Computer and Information Science; Vol. 8, No. 1; 2015
ISSN 1913-8989 E-ISSN 1913-8997
Published by Canadian Center of Science and Education

Kernel-Based Information Criterion

Somayeh Danafar[1], Kenji Fukumizu[3] & Faustino Gomez[2]

[1] Informatics Department, Università dellla Svizzera Italiana (USI), Lugano, Switzerland

[2] Istituto Dalle Molle di Studi sull'Intelligenza Artificiale (IDSIA)/SUPSI, Lugano, Switzerland

[3] The Institute of Statistical Mathematics (ISM), Tokyo, Japan

Correspondence: Somayeh Danafar, Informatics Department, USI, Lugano, Switzerland. E-mail: somayeh.danafar@usi.ch

Received: October 29, 2014 Accepted: November 3, 2014 Online Published: January 2, 2015

doi:10.5539/cis.v8n1p10 URL: http://dx.doi.org/10.5539/cis.v8n1p10

The research is financed by FNSNF grants (P1TIP2 148352, PBTIP2 140015).

Abstract

This paper introduces Kernel-based Information Criterion (KIC) for model selection in regression analysis. The kernel-based complexity measure in KIC efficiently computes the interdependency between parameters of the model using a novel variable-wise variance and yields selection of better, more robust regressors. Experimental results show superior performance on both simulated and real data sets compared to Leave-One-Out Cross-Validation (LOOCV), kernel-based Information Complexity (ICOMP), and maximum log of marginal likelihood in Gaussian Process Regression (GPR).

Keywords: kernel-based model selection, information criterion, kernel ridge regression, complexity measure

1. Introduction

Model selection is an important problem in many areas including machine learning. If a proper model is not selected, any effort for parameter estimation or prediction of the algorithm's outcome is hopeless. Given a set of candidate models, the goal of model selection is to select the model that best approximates the observed data and captures its underlying regularities. Model selection criteria are defined such that they strike a balance between the Goodness-of-fit (GoF), and the generalizability or complexity of the models.

$$\text{Model selection criterion} = \text{GoF} + \text{Complexity} \tag{1}$$

Goodness-of-fit measures how well a model captures the regularity in the data. Generalizability/complexity is the assessment of the performance of the model on unseen data or how accurately the model fits/predicts the future data. Models with higher complexity than necessary can suffer from overfitting and poor generalization, while models that are too simple will underfit and has low GoF [3].

Cross-validation [2], bootstrapping [8], Akaike Information Criterion (AIC) [1], and Bayesian Information Criterion (BIC) [13], are well known examples of traditional model selection. In re-sampling methods such as cross-validation and bootstraping, the generalization error of the model is estimated using Monte Carlo simulation [9]. In contrast with re-sampling methods, the model selection methods like AIC and BIC do not require validation to compute the model error, and are computationally efficient. In these procedures an Information Criterion is defined such that the generalization error is estimated by penalizing the model's error on observed data. A large number of information criteria have been introduced with different motivations that lead to different theoretical properties. For instance, the tighter penalization parameter in BIC favors simpler models, while AIC works better when the dataset has a very large sample size.

Kernel methods are strong, computationally efficient analytical tools that are capable of working on high dimensional data with arbitrarily complex structure. They have been successfully applied in wide range of applications such as classification, and regression. In kernel methods, the data are mapped from their original space to a higher dimensional feature space, the Reproducing Kernel Hilbert Space (RKHS). The idea behind this mapping is to transform the nonlinear relationships between data points in the original space into an easy-to-compute linear learning problem in the feature space. For example, in kernel regression the response variable is described as a linear combination of the embedded data.

Any algorithm that can be represented through dot products has a kernel evaluation. This operation, called kernelization, makes it possible to transform traditional, already proven, model selection methods into stronger, corresponding kernel methods. The literature on kernel methods has, however, mostly focused on kernel selection and on tuning the kernel parameters, but only limited work being done on kernel-based model selection [5, 7, 11, 18]. In this study, we investigate a kernel-based information criterion for ridge regression models. In Kernel Ridge Regression (KRR), tuning the ridge parameters to find the most predictive subspace with respect to the data at hand and the unseen data is the goal of the kernel model selection criterion.

In classical model selection methods the performance of the model selection criterion is evaluated theoretically by providing a consistency proof where the sample size tends to infinity and empirically through simulated studies for finite sample sizes (Note 1). Other methods investigate a probabilistic upper bound of the generalization error [16].

Proving the consistency properties of the model selection in kernel model selection is challenging. The proof procedure of the classical methods does not work here. Some reasons for that are: the size of the model to evaluate problems such as under/overfitting [3] is not apparent (for n data points of dimension p, the kernel is $n \times n$, which is independent of p) and asymptotic probabilities of generalization error or estimators are hard to compute in RKHS.

Researchers have kernelized the traditional model selection criteria and shown the success of their kernel model selection empirically. Kobayashi and Komaki [7] extracted the kernel-based Regularization Information Criterion (KRIC) using an eigenvalue equation to set the regularization parameters in Kernel Logistic Regression and Support Vector Machines (SVM). Rosipal et al. [11] developed Covariance Information Criterion (CIC) for model selection in Kernel Principal Component Analysis, because of its outperformed results compared to AIC and BIC in orthogonal linear regression. Demyanov et al. [5], provided alternative way of calculating the likelihood function in Akaike Information Criterion (AIC, [1] and Bayesian Information Criterion (BIC, [13]), and used it for parameter selection in SVMs using the Gaussian kernel.

As pointed out by van Emden [15], a desirable model is the one with the fewest dependent variables. Thus defining a complexity term that measures the interdependency of model parameters enables one to select the most desirable model. In this study, we define a novel variable-wise variance and obtain a complexity measure as the additive combination of kernels defined on model parameters. Formalizing the complexity term in this way effectively captures the interdependency of each parameter of the model. We call this method Kernel-based Information Criterion (KIC).

Model selection criterion in Gaussian Process Regression (GPR; [12]), and kernel-based Information Complexity (ICOMP; [18]) resemble KIC in using a covariance-based complexity measure. However, the methods differ because these complexity measures capture the interdependency between the data points rather than the model parameters.

Although we can not establish the consistency properties of KIC theoretically, we empirically evaluate the efficiency of KIC both on synthetic and real datasets obtaining state-of-the-art results compared to Leave-One-Out-Cross-Validation (LOOCV), kernel-based ICOMP, and maximum log marginal likelihood in GPR. The paper is organized as follows. In section 1, we give an overview of Kernel Ridge regression. KIC is described in detail in section 2. Section 3 provides a brief explanation of the methods to which KIC is compared, and in section 4 we evaluate the performance of KIC through sets of experiments.

2. Kernel Ridge Regression

In regression analysis, the regression model is in the form:

$$Y = \boldsymbol{X}\theta + \varepsilon, \tag{2}$$

where f can be either a linear or non-linear function.

In linear regression we have, $Y = \boldsymbol{X}\theta + \varepsilon$, where Y is an observation vector (response variable) of size $n \times 1$, $\boldsymbol{X}$ is a full rank data matrix of independent variables of size $n \times p$, and $\theta = (\theta_1, \dots, \theta_p)^T$, is an unknown vector of regression parameters, where T denotes the transposition. We also assume that the error (noise) vector ε is an n-dimensional vector whose elements are drawn i.i.d, $\mathcal{N}(0, \sigma^2 \mathrm{I})$, where I is an n-dimensional identity matrix and σ^2 is an unknown variance.

The regression coefficients minimize the squared errors, $\left\|\hat{f}-f\right\|^2$, between estimated function $\hat{f}$, and target function f. When $p>n$, the problem is ill-posed, so that some kind of regularization, such as Tikhanov regularization (ridge regression) is required, and the coefficients minimize the following optimization problem

$$\operatorname{argmin}\,(Y-\boldsymbol{X}\theta)^T(Y-\boldsymbol{X}\theta)+\alpha\theta^T\theta, \tag{3}$$

where α is the regularization parameter. The estimated regression coefficients in ridge regression $\hat{\theta}$ are:

$$\hat{\theta}=(\boldsymbol{X}^T\boldsymbol{X}+\alpha I)^{-1}\boldsymbol{X}^TY,$$
$$=\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T+\alpha I)^{-1}Y. \tag{4}$$

In *Kernel Ridge Regression (KRR),* the data matrix $\boldsymbol{X}$ is non-linearly transformed in RKHS using a feature map $\phi(\boldsymbol{X})$. The estimated regression coefficients based on $\phi(\cdot)$ are:

$$\hat{\theta}=\phi(\boldsymbol{X})^T(K+\alpha I)^{-1}Y, \tag{5}$$

where $K=\phi(\boldsymbol{X})\phi(\boldsymbol{X})^T$ is the kernel matrix. Equation 5 does not obtain an explicit expression for θ because of $\phi(\boldsymbol{X})$ (the kernel trick enables one to avoid explicitly defining $\phi(\cdot)$ that could be numerically intractable to be computed in RKHS if known), thus a ridge estimator is used (e.g. [18]) that excludes $\phi(\boldsymbol{X})$:

$$\theta^*=(K+\alpha I)^{-1}Y. \tag{6}$$

Using θ^* in the calculation of KRR is similar to regularizing the regression function instead of the regression coefficients, where the objective function is:

$$\hat{f}=\operatorname{argmin}_{f\in\mathcal{H}}\big(Y-f(\boldsymbol{X})\big)^T(y-f(\boldsymbol{X}))+\alpha\|f\|_{\mathcal{H}}^2, \tag{7}$$

and $\mathcal{H}$ denotes the relevant RKHS. For $f=K\theta$, and $\{(X_1, Y_1),\ldots, (X_n,Y_n)\}$ we have:

$$\hat{f}=\sum_{i=1}^{n}\theta_i k(\cdot,X_i), \tag{8}$$
$$\hat{\theta}=\operatorname{argmin}_{\hat{\theta}\in\mathbb{R}^n}\|Y-K\theta\|_2^2+\alpha\,\theta^TK\theta, \tag{9}$$

where k is the kernel function, and $\hat{\theta}=(K+\alpha I)^{-1}Y$.

3. Kernel-Based Information Criterion

The main contribution of this study is to introduce a new Kernel-based Information Criterion (KIC) for the model selection in kernel-based regression. According to equation (1) KIC balances between the goodness-of-fit and the complexity of the model. GoF is defined using a log-likelihood-based function (we maximize penalized log likelihood) and the complexity measure is a function based on the covariance function of the parameters of the model. In the next subsections we elaborate on these terms.

3.1 Complexity Measures

The definition of van Emden [15] for the complexity measure of a random vector is based on the interactions among random variables in the corresponding covariance matrix. A desirable model is the one with the fewest dependent variables. This reduces the information entropy and yields lower complexity. In this paper we focus on this definition of the complexity measures.

Considering a p-variate normal distribution $f(X)=f\big(X_1,..,X_p\big)=\mathcal{N}(\mu,\Sigma)$, the complexity of a covariance matrix Σ, is given by the Shannon's entropy [14],

$$C(\Sigma)=\sum_{j=1}^{P}J(X_j)-J(X_1,\ldots,X_P),$$
$$=\tfrac{1}{2}\textstyle\sum_{j=1}^{P}\log\sigma_{jj}-\tfrac{1}{2}\log|\Sigma|, \tag{10}$$

where $J\big(X_j\big),J(X_1,\ldots,X_P)$ are the marginal and joint entropy, and σ_{jj} is the j-th diagonal elements of Σ. $C(\Sigma)=0$ if and only if the covariates are independent. The complexity measure in equation (10) changes with orthonormal transformations because it is dependent on the coordinates of the random variable vectors $X_1,\ldots, X_p$ [4]. To overcome these drawbacks, Bozodgan and Haughton [4] introduced ICOMP information criterion with a complexity measure based on the maximal covariance complexity, which is an upper bound on the complexity measure in equation (10):

$$C(\Sigma) = \frac{1}{2}\log\left(\left(\frac{\mathrm{tr}(\Sigma)}{P}\right)^{P} / |\Sigma|\right) = \frac{P}{2}\log(\bar{\lambda}_a/\bar{\lambda}_g). \tag{11}$$

This complexity measure is proportional to the estimated arithmetic$(\bar{\lambda}_a)$ and geometric mean $(\bar{\lambda}_g)$of the eigenvalues of the covariance matrix. Larger values of $C(\Sigma)$, indicates higher dependency between random variables, and vice versa. Zhang [18] introduced a kernel form of this complexity measure $C(\Sigma)$, that is computed on kernel-based covariance of the ridge estimator:

$$\Sigma_{\theta^*} = \sigma^2(\mathrm{K} + \alpha\mathrm{I})^{-2}. \tag{12}$$

The complexity measure in Gaussian Process Regression (GPR; [12]) is defined as $\frac{1}{2}\log|\Sigma|$, a concept from the joint entropy $H(X_1, \dots, X_P)$ (as shown in equation 10).

In contrast to ICOMP and GPR, the complexity measure in KIC is defined using the Hilbert-Schmidt (HS) norm of the covariance matrix, $C(\Sigma) = ||\Sigma||_{HS}^2 = \mathrm{tr}(\Sigma^T\Sigma)$. Minimizing this complexity measure obtains a model with more independent variables. In the next sections, we explain in detail how to define the needed variable-wise variance in the complexity measure, and the computation of the complexity measure.

3.1.1 Variable-Wise Variance

In kernel-based model selection methods such as ICOMP, and GPR, the complexity measure is defined on a covariance matrix that is of size $n \times n$ for $\boldsymbol{X}$ of size $n \times p$. The idea behind this measure is to compute the interdependency between the model parameters, which independent of the number of the model parameters p. In the other words, the concept of the size of the model is hidden because of the definition of a kernel. To have a complexity measure that depends on p, we introduce variable-wise variance using an additive combination of kernels for each parameter of the model.

Let $\theta \in \mathcal{H}$ be the parameter vector of the kernel ridge regression:

$$\theta = Y(K + \alpha I)^{-1}\mathbf{k}(\cdot), \tag{13}$$

where $Y = (Y_1, \dots, Y_n)$ and $\mathbf{k}(\cdot) = (k(\cdot, X_1), \dots, k(\cdot, X_n))^T$, and $\boldsymbol{X} = \begin{bmatrix} X_1^1 & \cdots & X_p^1 \\ \vdots & \ddots & \vdots \\ X_1^n & \cdots & X_p^n \end{bmatrix}$. The solution of KRR is given by $f(x) = \langle\theta, k(\cdot, x)\rangle$. The quantity $\mathrm{tr}[\Sigma_\theta] = \sigma^2\mathrm{tr}[K(K + \alpha I)^{-2}]$ can be interpreted as sum of variances for the component-wise parameter vectors, if the following sum of component-wise kernel is introduced:

$$k(x, \tilde{x}) = \sum_{j=1}^{p} k_j(x_j, \tilde{x}_j), \tag{14}$$

where x_j and $\tilde{x}_j$ denote the j-th component of vectors x and $\tilde{x} \in \mathbb{R}^p$. With this sum kernel, the function $f \in \mathcal{H}$ can be written as:

$$f = g_1(x_1) + \cdots + g_p(x_p), \tag{15}$$

where g_j is a function in $\mathcal{H}_j$, the RKHS defined by k_j. The parameter θ in this case is given by

$$\theta = Y(K + \alpha I)^{-1}\left(\sum_j \mathbf{k}_j(\cdot)\right) = \sum_{j=1}^{p}\theta_j, \tag{16}$$

where $\theta_j = Y(K + \alpha I)^{-1}\mathbf{k}_j(\cdot)$, and thus g_j in equation 15 is equal to θ_j. Let V_j be the conditional covariance of θ_j or g_j given $(X^1, \dots, X^n)$. We have

$$V_j = \sigma^2\mathrm{tr}\left[K_j(K + \alpha I)^{-2}\right], \tag{17}$$

where $K_{j,ab} = K_j(X_j^a, X_j^b)$ be the Gram matrix with k_j. Since $K_{ab} = \sum_{j=1}^{P} K_{j,ab}$, we have

$$\sum_{j=1}^{p} V_j = \sigma^2\mathrm{tr}\,[K(K + \alpha I)^{-2}] = \mathrm{tr}[\Sigma_\theta]. \tag{18}$$

Formalizing the complexity term with variable-wise variance effectively captures the interdependency of each parameter of the model (measures the significance of the contribution by the variables) explicitly.

3.1.2 Hilbert-Schmidt Independence Criterion

Gretton et al. [6] introduced a kernel-based independence measure, namely the Hilbert-Schmidt Independence criterion (HSIC), which is explained here. Suppose $X \in \mathcal{X}$, and $Y \in \mathcal{Y}$ are random vectors with feature maps $\phi: \mathcal{X} \rightarrow \mathcal{U}$ and $\psi: \mathcal{Y} \rightarrow \mathcal{V}$, where $\mathcal{U}$, and $\mathcal{V}$ are RKHSs. The cross-covariance operator corresponding to the joint probability distribution P_{XY} is a linear operator, $\Sigma_{XY}: \mathcal{V} \rightarrow \mathcal{U}$ such that:

$$\Sigma_{XY} \coloneqq E_{XY}[(\phi(X)-\mu_X)\otimes(\psi(Y)-\mu_Y)], \tag{19}$$

where $\otimes$ denotes the tensor product, $\mu_X = E_X[u(X)] = E[k(\cdot, X)]$, and $\mu_Y = E_Y[v(Y)] = E[k(\cdot, Y)]$, for $u \in \mathcal{U}$, $v \in \mathcal{V}$, and associated kernel function k. The HSIC measure for separable RKHS $\mathcal{U}$, and $\mathcal{V}$ is the squared HS-norm of the cross-covariance operator and is denoted as:

$$\text{HSIC}(P_{XY}, \mathcal{U}, \mathcal{V}) \coloneqq \|\Sigma_{XY}\|_{HS}^2 = \text{tr}\left[\Sigma_{XY}{}^T \Sigma_{XY}\right]. \tag{20}$$

Theorem 1. *Assume $\mathcal{X}$, and $\mathcal{Y}$ are compact, for all $u \in \mathcal{U}$, and $v \in \mathcal{V}$, $\|u\|_\infty \leq 1$, and $\|v\|_\infty \leq 1$, $\|\Sigma_{XY}\|_{HS} = 0$ if and only if X, and Y are independent (Theorem 4 in [6]).*

By computing the HSIC on covariance matrix associated with model's parameters Σ_θ we can measure the independence between the parameters. Since Σ_θ is a symmetric positive semi-definite matrix, $\Sigma^T\Sigma = \Sigma^2$, and the trace of the HS norm of the covariance matrix is equal to:

$$\|\Sigma_\theta\|_{HS}^2 = \text{tr}\left[\Sigma_\theta{}^T\,\Sigma_\theta\right] = \textstyle\sum_{j=1}^{p} V_j^2,$$
$$= \sigma^4 \text{tr}[K(K+\alpha I)^{-2}K(K+\alpha I)^{-2}]\,. \tag{21}$$

3.2 Kernel-Based Information Criterion

KIC is defined as:

$$\text{KIC}(\hat{\Sigma}_{\hat{\theta}}) = -2 \text{ Penalized Log} - \text{Likelihood}(\hat{\theta}) + C(\hat{\Sigma}_{\hat{\theta}}), \tag{22}$$

where $C(\hat{\Sigma}_{\hat{\theta}}) = \left\|\hat{\Sigma}_{\hat{\theta}}\right\|_{HS}^2/\sigma^4$ is the complexity term based on equation 21. The normalization by σ^4 obtains a complexity measure that is robust to changes in variance (similar to ICOMP criterion). The minimum KIC defines the best model (Note 2). The Penalized log-likelihood (PLL) in KRR for normally distribution data is defined by:

$$\text{PLL}(\theta, \sigma^2) = -\tfrac{n}{2}\log 2\pi - \tfrac{n}{2}\log\sigma^2 - \tfrac{(Y-K\theta)^T(Y-K\theta)}{2\sigma^2} - \alpha\left(\tfrac{\theta^T K\theta}{2\sigma^2}\right), \tag{23}$$

The unknown parameters $\hat{\theta}$, and $\hat{\sigma}^2$ are calculated by minimizaing the KIC objective function.

$$\frac{\partial \text{KIC}}{\partial\theta} = 0 \rightarrow \hat{\theta} = (K+\alpha I)^{-1}Y, \tag{24}$$

$$\frac{\partial \text{KIC}}{\partial\sigma^2} = 0 \;\rightarrow \hat{\sigma}^2 = \tfrac{(Y-K\theta)^T(Y-K\theta)+\alpha\,\theta^T K\theta}{n}. \tag{25}$$

We also investigated the effect of using $\text{tr}[\Sigma_\theta^T\Sigma_\theta]$, and $\text{tr}[\Sigma_\theta]$ as complexity terms. The empirical results reported in section 5.2. on real datasets, and compared with KIC. We denote those information criteria as:

$$\text{KIC1} = -2\text{ PLL} + \sigma^2\text{tr}[K(K+\alpha I)^{-2}], \tag{26}$$

$$\text{KIC2} = -2\text{ PLL} + \sigma^4\text{tr}\,[K(K+\alpha I)^{-2}K(K+\alpha I)^{-2}]. \tag{27}$$

In both KIC1, and KIC2, similar to KIC, $\hat{\theta} = (K+\alpha I)^{-1}Y$, while because the complexity term is dependent on σ^2, $\hat{\sigma}^2$ for KIC1 is:

$$\frac{\partial \text{KIC}}{\partial\sigma^2} = 0 \rightarrow \tfrac{n}{\sigma^2} - \tfrac{D}{\sigma^4} + \text{tr}[K(K+\alpha I)^{-2}] = 0. \tag{28}$$

If we denote $\sigma^2 = Z$, $\hat{\sigma}^2$ is the solution of a quadratic optimization problem, $C(\Sigma)Z^2 + nZ - D = 0$, where $D = (Y-K\theta)^T(Y-K\theta) + \alpha\,\theta^T K\theta$. In the case of KIC2, the $\hat{\sigma}^2$ is the real root of the following cubic problem:

$$2C(\Sigma)Z^3 + nZ - D = 0, \tag{29}$$

where $C(\Sigma) = \text{tr}[K(K+\alpha I)^{-2}K(K+\alpha I)^{-2}]$.

4. Other Methods

We compared KIC with LOOCV [17], kernel-based ICOMP [18], and maximum log of marginal likelihood in GPR (abbreviated as GPR) [12] to find the optimal ridge regressors. The reason to compare KIC with ICOMP and GPR is that in all of these methods the complexity measure computes the interdependency of model parameters as a function of covariance matrix in different ways. LOOCV is a standard and commonly used methods for model selection.

LOOCV: Re-sampling model selection methods like cross-validation is time consuming [2]. For instance, the Leave-One-Out-Cross-Validation (LOOCV) has the computational cost of $l \times O(A) \times$ the number of parameter combinations (O(A) is the processing time of the model selection algorithm A) for $n-1$ training samples. To have cross-validation methods with faster processing time, the closed form formula for the risk estimators of the algorithm under special conditions are provided. We consider the kernel-based closed form of LOOCV for linear regression introduced by Wahba [17]:

$$\text{Squared Error}_{\text{LOOCV}} = \frac{\left\|[diag(I-H)]^{-1}\,[I-H]\,Y\right\|_2^2}{2}, \tag{30}$$

where $H=(K+\alpha I)^{-1}K$ is the hat matrix, and $diag(\cdot)$ denotes the diagonal elements of a matrix.

Maximizing the log of Marginal Likelihood in GPR: GPR is a kernel-based regression method. For a given training set $\{x_i\}_{i=1}^{n}$, and $y_i = f(x_i)+\epsilon$, a multivariate Gaussian distribution is defined on any function f such that, $P(f(x)) = \mathcal{N}(\mu,\Sigma)$, where Σ is a kernel. Marginal likelihood is used as the model selection criterion in GPR, since it balances between the lack-of-fit and complexity of a model. Maximizing the log of marginal likelihood obtains the optimal parameters for model selection. The log of marginal likelihood is denoted as:

$$\log P(y|x,\theta) = -\tfrac{1}{2}y^T\Sigma^{-1}y - \tfrac{1}{2}\log|\Sigma| - \tfrac{n}{2}\log 2\pi, \tag{31}$$

where $-y^T\Sigma y$ denotes the models fit, $\log|\Sigma|$, denotes the complexity , and $\frac{n}{2}\log 2\pi$ is a normalization constant. Without loss of generality in this paper the abbreviation GPR means the model selection criterion used in GPR.

ICOMP: The kernel-based ICOMP introduced in [18] is an information criterion to select the models and is defined as $\text{ICOMP} = -2\text{ PLL} + 2\,C(\Sigma)$, where $C(\Sigma)$, and Σ elaborated in equations 11, and 12.

5. Experiments and Results

In this section we evaluate the performance of KIC on synthetic, and real datasets, and compare with competing model selection methods.

5.1 Artificial Dataset

KIC was first evaluated on the problem of approximating $f(x) = sinc(x) = sin(\pi x)/\pi x$ from a set of 100 points sampled at regular intervals in [-6; 6]. To evaluate robustness to noise, normal random noise was added to the *sinc* function at two Noise-to-Signal (NSR) ratios: 5%, and 40%. Figure 1 shows the sinc function and the perturbed datasets. The following experiments were conducted: (1) shows how KIC balances between GoF and complexity, (2) shows how KIC and MSE on training sets change when the sample size and the level of noise in the data change (3) investigates the effect of using different kernels, and (4) evaluates the consistency of KIC in parameter selection. All experiments were run 100 times using randomly generated datasets, and corresponding test sets of size 1000.

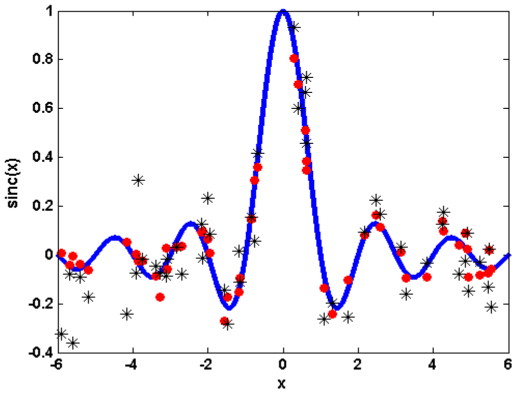


Figure 1. The simulated sinc function is depicted with solid line, and the noisy sinc function on 50 randomly selected data points with NSR=5%, and 40% are shown respectively with red dots, and black stars

Experiment 1. The effect of α on complexity, lack-of-fit and KIC values was measured by setting $\alpha = \{0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$, with KRR models being generated using a Gaussian kernel with different standard deviations, $\sigma = \{0.3, 1, 4\}$, computed over the 100 data points. The results are shown in Figure 2. The model generated with $\sigma = 0.3$ overfits, because it is overly complex, while $\sigma = 4$ gives a simpler model that underfits. As the ridge parameter α increases, the model complexity decreases while the goodness-of-fit is adversely affected. KIC balances between these two terms, which yields a criterion to select a model that has good generalization, as well as goodness of fit to the data.

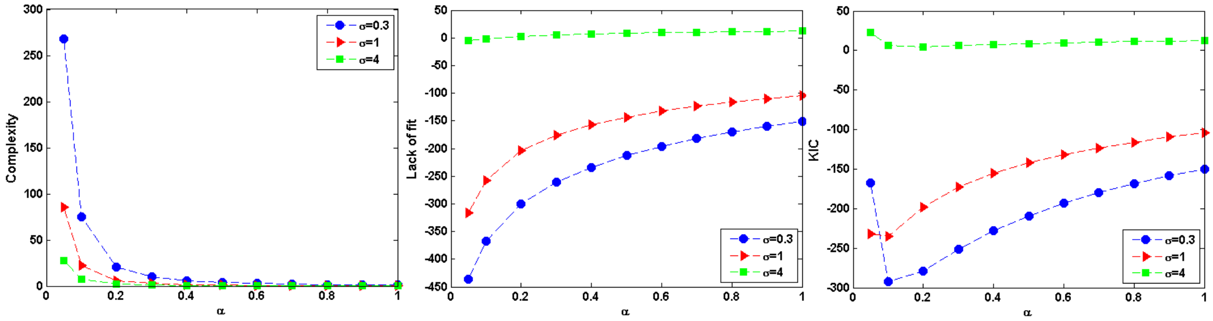
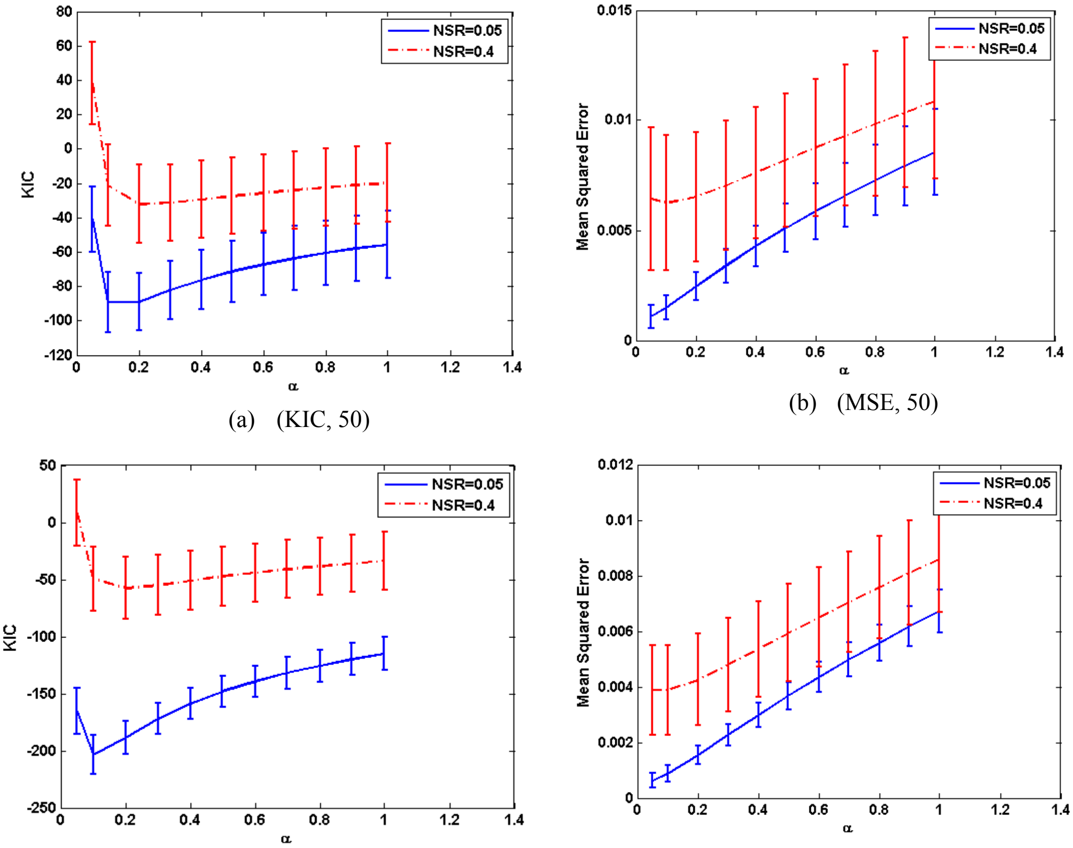


Figure 2. The complexity (left), goodness-of-fit (middle), and KIC values (right) of KRR models with respect to changes in α for different values of $\sigma = \{0.3,1,4\}$ in Gaussian kernels. KIC balances between the complexity and the lack of fit

Experiment 2. The influence of training sample size was investigated by comparing sample sizes, n, of 50, and 100, for a total of four sets of experiments: (n, NSR): (50, 5%), (50, 40%), (100, 5%), (100, 40%). The Gaussian kernel was used with $\sigma = 1$. The KIC value and Mean Squared Error $\left(\text{MSE}, \|\hat{f} - f\|^2/n\right)$, for different α ={0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0} is shown in Figure 3. The data with NSR=40% has larger MSE values, and larger error bars, and consequently larger KIC values compared to data with NSR=5%. In both cases, KIC and MSE change with similar profiles with respect to α. The sample size and the noise had no effect on KIC for selecting the best model (parameter α).

Experiment 3. The effect of using a Gaussian kernel, $k(x, y) = \exp\left(-\frac{\|x-y\|^2}{2\,\sigma^2}\right)$, versus the Cauchy kernel, $k(x, y) = 1/\left(1 + (\|x^\eta - y^\eta\|^2/\eta)\right)$, was investigated, where σ = 1, and $\eta = 2$ in the computation of the kernel-based model selection criteria ICOMP, KIC, GPR, and LOOCV. The results are reported in Figures 4 and 5. The graphs show box plots with markers at 5%, 25%, 50%, and 95% of the empirical distributions of MSE values. As expected, the MSE of all methods is larger when NSR is high, 0.4, and smaller for the larger of the two training sets (100 samples). LOOCV, ICOMP, and KIC performed comparably, and better than GPR using a Gaussian kernel for data with NSR = 0.05. In the other cases, the best results (smallest MSE) were achieved by KIC. All methods have smaller MSE values using the Gaussian kernel versus the Cauchy kernel. GPR with the Cauchy kernel obtains results comparable with KIC, but with a standard deviation close to zero.



(a) (KIC, 50)

(b) (MSE, 50)

(c) (KIC, 100)

(d) (MSE, 100)

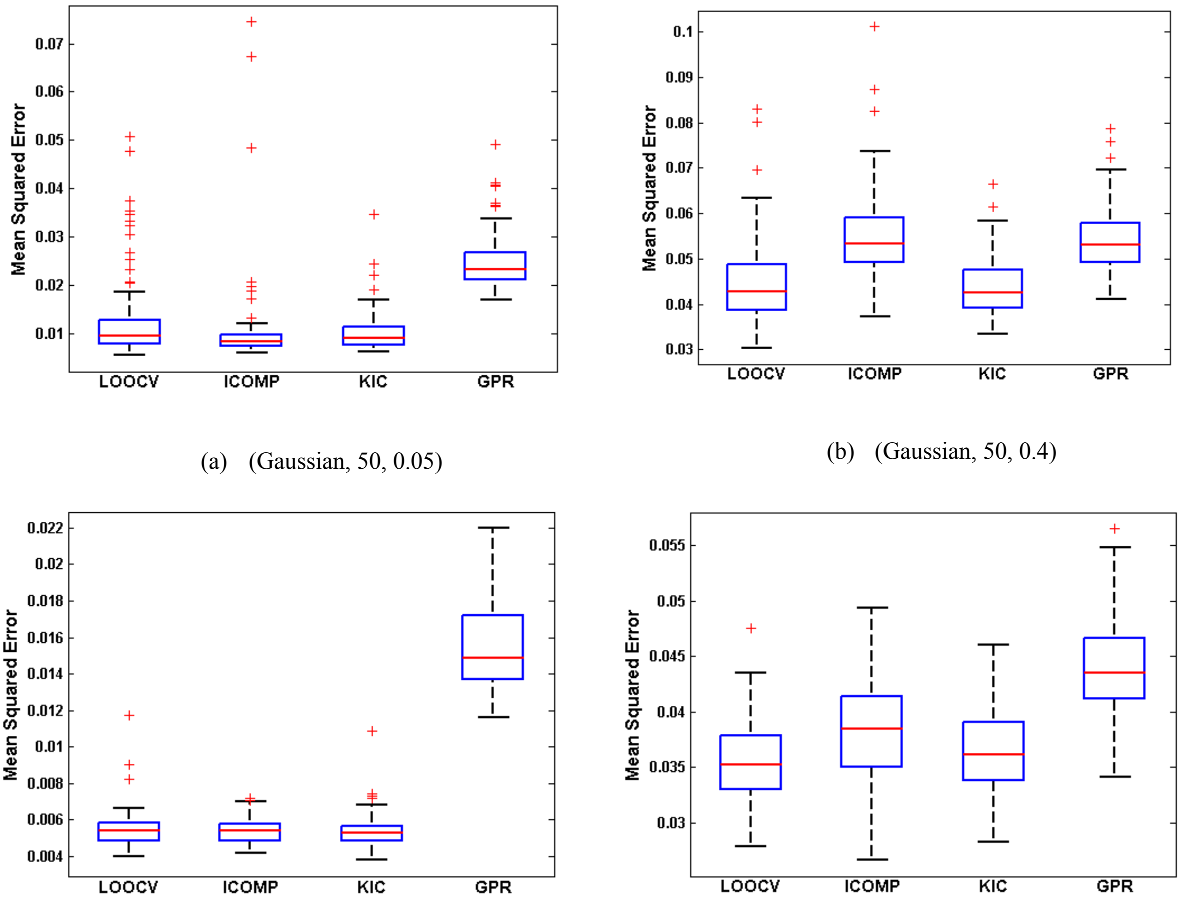
Figure 3. (a), and (b) represent changes of KIC, and MSE values with respect to α for 50 data points respectively, and (c) and (d) are corresponding diagrams for the data with 100 sample size. The solid lines represent NSR=0.05, and dashed lines NSR=0.4

Experiment 4. We assessed the consistency of selecting/tuning the parameters of the models in comparison with LOOCV. We considered four experiment of sample size, n = (50, 100), and NSR = (0.05, 0.4). The parameters to tune or select are $\alpha = \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1\}$, and $\sigma = \{0.4, 0.6, 0.8, 1, 1.2, 1.5\}$ for the Gaussian kernel. The frequency of selecting the parameters are shown in Figure 6 for LOOCV, and in Figure 7 for KIC. The more concentrated frequency shows the more consistent selecting criterion. The diagrams show that KIC is more consistent in selecting the parameters rather than LOOCV. LOOCV is also sensitive to sample size. It provides a more consistent result for benchmarks with 100 samples.

5.2 Abalon, Kin, and Puma Datasets

We used three benchmarks selected from the Delve datasets (www.cs.toronto.edu/~delve/data): (1) Abalone dataset (4177 instances, 7 dimensions), (2) Kin-family of datasets (4 datasets; 8192 instances, 8 dimensions), and (3) Puma-family of datasets (4 datasets; 8192 instances, 8 dimensions).

For the Abalone dataset, the task is to estimate the age of abalones. We used normalized attributes in range [0,1].



(a) (Gaussian, 50, 0.05)

(b) (Gaussian, 50, 0.4)

(c) (Gaussian, 100, 0.05)

(d) (Gaussian, 100, 0.4)

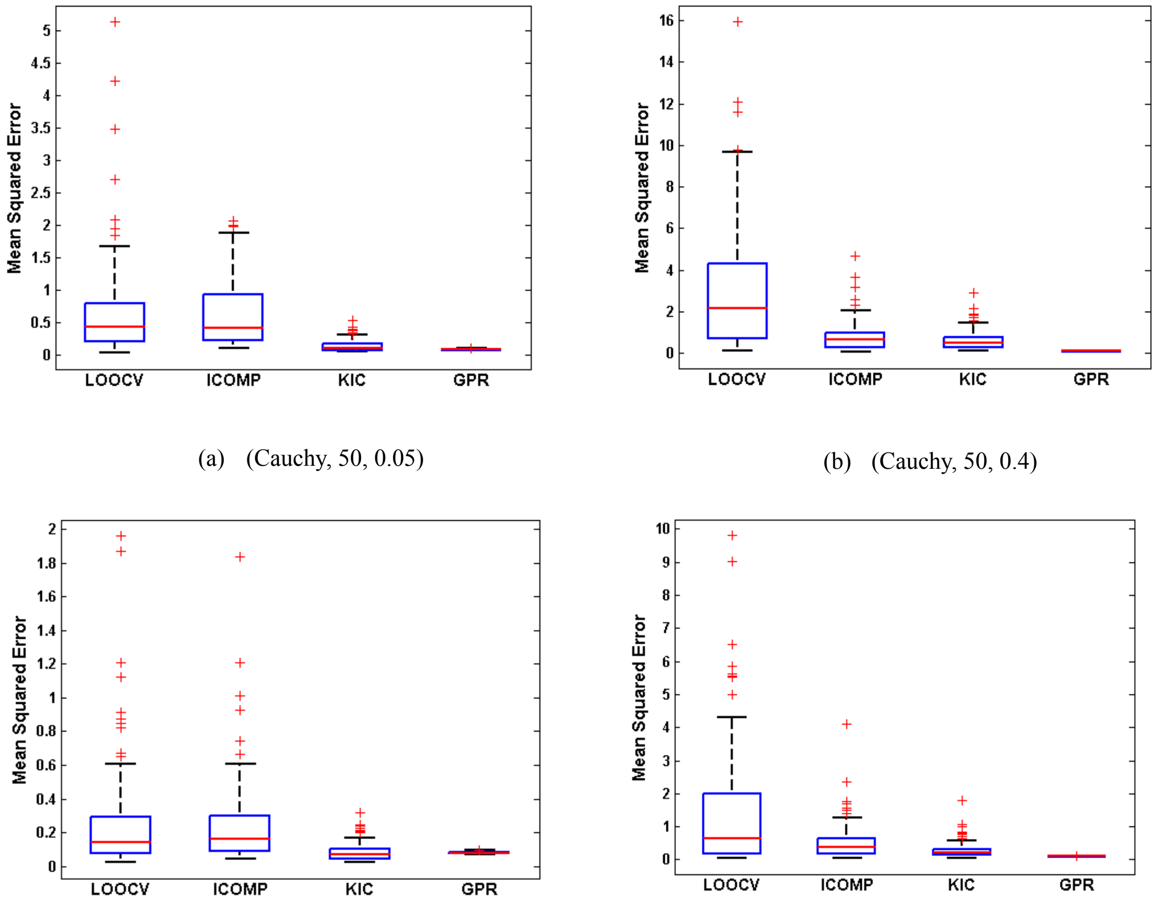
Figure 4. The graphs depict the Mean Squared Error values as results of using Gaussian kernel with σ = 1, in ICOMP, KIC, and GPR, compared with MSE of LOOCV. The results on simulated data with (n, NSR): (50, 0.05), (50, 0.4), (100, 0.05), and (100, 0.4) are shown in (a), (b), (c), and (d) respectively. KIC gives the best results

The experiment is repeated 100 times to obtain the confidence interval. In each trial 100 samples were selected randomly as the training set and the remaining 4077 samples as the test set. The Kin-family and Puma-family datasets are realistic simulations of a robot arm taking into consideration combinations of attributes such as whether the arm movement is nonlinear (n) or fairly linear (f), and whether the level of noise (unpredictability) in the data is: medium (m), or high (h). The Kin-family includes: kin-8fm, kin-8fh, kin-8nm, kin-8nh datasets, and the Puma-family contains: puma-8fm, puma-8fh, puma-8nm, and puma-8nh datasets.

In the Kin-family of datasets, having the angular positions of an 8-link robot arm, the distance of the end effector of the robot arm from a starting position is predicted. The angular position of a link of the robot arm is predicted given the angular positions, angular velocities, and the torques of the links.

We compared KIC1 (26), KIC2 (27), and KIC with LOOCV, ICOMP, and GPR on the three datasets. The results are shown as box-plots in Figures 8, 9, and 10 for Abalone, Kin-family, and Puma-family datasets, respectively. The best results across all three datasets were achieved using KIC, and the second best results were for LOOCV.

For the Abalone dataset, comparable results were achieved for KIC and LOOCV, that are better than ICOMP, and the smallest MSE value obtained by GPR. KIC1, and KIC2 had similar MSE values, which are larger than for the other methods.



(a) (Cauchy, 50, 0.05)

(b) (Cauchy, 50, 0.4)

(c) (Cauchy, 100, 0.05)

(d) (Cauchy, 100, 0.4)

Figure 5. The graphs show the MSE values resulting from using a Cauchy kernel with $\eta = 2,$ in ICOMP, KIC, and GPR, compared with the MSE of LOOCV. The results on simulated data with (n, NSR): (50, 0.05), (50, 0.4), (100, 0.05), and (100, 0.4) are shown in (a), (b), (c), and (d) respectively. KIC and GPR give the best result

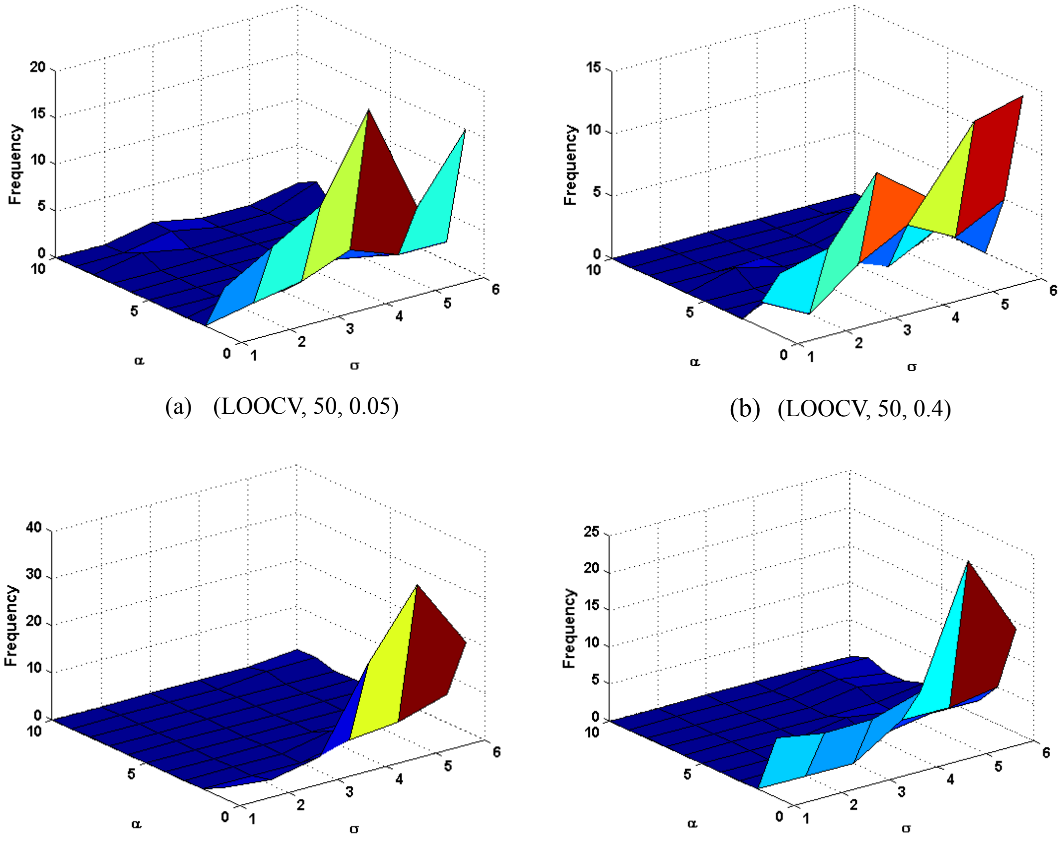
For the Kin-family datasets, except for kin-8fm, KIC gets better results than GPR, ICOMP, and LOOCV. KIC1, and KIC2 obtain better results than GPR, and LOOCV for kin-8fm, and kin-8nm, which are datasets with medium level of noise, but larger MSE value for datasets with high noise (kin-8fh, and kin-8nh).

For the Puma-family datasets, KIC got the best results on all datasets except for on puma-8nm, where the smallest MSE was achieved by LOOCV. The result of KIC is comparable to ICOMP and better than GPR for puma-8nm dataset. For puma-8fm, puma-8fh, and puma-8nh, although the median of MSE for LOOCV and GPR are comparable to KIC, KIC has a more significant MSE (smaller interquartile in the box bots). The median MSE value for KIC1, and KIC2 are closer to the median MSE values of the other methods on puma-8fm, and puma-8nm, where the noise level is moderate compared to puma-8fh, and puma-8nh, where the noise level is high.

The sensitivity of KIC1, and KIC2 to noise is due to the existence of variance in their formula. KIC2 has a larger interquartile of MSE than KIC1 in datasets with high noise, which highlights the effect of 4 in its formula (equation 27) rather than 2 in equation (26).

6. Conclusion

We introduced a kernel-based information criterion (KIC) for model selection in regression analysis. The complexity measure in KIC is defined on a novel variable-wise variance which explicitly computes the interdependency of each parameter involved in the model; whereas in methods such as kernel-based ICOMP and GPR, this interdependency is defined on a covariance matrix, which obscures the true contribution of the model parameters. We provided empirical evidence showing how KIC outperforms LOOCV (with kernel-based closed form formula of the estimator), Kernel-based ICOMP, and GPR, on both artificial data and real benchmark datasets: Abalon, Kin-family, and Puma-family. In these experiments, KIC efficiently balances the goodness of fit and complexity of the model, is robust to noise (although for higher noise we have larger confidence interval as expected) and sample size, is consistent in tuning/selecting the ridge and kernel parameters, and has significantly smaller or comparable Mean Squared Error values with respect to competing methods, while yielding stronger regressors. The effect of using different kernels was also investigated since the definition of a proper kernel plays an important role in kernel methods. KIC had superior performance using different kernels and for the proper one obtains smaller MSE.



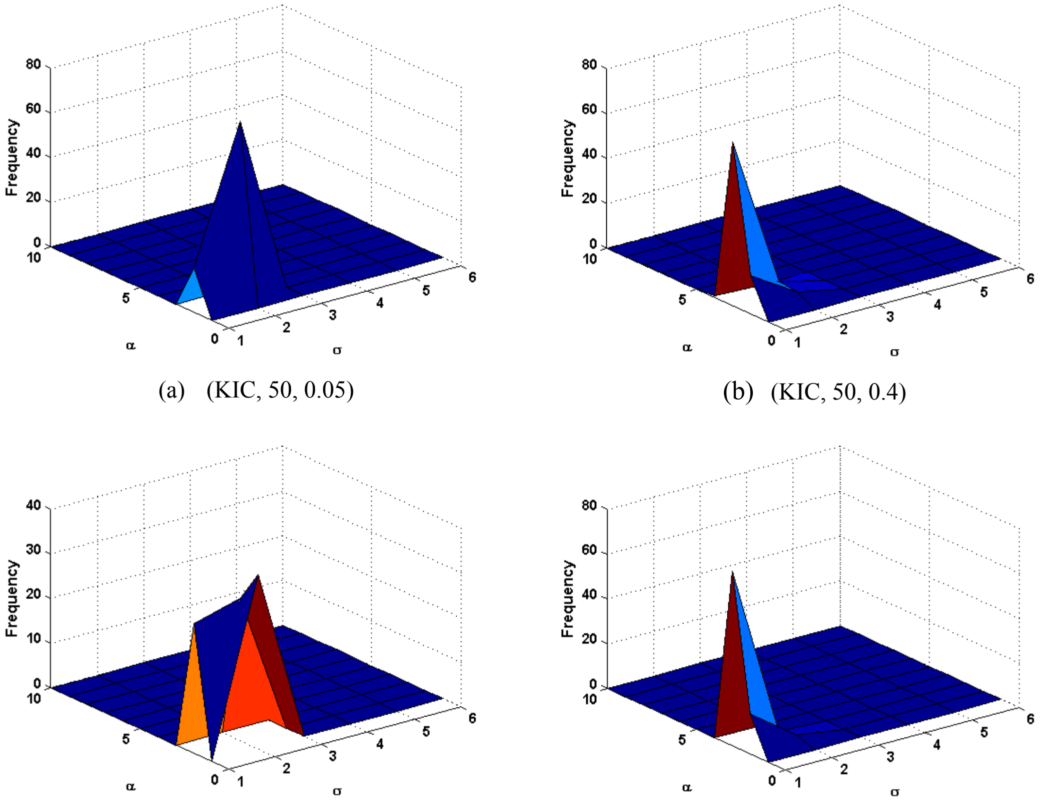
(a) (LOOCV, 50, 0.05)

(b) (LOOCV, 50, 0.4)

(c) (LOOCV, 100, 0.05)

(d) (LOOCV, 100, 0.4)

Figure 6. The frequency of selecting parameters α, σ, and by LOOCV method via running 100 trials on artificial benchmarks with (n, NSR) = (50, 0.05), (50, 0.4), (100, 0.05), (100, 0.4) are shown in (a), (b), (c), and (d) respectively



(a) (KIC, 50, 0.05)

(b) (KIC, 50, 0.4)

(c) (KIC, 100, 0.05)

(d) (KIC, 100, 0.4)

Figure 7. The frequency of selecting parameters α, and σ by KIC via running 100 trials on artificial benchmarks with (n, NSR) = (50, 0.05), (50, 0.4), (100, 0.05), (100, 0.4) are shown in (a), (b), (c), and (d) respectively

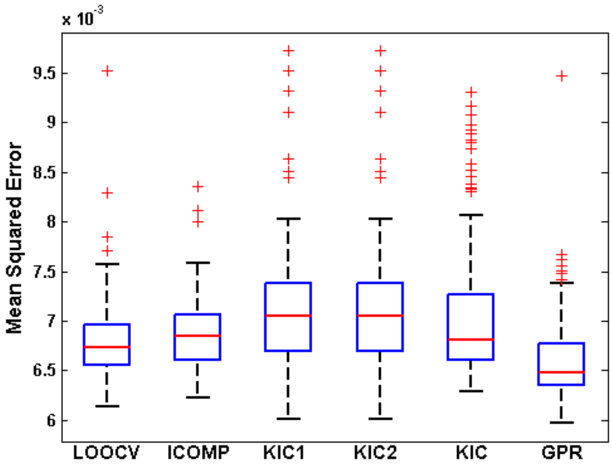
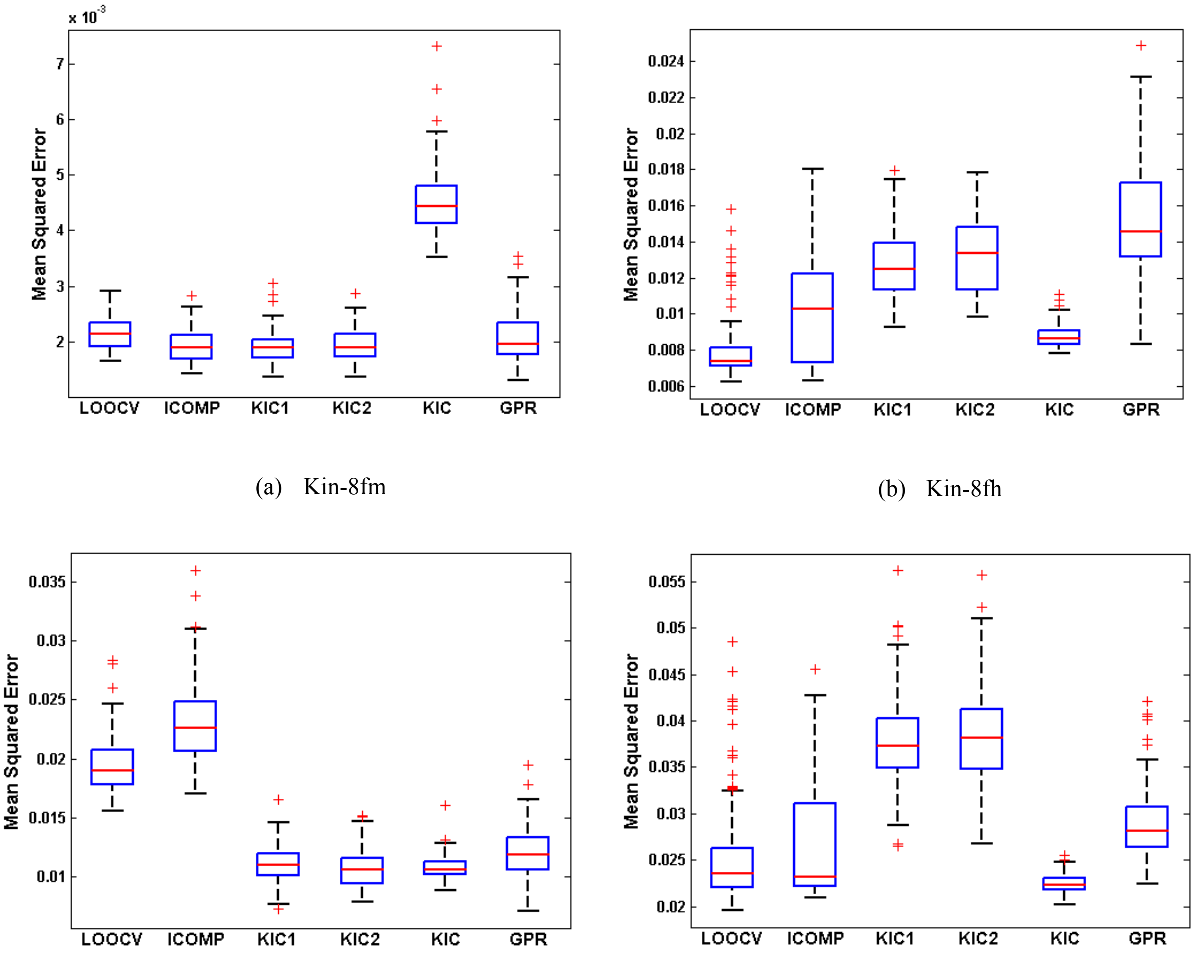


Figure 8. The results of LOOCV, ICOMP, KIC1, KIC2, KIC, and GPR on Abalone dataset. Comparable results achieved by LOOCV, and KIC, with smaller MSE value rather than ICOMP, and the best results by GPR



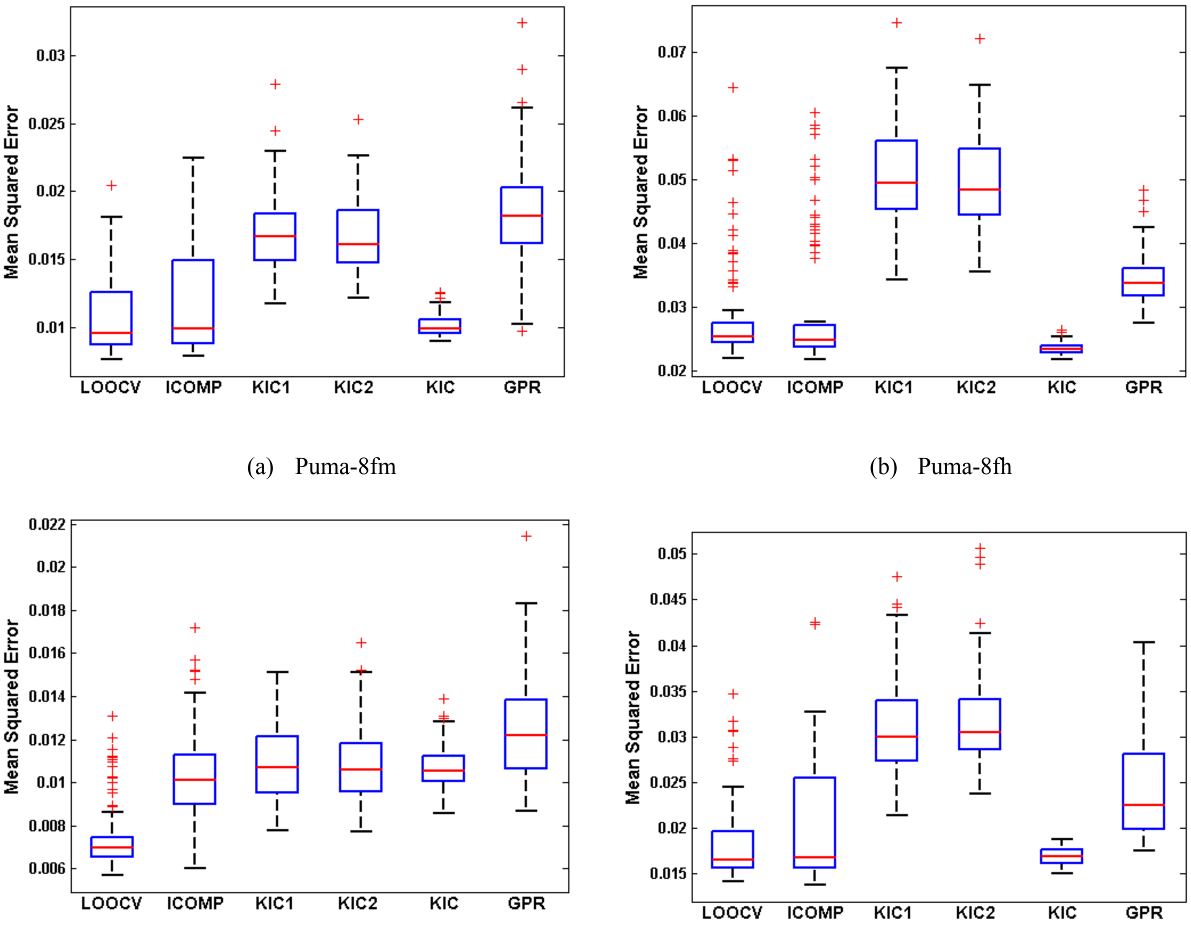
(a) Kin-8fm

(b) Kin-8fh

(c) Kin-8nm

(d) Kin-8nh

Figure 9. The results of LOOCV, ICOMP, KIC1, KIC2, KIC, and GPR on kin-family of datasets are shown using box plots. (a), (b), (c), and (d) are the results on kin-8fm, kin-8fh, kin-8nm, kin-8nh, respectively



(a) Puma-8fm

(b) Puma-8fh

(c) Puma-8nm

(d) Puma-8nh

Figure 10. The results of LOOCV, ICOMP, KIC1, KIC2, KIC, and GPR on puma-family of datasets are shown using box plots. (a), (b), (c), and (d) are the results on puma-8fm, puma-8fh, puma-8nm, puma-8nh, respectively

Acknowledgments

We want to thank Arthur Gretton, and Zoltán Szabó for the fruitful discussions. Kenji Fukumizu was supported by MEXT Grant-in-Aid for Scientific Research on Innovative Areas 25120012.

References

Akaike, H. (1973). *Information theory and an extension of the maximum likelihood principle.* In Petrov, B. N., Csaki, B. F., (Ed.), *Second Int. Symposium on Information Theory* (pp. 267-281). http://dx.doi.org/10.1007/978-1-4612-1694-0_15

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys, 4*, 40-79. http://dx.doi.org/10.1214/09-ss054

Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning*. Springer New York.

Bozdogan, H., & Haughton, D. (1998). Information complexity criteria for regression models. *Computational statistics and data analysis, 28*(1), 51-76. http://dx.doi.org/10.1016/s0167-9473(98)00025-5

Demyanov, S., Bailey, J., Ramamohanarao, K., & Leckie, C. (2012). AIC and BIC based approaches for SVM parameter value estimation with RBF kernels. In ACML, JMLR Proceedings (vol. 25, pp. 97-112). Retrieved from http://jmlr.org/proceedings/papers/v25/demyanov12/demyanov12.pdf

Gretton, A., Bousquet, O., Smola, A., & Schölkopf, B. (2005). *Measuring statistical dependence with Hilbert-Schmidt norms*, In: Proceedings algorithmic learning theory (pp. 63-77). Springer-Verlag.

http://dx.doi.org/10.1007/11564089_7

Kobayashi, K., & Komaki, F. (2005). *Information Criteria for Kernel Machines*. Technical report, Uni. of Tokyo. Retrieved from http://www.keisu.t.u-tokyo.ac.jp/research/techrep/data/2005/METR05-23.pdf

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *IJCAI, 14*, 1137-1145. Retrieved from http://www.cs.iastate.edu/~jtian/cs573/Papers/Kohavi-IJCAI-95.pdf

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1993). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics, 21*(6). http://dx.doi.org/10.1063/1.1699114

Nishii, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *Annals of Statistics, 12*, 758-765. http://dx.doi.org/10.1214/aos/1176346522

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian processes for machine learning*. The MIT Press.

Rosipal, R., Girolami, M., & Treja, L., J. (2001). *On Kernel Principal Component Regression with covariance inflation criterion for model selection*. Technical report, Uni. of Paisley. Retrieved from http://aiolos.um.savba.sk/~roman/Papers/tr01_2.pdf

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461-464. http://dx.doi.org/10.1214/aos/1176344136

Shannon, C. E. (2001). A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review, 5*(1), 3-55. http://dx.doi.org/10.1145/584091.584093

Van Emden, M. (1975). An analysis of complexity. *Mathematical center Tracts, 35*, 1-86. Retrieved from http://books.google.ch/books/about/An_Analysis_of_Complexity.html?id=vZCgPAAACAAJ&redir_esc=y

Vapnik, V., Chapelle, O. (2000). Bounds on error expectation for support vector machines. *Neural Computation, 12*(9), 2013-2036. http://dx.doi.org/10.1162/089976600300015042

Wahba, G. (1990). Spline models for observational data. Siam, 59. http://dx.doi.org/10.1137/1.9781611970128

Zhang, R. (2007). *Model selection techniques for kernel-based regression analysis using information complexity measure and genetic algorithms*. PhD Dissertation, Uni. of Tennessee. Retrieved from http://etd.utk.edu/2007/ZhangRui.pdf

Notes

Note 1. The reader is referred to Nishii [10] for a detailed discussion on the asymptotic properties of a wide range of information criteria for linear regression.

Note 2. Under the normality assumption for random errors, we can use the least-squares instead of the log-likelihood in the formula. Therefore KIC can also be written as $\mathrm{KIC}(\hat{\Sigma}_{\hat{\theta}}) = 2\ \mathrm{Penalized\ Least - Squares}(\hat{\theta}) + \mathrm{C}(\hat{\Sigma}_{\hat{\theta}})$.

Copyrights