernel with standard deviation 1 is employed, and the ridge parameter λ is selected from $\{10^{-3}, 10^{-2}, 10^{-1}, \ldots, 10^{3}\}$. As ridge parameter selection strategies, we compare SIC, RSIC, the leave-one-out cross-validation, and an empirical Bayesian method [1]. The noise variance σ^2 is estimated by

$$\hat{\sigma}_\lambda^2 = \|\boldsymbol{K}\boldsymbol{X}_\lambda \boldsymbol{y} - \boldsymbol{y}\|^2 / (n - \mathrm{trace}\,(\boldsymbol{K}\boldsymbol{X}_\lambda)). \quad (17)$$

The simulation is repeated 100 times for each data set, randomly selecting the training set $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{100}$ in each trial. Note that the test samples also vary in each trial.

Simulation results are summarized in Table 1, showing 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 the data sets with high noise (*Bank-8fh, Bank-8nh, Kin-8fh*, and *Kin-8nh*), while RSIC is fairly comparable to SIC for the data sets with medium noise (*Bank-8nm, Kin-8fm*, and *Kin-8nm*). Therefore, RSIC is shown to 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. From this result, we conjecture that RSIC should be regarded as a practical model selection criterion for choosing the ridge parameter.

5 Conclusions

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 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. Further theoretical analysis and experimental evaluation are included in an extended version [4].

References

[1] H. Akaike. Likelihood and the Bayes procedure. In N. J. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, editors, *Bayesian Statistics*, pages 141–166, Valencia, 1980. University Press.

[2] C. E. Rasmussen, R. M. Neal, G. E. Hinton, D. van Camp, M. Revow, Z. Ghahramani, R. Kustra, and R. Tibshirani. The DELVE manual, 1996.

[3] C. Stein. 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*, volume 1, pages 197–206, Berkeley, CA., USA, 1956. University of California Press.

[4] M. Sugiyama, M. Kawanabe, and K.-R. Müller. Trading variance reduction with unbiasedness — The regularized subspace information criterion for robust model selection in kernel regression. *Neural Computation*, 2004. to appear.

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

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

[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, Inc., New York, 1998.