Model Selection in Regression Problems Based on the Modulus of Continuity

(Revision of Research Report 04-23)

Imhoi Koo and Rhee Man Kil
Division of Applied Mathematics,
Korea Advanced Institute of Science and Technology
373-1 Guseong-dong, Yuseong-gu, Daejeon 305-701, Korea
haihaba@amath.kaist.ac.kr; rmkil@kaist.ac.kr

Key-words: Model selection, Regression, Modulus of continuity, Expected Risk

Abstract

This paper presents a new method of model selection in regression problems based on the modulus of continuity. For this purpose, the bounds on the expected risks are suggested using the modulus of continuity for the target and estimation functions. As a result, the suggested bounds are described by learned parameters of regression models and the model selection criteria referred to as the modulus of continuity information criteria (MCIC) are suggested for the selection of optimal structure of regression models. Through the simulation for function approximation, we have shown that the suggested model selection can provide a better performance than other statistical methods of model selection such as AIC or BIC from the view points of the risks and the number of parameters.

1 Introduction

Model selection is an important issue of selecting a reasonable network size to get the optimal performance of regression models. The regression of real-valued functions is performed for a given finite number of samples. In this regression, the proper network size (or number of parameters) in a regression model is hard to decide since the performance of a regression model measured for the entire distribution of samples (not just given samples) should be optimized. For the performance of regression models, the loss function of error square is usually measured and the expectation of the loss function for the entire sample distribution is considered. This expected risk can be decomposed by the bias and variance terms of regression models. If we increase the number of parameters, the bias term is decreased but the variance term is increased or vice versa. If the number of parameters is so small that the performance is not optimal due to large bias term, it is called the under-fitting of regression models. If the number of parameters is too large so that the performance is not optimal due to large variance term, it is called the over-fitting of regression models. So we need the trade-off between the under-fitting and over-fitting of regression models. Here, the important issue is measuring the model complexity related to the variance term. The statistical methods of model selection are to use a penalty (correction) term as the measure of model complexity. The well known criteria using this penalty term are Akaike information criteria (AIC)[1], Bayesian information criteria (BIC)[2], generalized cross-validation (GCV)[3], minimum description length (MDL)[4, 5], risk inflation criteria (RIC)[6], etc. These methods can be easily fit to linear regression models when a large number of samples is available. However, they suffer the difficulty to select the optimal structure of the estimation networks in the case of nonlinear regression models and/or the small number of samples. Recently, Vapnik[7] proposed a model selection method based on the structural risk minimization (SRM) principle. One of characteristics of this method is that the model complexity is described by the structural information such as the VC dimension of the estimation network. This method can be applied to nonlinear models and also regression models trained for a small number of samples. Chapelle et al.[8] and Cherkasky et al.[9] showed that the SRM based model selection is able to outperform other statistical methods such as AIC or BIC. However, these methods require the actual VC dimension of the hypothesis space associated with the estimation network, which is not easy to determine in the case of nonlinear regression models. From this point of view, we present the bounds on the expected risks in the sense of the modulus of continuity (MC) representing a measure of continuity for the given function. Lorentz[10] applied the MC to function approximation theories. In our method, this measure is applied to determine the bounds on expected risks. For the estimation of these bounds, the MC is analyzed for both the target and estimation functions. As a result, the suggested bounds are described by learned parameters of regression models and the model selection criteria referred to as the modulus of continuity information criteria (MCIC) are suggested for the selection of optimal structure of regression models. One of characteristics in the suggested MCIC is that it can be estimated directly from the given samples. Through the simulation for function approximation, we have shown that the suggested method of model selection can provide the better performance than other statistical methods of model selection such as AIC or BIC from the view points of the risks and the number of parameters.

In section 2, we introduce the various model selection criteria based on statistics and VC dimension based approaches. Section 3 describes the model selection criteria based on the modulus of continuity. Section 4 describes how we can estimate the modulus of continuity for various estimation functions. In section 5, we compare the optimality of estimation models in the L_1 and L_2 sense. Section 6 describes the simulation results for function approximation based on various model selection criteria including the suggested method. Finally, section 7 presents the conclusion.

2 Model Selection Criteria for Regression Models

Consider the regression problem of estimating a function f in $C^1(X,\mathbb{R})$ or $C^2(X,\mathbb{R})$, where X is the compact subset of Euclidean space $\mathbb{R}^d$ ($d \geqslant 1$) and $C^k(X,\mathbb{R})$ is a class of functions having continuous kth derivative on X. The observed output y for $x \in X$ can be represented by

$$y(x) = f(x) + \epsilon \qquad (1)$$

where $f(x)$ is the target function and ϵ is a random noise with mean zero and variance σ^2. Here, the input and output samples (x_i, y_i), $i = 1, \cdots, N$ are randomly generated according to the distribution $P(x)$, $x \in X$:

$$y_i = f(x_i) + \epsilon_i, \quad x_i \in X \tag{2}$$

where ϵ_i, $i = 1, \cdots, N$ are independent and identically distributed (*i. i. d.*) random variables with zero mean and variance σ. For these samples, our goal is to construct an estimation network (or function) $f_n(x)$ which minimizes the expected risk (or true risk)

$$R(f_n) = \int_{X \times \mathbb{R}} L(y, f_n(x)) dP(x, y) \tag{3}$$

with respect to the number of parameters n, where $L(y, f_n(x))$ is a given loss functional, usually the square loss function $L(y, f_n(x)) = (y - f_n(x))^2$ for regression problems. In general, we can construct an estimation function as

$$f_n(x) = \sum_{k=1}^{n} w_k \phi_k(x) \tag{4}$$

where w_k and ϕ_k represent the kth weight value and kernel function respectively.

To minimize the expected risk (3), we have to identify the distribution $P(x, y)$ but it is usually unknown. Rather, the parameters to minimize the empirical risk

$$R_{emp}(f_n) = \frac{1}{N} \sum_{i=1}^{N} L(y_i, f_n(x_i)) \tag{5}$$

are obtained for the given samples. If we increase the number of parameters, the empirical risk of (5) is decreased but the variance term related to the complexity of regression models is increased or vice versa. So we have to make the reasonable trade-off between the under-fitting and over-fitting of regression models. This problem is referred to as the model selection, that is, we have to determine the optimal number of parameters (or complexity) while avoiding the under-fitting or over-fitting of regression models. To check whether under-fitting or over-fitting of regression models happens in the course of learning, we have to analyze the expected risks for the given parameters in regression models. Here, let us represent the estimate of the expected risk as

$$\hat{R}(f_n) = R_{emp}(f_n) T(n, l) \tag{6}$$

where $T(n, l)$ is a penalty (correction) term, n is the number of parameters (or model complexity), and l is the number of sample.

The well known statistical model selection criteria such as AIC and BIC can be described using the form of (6), they are

$$T_{AIC}(n, l) = \frac{1 + \frac{n}{l}}{1 - \frac{n}{l}} \quad \text{and} \tag{7}$$

$$T_{BIC}(n, l) = 1 + \frac{\ln l}{2} \frac{\frac{n}{l}}{\left(1 - \frac{n}{l}\right)} \tag{8}$$

where T_{AIC} and T_{BIC} represent the penalty factors of the AIC and BIC criteria respectively. These model selection criteria come from the asymptotic analysis for linear models in L_2 sense.

A good measure of model complexity in nonlinear models is the VC dimension of the hypothesis space associated with the estimation network. The VC dimension can represent the capacity of the estimation network from the view point of the number of samples. As the hypothesis space is increased, the empirical risk can be decreased but the confidence interval associated with the complexity of the estimation network, is increased. In this sense, we need to make the proper trade-off between these two terms. The structural risk minimization (SRM) principle considers both the empirical risk and the complexity of the regression model to decide the optimal structure of the regression model. In this approach, the estimate of expected risks for the VC dimension d associated with the hypothesis space defined by f_n, holds the following inequality with a probability of at least $1-\delta$[7, 9]:

$$\hat{R}(f_n) \leqslant R_{emp}(f_n)\left(1 - c\sqrt{\frac{d\left(1+\ln\frac{l}{d}\right) - \ln\delta}{l}}\right)_+^{-1} \tag{9}$$

where c is a constant dependent on the norm and tails of the loss function distribution and $u_+ = \max\{u, 0\}$.

In the case of nonlinear estimation network, there is some difficulty in determining the VC dimension. If the class of basis function $\{\phi_k(x)\}$, $k = 1, \cdots, n$ is orthogonal with respect to the probability measure $P(x)$, the form of (9) can be described in a way that is easier form to calculate. For instance, if $\phi_k(x)$ represents the kth orthogonal polynomial and a continuous target function $f(x)$ in some Hilbert space with respect to the basis functions $\{\phi_1(x), \cdots, \phi_n(x)\}$ is given to be approximated, the estimated expected risk[8] can be determined by

$$E[R(f_n)] = E[R_{emp}(f_n)]\left(1 + \frac{E[\sum_{i=1}^{n} 1/\lambda_i]}{l}\right)\left(1 - \frac{d}{l}\right)^{-1} \tag{10}$$

where $\{\lambda_1, \cdots, \lambda_n\}$ are the eigenvalues of the $n \times n$ covariant matrix C in which the pqth entry is given by $\frac{1}{l}\sum_{i=1}^{l}\phi_p(x_i)\phi_q(x_i)$.

Furthermore, for the experimental set up, Chapelle et al.[8] suggest the following bound with the confidence parameter $\delta = 0.1$:

$$E[R(f_n)] \leqslant E[R_{emp}(f_n)]T_{SEB}(n, l) \tag{11}$$

where

$$T_{SEB}(n, l) = \frac{1 + \frac{n}{lk}}{1 - \frac{n}{l}} \quad \text{and} \tag{12}$$

$$k = \left(1 - \sqrt{\frac{n\left(1 + \ln\frac{2l}{n}\right) + 4}{l}}\right)_+ . \tag{13}$$

The risk function of (11) was applied to the model selection of regression problems and they showed that their method outperformed the model selection using the statistical criteria such as AIC or BIC.

3 Model Selection Criteria Based on the Modulus of Continuity

The modulus of continuity is a measure of continuity for a given function. First, we will suggest the bounds on expected risks based on this measure. Here, we assume that X is a compact subset of Euclidean space $\mathbb{R}$. Then, the measure of continuity $w(f,h)$ of a function $f \in C(X)$ can be described by the following form:

$$\omega(f,h) = \max_{x,x+t \in X, |t| \leqslant h} |f(x+t) - f(x)| \tag{14}$$

where h is a positive constant. This modulus of continuity of f has the following properties:

- $\omega(f,h)$ is continuous at $h \geqslant 0$ for each f,
- $\omega(f,h)$ is positive and increasing for $h > 0$, and
- $\omega(f,h)$ is sub-additive, that is, $\omega(f,h_1+h_2) \leqslant \omega(f,h_1) + \omega(f,h_2)$ for each f.

As a function of f, the modulus of continuity has the properties of a semi-norm:

$$\omega(af,h) \leqslant |a|\omega(f,h) \quad \text{and}$$

$$\omega(f_1+f_2,h) \leqslant \omega(f_1,h) + \omega(f_2,h).$$

The famous example of modulus of continuity of a function f is that f is defined on $A = [a,b]$ and satisfies a Lipschitz condition with constant $M \geqslant 0$ and exponent α, denoted by $\text{Lip}_M \alpha, 0 < \alpha \leqslant 1$, that is,

$$|f(x) - f(y)| \leqslant M|x-y|^\alpha, \quad x,y \in A. \tag{15}$$

In this case, the modulus of continuity is given by

$$\omega(f,h) \leqslant Mh^\alpha. \tag{16}$$

One of applications of the modulus of continuity is the analysis of a degree of approximation. For example, the n-th degree of approximation of a function $f \in C^*[-\pi,\pi]$, a set of all periodic and continuous functions on $[-\pi,\pi]$, by trigonometric polynomials T_n of degree n is defined by

$$E_n^*(f) = \min_{T_n} \|f - T_n\|_\infty. \tag{17}$$

For the above degree of approximation, Jackson[11, 12] suggested the following theorem:
Theorem (degree of approximation for trigonometric polynomials)
There is a constant $M \geqslant 0$ such that the following inequality

$$E_n^*(f) \leqslant M\omega\left(f, \frac{1}{n}\right), n = 1, 2, \cdots$$

holds true for every function $f \in C^*[-\pi,\pi]$.

For some algebraic polynomials $P_n(x)$, a similar result can be obtained for some f satisfying $\omega(f,h)$, that is,

$$|f(x) - P_n(x)| \leqslant M\omega\left(f, \frac{1}{n}\right), \quad -1 \leqslant x \leqslant 1 \tag{18}$$

from Jackson's theorem.

With these properties of the modulus of continuity, we will describe the relationship between the expected (or true) and empirical risks using the modulus of continuity. Here, let us consider the loss function for the observed model y and the estimation function $f_n(x)$ with n parameters as $L(y, f_n) = |y - f_n(x)|$, that is, the expected and the empirical risks are defined by the following L_1 measure:

$$R(f_n)_{L_1} = \int_{X\times\mathbb{R}} |y - f_n(x)| dP(x,y) \quad \text{and} \tag{19}$$

$$R_{emp}(f_n)_{L_1} = \frac{1}{N}\sum_{i=1}^{N} |y_i - f_n(x_i)|. \tag{20}$$

For these risks, we suggest the following theorem based on the modulus of continuity:

Theorem 1 *Let us consider that the target function f of (1) is approximated by the estimation function f_n of (4), that is, a linear combination of weight parameters w_k, $k = 1,\cdots,n$ and basis functions ϕ_k, $k = 1,\cdots,n$ for the given samples (x_i, y_i), $i = 1,\cdots,N$ generated by (2). Then, the expected risk in the L_1 sense is bounded by the following inequality with a probability of at least $1-\delta_1-\delta_2$:*

$$R(f_n)_{L_1} \leqslant R_{emp}(f_n)_{L_1} + \frac{1}{N^2}\sum_{i,j=1}^{N}(|y_i - y_j| + |f_n(x_i) - f_n(x_j)|)$$
$$+ (\omega(f_n, h_0) + C)\sqrt{\frac{1}{2N}\ln\frac{2}{\delta_1}} \quad \textit{and} \tag{21}$$

$$C = |f_n(x_0) - f_n(x_0^{'})| + 2\|f\|_\infty + 2\sigma\sqrt{\frac{1}{\delta_2}} \quad \textit{for } x_0, x_0^{'} \in \{x_1,\cdots,x_N\} \tag{22}$$

where $w(f_n, h_0)$ represents the modulus of continuity of the estimation function f_n, h_0 represents a constant independent upon the target and estimation functions.

Proof. Before the description of main proof, let us introduce the Hoeffding inequality[14]:

Given *i. i. d.* random variables $Y_1,\ldots,Y_N$, let us define a new random variable

$$S_N = \frac{1}{N}\sum_{i=1}^{N} Y_i$$

and we assume that there exist real numbers a_i and b_i for $i = 1,\ldots,N$ such that $Pr\{Y_i \in [a_i, b_i]\} = 1$. Then, for any $\epsilon > 0$ we have

$$Pr\{E[S_N] - S_N \geqslant \epsilon\} \leqslant \exp\left(-\frac{2\epsilon^2 N^2}{\sum_{i=1}^{N}(b_i - a_i)^2}\right).$$

First, let us consider the noiseless case, that is, $y = f(x)$ in (1). For the input samples $x_1, \cdots, x_N$, an event A is defined by

$$\frac{1}{N}\sum_{i=1}^{N}\int_X |f_n(x) - f_n(x_i)|dP(x) - \frac{1}{N}\sum_{j=1}^{N}\frac{1}{N}\sum_{i=1}^{N}|f_n(x_j) - f_n(x_i)| \geqslant \epsilon$$

where the first and second terms represent the average over the expectation of $|f_n(x) - f_n(x_i)|$ and the unbiased estimator of the first term respectively.

Then, from the Hoeffding inequality, the probability of an event A is bounded by

$$Pr\{A\} \leqslant \exp\left(\frac{-2\epsilon^2 N}{\left(\max_{x\in X}\frac{1}{N}\sum_{i=1}^{N}|f_n(x) - f_n(x_i)|\right)^2}\right).$$

For the denominator in the argument of the exponent, we can consider the following inequality:

$$\begin{aligned}\max_{x\in X}\frac{1}{N}\sum_{i=1}^{N}|f_n(x) - f_n(x_i)| &\leqslant \frac{1}{N}\sum_{i=1}^{N}\max_{x\in X}|f_n(x) - f_n(x_i)| \\ &\leqslant \max_i\max_{x\in X}|f_n(x) - f_n(x_i)|.\end{aligned}$$

Let $x_i' = \arg\max_{x\in X}|f_n(x) - f_n(x_i)|$, $x_{i'} = \arg\min_j d(x_j, x_i')$, and $h_i = d(x_i', x_{i'})$ where $d(x, y)$ represents the distance measure defined by $d(x, y) = |x - y|$. Then,

$$\begin{aligned}\max_{x\in X}\frac{1}{N}\sum_{i=1}^{N}|f_n(x) - f_n(x_i)| &\leqslant \max_i\left(|f_n(x_i') - f_n(x_{i'})| + |f_n(x_{i'}) - f_n(x_i)|\right) \\ &\leqslant \max_i\left(\omega(f_n, h_i) + |f_n(x_{i'}) - f_n(x_i)|\right) \\ &\leqslant \omega(f_n, h_0) + |f_n(x_0') - f_n(x_0)|\end{aligned}$$

where $h_0 \in \{h_1, \ldots, h_N\}$ and $x_0, x_0' \in \{x_1, \ldots, x_N\}$ satisfy

$$\omega(f_n, h_i) + |f_n(x_i) - f_n(x_j)| \leqslant \omega(f_n, h_0) + |f_n(x_0) - f_n(x_0')| \tag{23}$$

for $i, j = 1, \cdots, N$. For the illustration of this concept, refer to Figure 1.

Thus, the probability of an event A is bounded by

$$Pr\{A\} \leqslant \exp\left(\frac{-2\epsilon^2 N}{(\omega(f_n, h_0) + |f_n(x_0) - f_n(x_0')|)^2}\right).$$

Here, let us set

$$\frac{\delta_1}{2} = \exp\left(\frac{-2\epsilon^2 N}{(\omega(f_n, h_0) + |f_n(x_0) - f_n(x_0')|)^2}\right).$$

Then, with probability at least $1 - \delta_1/2$, we have

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}\int_X |f_n(x) - f_n(x_i)|dP(x) \leqslant{}& \frac{1}{N^2}\sum_{i,j=1}^{N}|f_n(x_i) - f_n(x_j)| \\ &+ \sqrt{\frac{1}{2N}\ln\frac{2}{\delta_1}}\left(\omega(f_n, h_0) + |f_n(x_0) - f_n(x_0')|\right).\end{aligned} \tag{24}$$

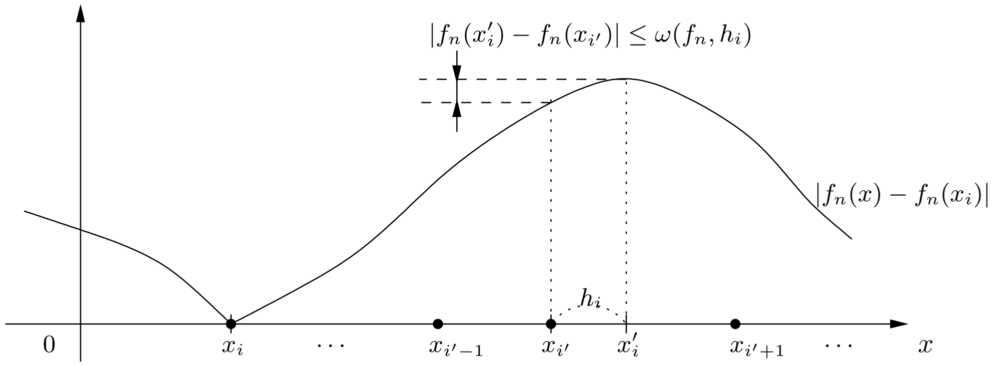


Figure 1: The plot of $|f_n(x) - f_n(x_i)|$ versus x: the value of $|f_n(x) - f_n(x_i)|$ is maximum at x_i' and this maximum value is decomposed by two factors: one is the value of $|f_n(x) - f_n(x_i)|$ at a sample point $x_{i'}$ and another is the modulus of continuity $\omega(f_n, h_i)$ with respect to h_i. The value h_i is chosen by the distance $d(x_i', x_{i'})$.

On the other hand, for the target function f, we can apply a similar method. As a result, with a probability of at least $1 - \delta_1/2$, the following inequality holds:

$$\frac{1}{N}\sum_{i=1}^{N}\int_X |f(x) - f(x_i)|dP(x) \leqslant \frac{1}{N^2}\sum_{i,j=1}^{N}|f(x_i) - f(x_j)| + 2\|f\|_\infty\sqrt{\frac{1}{2N}\ln\frac{2}{\delta_1}}. \quad (25)$$

Let us consider the difference between the expected and empirical errors of $|f(x) - f_n(x)|$:

$$\begin{aligned}\int_X |f(x) - f_n(x)|dP(x) \;&-\; \frac{1}{N}\sum_{i=1}^{N}|f(x_i) - f_n(x_i)| \\ &= \frac{1}{N}\sum_{i=1}^{N}\int_X (|f(x) - f_n(x) - f(x_i) + f_n(x_i) + f(x_i) - f_n(x_i)| \\ &\quad -|f(x_i) - f_n(x_i)|)dP(x) \\ &\leqslant \frac{1}{N}\sum_{i=1}^{N}\int_X |f(x) - f_n(x) - f(x_i) - f_n(x_i)|dP(x) \\ &\leqslant \frac{1}{N}\sum_{i=1}^{N}\int_X (|f(x) - f(x_i)| + |f_n(x) - f_n(x_i)|)dP(x).\end{aligned}$$

Then, from (24) and (25), the difference between the true and empirical risks is bounded by the following inequality with a probability of at least $1-\delta_1$:

$$\begin{aligned}\int_X |f(x)-f_n(x)|dP(x) - \frac{1}{N}\sum_{i=1}^{N}|f(x_i)-f_n(x_i)| \\ \leqslant \frac{1}{N^2}\sum_{i,j=1}^{N}\left(|f(x_i)-f(x_j)|+|f_n(x_i)-f_n(x_j)|\right) \\ +\sqrt{\frac{1}{2N}\ln\frac{2}{\delta_1}}\left(\omega(f_n,h_0)+|f_n(x_0)-f_n(y_0)|+2\|f\|_\infty\right). \quad (26)\end{aligned}$$

Second, let us consider the noisy condition, that is, $y=f(x)+\epsilon$ in (1). Here, we assume that for the output samples $y_1,\cdots,y_N$, the noise terms $\epsilon_i,\ldots,\epsilon_N$ are *i. i. d.* random variables with mean 0 and variance σ. We will define the event B as

$$|\epsilon|\geqslant a$$

where a is a positive constant. Then, from the Chebyshev inequality,

$$Pr\{B\}\leqslant\frac{\sigma^2}{a^2}.$$

Let us set

$$\delta_2=\frac{\sigma^2}{a^2}.$$

Then, with a probability of at least $1-\delta_2$,

$$|\epsilon|\leqslant\sigma\sqrt{\frac{1}{\delta_2}}.$$

This implies that with a probability of at least $1-\delta_2$,

$$|y|\leqslant|f(x)|+|\epsilon|\leqslant\|f\|_\infty+\sigma\sqrt{\frac{1}{\delta_2}}.$$

Here, let us define the event C as

$$\frac{1}{N}\sum_{i=1}^{N}\int_{\mathbb{R}}|y-y_i|dP(y)-\frac{1}{N}\sum_{i=1}^{N}\frac{1}{N}\sum_{j=1}^{N}|y_j-y_i|>\epsilon.$$

Then, from the Hoeffding inequality, we obtain

$$\begin{aligned}Pr\{C|B^c\} &\leqslant \exp\left\{\frac{-2\epsilon^2N}{\left(\max_{y\in\mathbb{R}}\frac{1}{N}\sum_{i=1}^{N}|y-y_i|\right)^2}\right\} \\ &\leqslant \exp\left\{\frac{-\epsilon^2N}{2(\|f\|_\infty+\sigma\sqrt{1/\delta_2})^2}\right\}.\end{aligned}$$

Let us set

$$\frac{\delta_1}{2} = \exp\left\{\frac{-\epsilon^2 N}{2(\|f\|_\infty + \sigma\sqrt{1/\delta_2})^2}\right\}.$$

Then, with a probability of at least $1 - \delta_1/2 - \delta_2$,

$$\frac{1}{N}\sum_{i=1}^{N}\int |y - y_i| dP(y) - \frac{1}{N^2}\sum_{i,j=1}^{N}|y_i - y_j| \leqslant \sqrt{\frac{2}{N}\ln\frac{2}{\delta_1}}\left(\|f\|_\infty + \sigma\sqrt{\frac{1}{\delta_2}}\right) \tag{27}$$

since

$$\begin{aligned} Pr\{C^c\} &\geqslant Pr\{C^c, B^c\} \\ &\geqslant Pr\{C^c|B^c\}Pr\{B^c\} \\ &\geqslant \left(1 - \frac{\delta_1}{2}\right)(1 - \delta_2) \\ &> 1 - \frac{\delta_1}{2} - \delta_2. \end{aligned}$$

Similar to (26), the difference between the expected and empirical risks of $|y - f_n(x)|$ is bounded by

$$\begin{aligned} \int_{X\times\mathbb{R}} |y - f_n(x)| dP(x,y) \quad &- \quad \frac{1}{N}\sum_{i=1}^{N}|y_i - f_n(x_i)| \\ &\leqslant \quad \frac{1}{N}\sum_{i=1}^{N}\int_{X\times\mathbb{R}} |y - y_i| + |f_n(x) - f_n(x_i)| dP(x,y). \end{aligned}$$

Finally, from (24) and (27), with a probability of at least $1 - \delta_1 - \delta_2$

$$\begin{aligned} \int_{X\times\mathbb{R}} |y - f_n(x)| dP(x,y) \quad &- \quad \frac{1}{N}\sum_{i=1}^{N}|y_i - f_n(x_i)| \\ &\leqslant \quad \frac{1}{N^2}\sum_{i,j=1}^{N}(|y_j - y_i| + |f_n(x_j) - f_n(x_i)|) \\ &\qquad + (\omega(f_n, h_0) + C)\sqrt{\frac{1}{2N}\ln\frac{2}{\delta_1}} \end{aligned} \tag{28}$$

where $C = |f_n(x_0) - f_n(y_0)| + 2\|f\|_\infty + 2\sigma\sqrt{1/\delta_2}$.

This completes the proof. □

This theorem states that the expected risk $R(f_n)_{L_1}$ is bounded by the empirical risk $R_{emp}(f_n)_{L_1}$, the second term of (21) describing the variation of output samples and estimation functions for the given input samples, and the third term dependent upon the modulus of continuity of estimation function plus some coefficients associated with target function, noise variance and the number of samples.

First, let us consider the second term. This term can be further break down to the empirical risk and the term depends on the target function. The next corollary shows the bounds on expected risks including this decomposition:

Corollary 1 *Let H_y be the $N \times N$ matrix in which the ij-th entry is given by $|y_i - y_j|$. Then, the following inequality holds with probability at least $1 - \delta_1 - \delta_2$:*

$$R(f_n)_{L_1} \leqslant 3R_{emp}(f_n)_{L_1} + \frac{2}{N}\max\{\lambda_i\} + (\omega(f_n, h_0) + C)\sqrt{\frac{1}{2N}\ln\frac{2}{\delta_1}} \tag{29}$$

where λ_i represents the ith eigenvalue of the matrix H_y.

Proof. Let the H_y be a matrix in which the ijth element is given by $|y_i - y_j|$ and an N dimensional vector $\mathbf{a}$ be given by

$$\mathbf{a} = \frac{1}{\sqrt{N}}(1, \cdots, 1)^T.$$

Then,

$$\frac{1}{N}\sum_{i,j=1}^{N}|y_i - y_j| = \mathbf{a}^T H_y \mathbf{a}.$$

Here, the matrix H_y can be decomposed by

$$H_y = E\Lambda E^T = \sum_{i=1}^{N}\lambda_i \mathbf{e}_i \mathbf{e}_i^T$$

where E represents a matrix in which the ith column vector is the ith eigenvector $\mathbf{e}_i$ and Λ represents the diagonal matrix in which the ith diagonal element is the ith eigenvalue λ_i. Thus,

$$\frac{1}{N}\sum_{i=1}^{N}|y_i - y_j| = \sum_{i=1}^{N}\lambda_i(\mathbf{a}^T\mathbf{e}_i)^2 \leqslant \max_i\{\lambda_i\}.$$

Now, let us consider the following inequality:

$$\begin{aligned}
\frac{1}{N^2}\sum_{i,j=1}^{N}|f_n(x_i) - f_n(x_j)| \quad &\leqslant \quad \frac{1}{N^2}\sum_{i,j=1}^{N}|f_n(x_i) - y_i| \\
&\quad + \frac{1}{N^2}\sum_{i,j=1}^{N}|y_i - y_j| + \frac{1}{N^2}\sum_{i,j=1}^{N}|y_j - f_n(x_j)| \\
&\quad = 2R_{emp}(f_n)_{L_1} + \frac{1}{N^2}\sum_{i,j=1}^{N}|y_i - y_j| \\
&\quad \leqslant 2R_{emp}(f_n)_{L_1} + \frac{1}{N}\max_i\{\lambda_i\}.
\end{aligned}$$

Therefore, from (21) of theorem 1 and the above inequality, the following inequality holds with a probability of at least $1 - \delta_1 - \delta_2$:

$$R(f_n) \leqslant 3R_{emp}(f_n) + \frac{2}{N}\max_i\{\lambda_i\} + (\omega(f_n, h_0) + C)\sqrt{\frac{1}{2N}\ln\frac{2}{\delta_1}}.$$

This completes the proof. □

This corollary states that the terms dependent upon estimation function f_n in (29) is the empirical risk $R_{emp}(f_n)$ and the modulus of continuity $w(f_n, h_0)$ since the eigenvalue λ_i of H_f and a constant C have insignificant influence to the shape of expected risks in accordance with varying the number of parameters n. The bounds on expected risks of (29) seem to be overestimated since the empirical risk is multiplied by 3. However, for our purpose of finding the model selection criteria, the estimation of the tight bound on the expected risk is not essential. Rather, the coefficient ratio between the empirical risk and modulus of continuity terms plays an important role for model selection since the optimal number of parameters can be determined by the minimum value of the estimated expected risks which are dependent upon these two terms. From this point of view, we can consider the model selection criteria referred to as the modulus of continuity information criteria (MCIC) as follows:

$$\mathrm{MCIC}(n) = R_{emp}(f_n)_{L_1} + \frac{\omega(f_n, h_0)}{3}\sqrt{\frac{1}{2N}\ln\frac{2}{\delta_1}}. \tag{30}$$

Suppose we have fixed number of samples N. Then, as the number of parameters n increases, the empirical risk $R_{emp}(f_n)$ decreases while the modulus of continuity $\omega(f_n, h_0)$ increases since the estimation function f_n becomes more complex function. From this point of view, we need to make a trade-off between over-fitting and under-fitting of regression models using the MCIC(n) for the optimization of regression models.

When we use the MCIC(n), a constant h_0 in (30) can be determined by the distance h_i between the point $x_i' = \arg\max_{x\in X}|f_n(x) - f_n(x_i)|$ and the nearest sample point $x_{i'} \in \{x_1, \ldots, x_N\}$ satisfying the condition of (23). In general, h_0 lies within the following bound:

$$\frac{1}{2}\min_i \min_{j\neq i} d(x_i, x_j) \leqslant h_0 \leqslant \frac{1}{2}\max_i \min_{j\neq i} d(x_i, x_j). \tag{31}$$

If samples are evenly distributed in $[a, b]$, we can set h_0 as

$$h_0 = \frac{1}{2}\frac{b-a}{N+1}. \tag{32}$$

That is, the half length of the equidistance of $N+1$ intervals on $[a, b]$.

In general, the modulus of continuity term in (30) increases as n increases. However, if the selected estimation function f_n is optimal for every n, we can show that the modulus of continuity term decreases as n increases. The following theorem states this phenomenon:

Theorem 2 *Let $f_n \in T_n$ be the polynomials of degree n approximating a function $f \in C^*[-\pi, \pi]$, that is, $\|f - f_n\|_\infty = \inf_{f_n\in T_n}\|f - f_n\|_\infty = E_n^*(f)$. Then, the following inequality holds with a probability of at least $1 - \delta_1$,*

$$R(f_n)_{L_1} \leqslant R_{emp}(f_n)_{L_1} + \omega\left(f, \frac{1}{n}\right)\sqrt{\frac{1}{2N}\ln\frac{1}{\delta_1}}. \tag{33}$$

Furthermore, the same result can be obtained provided that f_n is the algebraic polynomials P_n approximating a function $f \in C[-1, 1]$.

Proof. In the classes of both algebraic and trigonometric polynomials of order n, the L_∞ norm of $f - f_n$ is bounded by

$$E_n^*(f) = \inf_{f_n}\|f - f_n\|_\infty \leqslant \omega(f, \frac{1}{n})$$

where f_n represents the polynomials of the best approximation of $f \in C^*[-\pi,\pi]$ or $C[-1,1]$.

From the Hoeffding inequality, we obtain

$$\begin{aligned} Pr & \left\{ \int_X |f(x) - f_n(x)| dP(x) - \frac{1}{N} \sum_{i=1}^{N} |f(x_i) - f_n(x_i)| > \epsilon \right\} \\ & \leqslant \exp\left\{ -\frac{2\epsilon^2 N}{(\max|f(x) - f_n(x)|)^2} \right\} \\ & \leqslant \exp\left\{ -\frac{2\epsilon^2 N}{\omega(f, 1/n)^2} \right\}. \end{aligned}$$

Here, we set

$$\delta_1 = \exp\left\{ -\frac{2\epsilon^2 N}{\omega(f, 1/n)^2} \right\}.$$

Then, with a probability of at least $1 - \delta_1$, the following inequality holds:

$$\int_X |f(x) - f_n(x)| dP(x) \leqslant \frac{1}{N} \sum_{i=1}^{N} |f(x_i) - f_n(x_i)| + \omega(f, \frac{1}{n}) \sqrt{\frac{1}{2N} \ln \frac{1}{\delta_1}}.$$

This completes the proof. □

Note that the modulus of continuity term in (33) is dependent upon the target function f, not the estimation function f_n when f_n is the best approximation polynomial. Therefore, as we increase the number of parameters n, the modulus of continuity term in (33) tends to decrease, that is, the larger the hypothesis space is, the smaller the confidence interval of the expected risk is.

The distinctive characteristics of the suggested model selection criteria MCIC(n) can be described as follows:

- The suggested MCIC can be applied to nonlinear models with an arbitrary sample distribution while the AIC and BIC are good measures of linear models with a large number of samples.

- The suggested MCIC depends on the modulus of continuity for a specific hypothesis generated by a learning algorithm and also on the sample distribution, while the VC dimension or degree of freedoms based methods consider the structural information only, that is, the complexity of a hypothesis space and the number of samples, and does not depend on a specific hypothesis and the sample distribution.

Considering these characteristics, the MCIC can be a good measure for model selection when the regression model is trained using a certain learning algorithm for the given distribution of training samples, since the regression model is selected using the modulus of continuity (or in a sense, the smoothness) for the estimation function generated by a learning algorithm, and the term h_0 in the argument of the modulus of continuity is dependent upon the sample distribution.

4 Modulus of Continuity for Estimation Functions

For regression models, we will suggest a method of estimating the modulus of continuity. First, let f satisfy the Lipschitz condition Lipα, $0 < \alpha \leqslant 1$, that is, the condition of (15). Then, for every $f \in C^1(a, b)$, we have $\omega(f, h) \leqslant \|f'\|_\infty h$. For this class of functions, Marchaud[13] showed the following theorem:

Theorem (modulus of continuity for $f \in C^1$)

Let l be the length of domain X. Then, the modulus of continuity for $f \in C^1$ is bounded by the following inequality:

$$\omega(f,h) \leqslant h \int_h^{\frac{l}{2}} \frac{\omega_2(f,u)}{u^2} du + \frac{2h}{l} \|f\|_\infty. \tag{34}$$

where $\omega_2(f, u)$ is defined by

$$\omega_2(f,u) = \max_{x, x\pm t \in X, |t| \leqslant u} |f(x+t) - 2f(x) + f(x-t)|. \tag{35}$$

If the given function f is Lip*1, then f is qusi-smooth, that is, $\omega_2(f,h) = \mathcal{O}(h)$, and it follows that*

$$\omega(f,h) \leqslant \mathcal{O}\left(h \left| \ln \frac{1}{h} \right| \right). \tag{36}$$

From the bounds of modulus of continuity described in (15) or (34), the modulus of continuity for $f \in C^1$ can be obtained by the Lip1 condition or Marchaud theorem, that is,

$$\omega(f,h) \leqslant \|f'\| h \text{ or}$$
$$\omega(f,h) \leqslant 2h\|f'\| \ln \frac{l}{2h} + \frac{2h}{l}\|f\|_\infty.$$

If we compare the two methods, we can identify that the bound of the modulus of continuity $w(f, h)$ using the Lipschitz condition is less than that of using the Marchaud's method when h is small as illustrated in Figure 2: $w(f, h)$ using the Lipschitz condition is less than that of using the Marchaud's method for $h < h'$ in which h' is a constant between 0 and $l/2$ satisfying

$$\|f'\|h' = 2h'\|f'\| \ln \frac{l}{2h'} + \frac{2h'}{l}\|f\|_\infty.$$

However, the Marchaud method has a useful sample property which is dependent on the given domain length of X. Here, we assume that the samples are densely distributed in X so that we can take h_0 as a small value. In this case, the modulus of continuity using the Lipschitz condition can be a good candidate for estimating the modulus of continuity for regression models.

The modulus of continuity for the estimation function f_n is dependent upon the basis function ϕ_k in (4). Here, we consider the cases of estimating $w(f_n, h)$ when the basis functions are trigonometric polynomials, Gaussian functions, and sigmoid functions:

Figure 2: A plot of $w(f, h)$ versus h: the modulus of continuity using the Lipschitz condition is tighter than the modulus of continuity using the Marchaud method when $h < h'$ and the point h' is dependent upon the ratio value $\frac{\|f\|}{\|f'\|}$.

Case 1: Consider the case of the estimation function f_n with trigonometric polynomials:

$$\left\{\frac{1}{2}, \cos px, \sin px\right\}.$$

Applying the mean value theorem to ϕ_k, we get

$$\begin{aligned} \omega(\phi_k, h) &\leqslant \|\phi_k'\|_\infty h \\ &\leqslant \left\lfloor \frac{k}{2} \right\rfloor h, \quad \text{for} \quad k = 1, \cdots, n. \end{aligned}$$

Therefore, the modulus of continuity for f_n can be determined by

$$\omega(f_n, h) \leqslant \sum_{k=0}^{n} h|w_k| \cdot \left\lfloor \frac{k}{2} \right\rfloor. \tag{37}$$

Case 2: Consider the case of the estimation function f_n with radial basis functions:

$$\phi_k(x) = \exp\left(-\frac{(x-t_k)^2}{2\sigma_k^2}\right) \quad \text{for} \quad k = 1, \cdots, n$$

where t_k and σ_k^2 represent the center and width of the basis function ϕ_k respectively. Applying the mean value theorem to ϕ_k, we get

$$\begin{aligned} \omega(\phi_k, h) &\leqslant \|\phi_k'\|_\infty h \\ &\leqslant \frac{h}{\sigma_k} \exp\left(-\frac{1}{2}\right) \quad \text{for} \quad k = 1, \cdots, n \end{aligned}$$

since ϕ_k' has the maximum and minimum at $x = t_k - \sigma_k$ and $x = t_k + \sigma_k$ respectively. Therefore, the modulus of continuity for f_n can be determined by

$$\omega(f_n, h) \leqslant h \sum_{k=1}^{n} \frac{1}{\sigma_k} \exp(-1/2)|w_k|. \tag{38}$$

Case 3: Consider the case of the estimation function f_n with sigmoid functions:

$$\phi_k(x) = \frac{1}{1 + \exp(-a_k x + b_k)} \quad \text{for } k = 1, \cdots, n$$

where $a_k (> 0)$ and b_k represent the adaptable parameters of the basis function ϕ_k.

Applying the mean value theorem to ϕ_k, we get

$$\begin{aligned} w(\phi_k, h) &\leqslant \|\phi_k'\|_\infty h \\ &\leqslant \frac{h}{4} a_k \text{ for } k = 1, \cdots, n \end{aligned}$$

since ϕ_k' has the maximum value at $x = b_k/a_k$.

Therefore, the modulus of continuity for f_n can be determined by

$$\omega(f_n, h) \leqslant \frac{h}{4} \sum_{k=1}^{n} |w_k| a_k. \tag{39}$$

5 Model Selection Using $R(f_n)_{L_1}$ versus $R(f_n)_{L_2}$

The modulus of continuity explained in theorem 1 is derived in the L_1 sense, that is, the loss function $L(y, f(x)) = |y - f(x)|$. In general, we can consider the modulus of continuity in the L_p sense as explained in the following definition:

Definition (the modulus of continuity in the L_p sense)
If the Lebesgue measurable function $f(x) \in L_p[a,b]$, then L_p integral modulus of continuity is defined by

$$\omega(f, h)_{L_p[a,b]} = \sup_{|t| \leqslant h} \left(\int_a^b |f(x+t) - f(x)|^p dx \right)^{1/p} \tag{40}$$

since we consider the domain Ω to be a probability space, we need the new definition of modulus of continuity of f replacing Lebesgue measure. Let $(\Omega, \mathcal{B}, P)$ be a probability space. Then, the L_p-modulus of $f \in L_p(\Omega)$ on the space Ω is defined by

$$\omega(f, h)_{L_p, P} = \sup_{|t| \leqslant h} \left(\int_\Omega |f(x+t) - f(x)|^p dP(x) \right)^{1/p}. \tag{41}$$

For convenience's sake, we will describe $\omega(f, h)_{L_p, P}$ as $\omega(f, h)_{L_p}$. Here, we can describe the distance between f and f_n in the L_p measure using $\omega(f, h)_{L_p}$ in the following theorem:

Theorem 3 *Let the value h satisfy the condition that $x_i + t \in X$, $i = 1, ..., N$, and $|t| \leqslant h$. Then, the distance between f and f_n in the L_p $(p \geqslant 1)$ sense is bounded by*

$$\|f - f_n\|_p \leqslant \mathcal{O}\left(\omega(f - f_n, h)_{L_p} + \left(\frac{1}{N} \sum_{i=1}^{n} |f(x_i) - f_n(x_i)|^p \right)^{1/p} + \left(\frac{1}{\sqrt{N}} \right)^{1/p} \right).$$

Proof. The following inequality holds from the Minkowski inequality:

$$\begin{aligned} \|f - f_n\|_p &\leqslant \|f - f_n - (f(\cdot + t) - f_n(\cdot + t))\|_p + \|f(\cdot + t) - f_n(\cdot + t)\|_p \\ &\leqslant \omega(f - f_n, h)_{L_p} + \|f(\cdot + t) - f_n(\cdot + t)\|_p. \end{aligned}$$

For $i.\ i.\ d.$ random variables $x_i + t \in X,\ i = 1, \ldots, N$, the following inequality holds from the Hoeffding inequality:

$$\begin{aligned}
&Pr\left\{\int_X |f(x+t) - f_n(x+t)|^p dP(x) - \frac{1}{N}\sum_{i=1}^{N} |f(x_i+t) - f_n(x_i+t)|^p > \epsilon\right\}\\
&\leqslant \exp\left\{\frac{-\epsilon^2 N}{\max_{x\in X} |f(x+t) - f_n(x+t)|^{2p}}\right\}\\
&\leqslant \exp\left\{\frac{-\epsilon^2 N}{\|f(\cdot+t) - f_n(\cdot+t)\|_\infty^{2p}}\right\}.
\end{aligned}$$

Let us set

$$M = \max\left\{\|f - f_n\|_\infty, \|f(\cdot+t) - f_n(\cdot+t)\|_\infty\right\} \quad \text{and}$$

$$\delta_1 = \exp\left\{\frac{-\epsilon^2 N}{M^{2p}}\right\}.$$

Then, with a probability of at least $1 - \delta_1/2$, the following inequality holds:

$$\|f(x+t) - f_n(x+t)\|_p \leqslant \left(\frac{1}{N}\sum_{i=1}^{N} |f(x_i+t) - f_n(x_i+t)|^p + \sqrt{\frac{1}{N}\ln\frac{2}{\delta_1}}M^p\right)^{1/p}.$$

The empirical part of the above equation can be redescribed in the following form with a probability of at least $1 - \delta_1/2$:

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N} |f(x_i+t) - f_n(x_i+t)|^p\\
&\leqslant \frac{1}{N}\sum_{i=1}^{N} \left[|f(x_i) - f_n(x_i)| + |f(x_i+t) - f_n(x_i+t) - (f(x_i) - f_n(x_i))|\right]^p\\
&\leqslant \frac{2^p}{N}\sum_{i=1}^{N} \left[|f(x_i) - f_n(x_i)|^p + |f(x_i+t) - f_n(x_i+t) - (f(x_i) - f_n(x_i))|^p\right]\\
&\leqslant \frac{2^p}{N}\sum_{i=1}^{N} |f(x_i) - f_n(x_i)|^p + 2^p\omega(f - f_n, h)_{L_p}^p + 2^{p+1}M^p\sqrt{\frac{-\ln(\delta_1/2)}{2N}}
\end{aligned}$$

for $p \geqslant 1$ and $0 < |t| \leqslant h$.

Therefore, with a probability of at least $1 - \delta_1$, the following inequality holds:

$$\|f - f_n\|_p \leqslant \mathcal{O}\left(\omega(f - f_n, h)_{L_p} + \left(\frac{1}{N}\sum_{i=1}^{n} |f(x_i) - f_n(x_i)|^p\right)^{1/p} + \left(\frac{1}{\sqrt{N}}\right)^{1/p}\right).$$

This completes the proof. □

This theorem states that the distance between f and f_n in the L_p sense is represented by the newly defined modulus of continuity $\omega(f - f_n, h)_{L_p}$ whose value changes with respect to f_n as n varies. However, it is difficult to calculate the modulus of continuity in the L_p sense because the probability distribution P in (41) is not known in general. From this point of view, we consider the model selection criteria of (30) based on the modulus of continuity in the L_1 sense. In some cases, the regression models of the L_1 sense optimality are also consistent with the L_2 sense optimality. The next theorem explains the relationship between the L_1 and L_2 sense optimality:

Theorem 4 *Let*

$$E[|g|] = \int |g| d\mu \leqslant E[|h|] = \int |h| d\mu. \tag{42}$$

Then,

$$E[|g|^2] = \int |g|^2 d\mu \leqslant E[|h|^2] = \int |h|^2 d\mu \tag{43}$$

under the following condition:

$$\begin{aligned} &(a)\ Var(|g|) &\leqslant&\ Var(|h|)\ \textit{or} &(44)\\ &(b)\ Var(|g|) &\geqslant&\ Var(|h|)\ \textit{and} &\\ &E[|h|] - E[|g|] &\geqslant&\ -\int |g| d\mu + \sqrt{\left(\int |g| d\mu\right)^2 + Var(|g|) - Var(|h|)}. &(45)\end{aligned}$$

Proof. (a) Let us consider the difference between two variances, $Var(|h|)$ and $Var(|g|)$:

$$\begin{aligned} Var(|h|) - Var(|g|) &= \int (|h| - E[|h|])^2 d\mu - \int (|g| - E[|g|])^2 d\mu \\ &= \int |h|^2 d\mu - \int |g|^2 d\mu - E^2[|h|] + E^2[|g|]. \end{aligned}$$

This implies that

$$\int |h|^2 d\mu - \int |g|^2 d\mu = Var(|h|) - Var(|g|) + E^2[|h|] - E^2[|g|] \geqslant 0$$

from the given conditions.

(b) Let us consider

$$c = E[|h|] - E[|g|] = \int (|h| - |g|) d\mu.$$

Then,

$$\begin{aligned} \int |h|^2 - |g|^2 d\mu &= \mathrm{Var}(|h|) - \int |g|^2 d\mu + \left(\int |h| d\mu\right)^2 \\ &= \mathrm{Var}(|h|) - \int |g|^2 d\mu + \left(\int |g| d\mu + c\right)^2 \\ &= \mathrm{Var}(|h|) - \mathrm{Var}(|g|) + 2c \int |g| d\mu + c^2 \\ &= \left(c + \int |g| d\mu\right)^2 - \left(\left(\int |g| d\mu\right)^2 + \mathrm{Var}(|g|) - \mathrm{Var}(|h|)\right). \end{aligned}$$

Therefore, if

$$c \geqslant -\int |g| d\mu + \sqrt{\left(\int |g| d\mu\right)^2 + \mathrm{Var}(|g|) - \mathrm{Var}(|h|)},$$

then the following inequality holds:

$$\int |h|^2 d\mu - \int |g|^2 d\mu \geqslant 0.$$

This completes the proof. □

In the regression models, square loss functions, that is, $L(y, f_n(x)) = (y - f_n(x))^2$ are usually used since the minimization of the risk in the L_2 sense is optimal under the assumption that the estimation function is unbiased and the error term $y - f_n(x)$ has normal distribution. Under this assumption, we can show that the optimality in the L_1 sense is also optimal in the L_2 sense, that is, the optimization of regression model in the L_1 sense is also the maximum likelihood estimate. Here, let us assume that

$$g = y - f_n \sim N(0, \sigma_n^2) \quad \text{and}$$

$$h = y - f_m \sim N(0, \sigma_m^2).$$

Then, the expectations for $|g|$ and $|h|$ are given by

$$E[|g|] = \int_{-\infty}^{+\infty} \frac{|x|}{\sqrt{2\pi}\sigma_n} \exp(\frac{-x^2}{2\sigma_n^2}) dx = \frac{2\sqrt{2}\sigma_n}{\sqrt{\pi}} \quad \text{and}$$

$$E[|h|] = \int_{-\infty}^{+\infty} \frac{|x|}{\sqrt{2\pi}\sigma_m} \exp(\frac{-x^2}{2\sigma_m^2}) dx = \frac{2\sqrt{2}\sigma_m}{\sqrt{\pi}}$$

respectively. Therefore, if $E[|g|] \leqslant E[|h|]$, then $\sigma_n \leqslant \sigma_m$ and the condition of (a) in theorem 4 is satisfied.

If $E[|g|]$ is smaller than $E[|h|]$ but $Var(|g|)$ is larger than $Var(|h|)$, the optimality in the L_1 or L_2 sense can be different. In the case that the difference between $E[|h|]$ and $E[|g|]$ is larger than the square root of the difference between $Var(|g|)$ and $Var(|h|)$, we can guarantee that the optimality in the L_1 or L_2 sense is the same since the inequality of (45) is satisfied.

6 Simulation

The simulation for function approximation was performed using the regression model with trigonometric polynomial functions. In this regression, various model selection methods such as the AIC, BIC, VC dimension, and the suggested MCIC based methods were tested. As the VC dimension based method, we choose the smallest eigenvalue based (SEB) method[8] suggested by Chapelle et al. Here, the trigonometric polynomial functions were selected as the basis functions since this network could be considered as a linear regression model so that the VC dimension of the network was easy to find out and the optimization of the network parameters could be easily solved. This makes it easier to compare the suggested method with other methods. In this simulation, the data (x_i, y_i) were generated by (1). Here, the input value x_i was uniformly distributed within the interval of $[-\pi, \pi]$ and the noise has the normal distribution of mean 0 and standard variance $\sigma = 0, 0.025, 0.05,$ and 0.1 respectively. As the target functions, we considered the following functions:

1) the step function defined by

$$f(x) = \begin{cases} 1 & \text{if } x \geqslant 0 \\ 0 & \text{otherwise,} \end{cases} \tag{46}$$

2) the sine square function defined by

$$f(x) = \sin^2(x), \tag{47}$$

3) the sinc function defined by

$$f(x) = \frac{\sin(\pi x)}{\pi x}, \tag{48}$$

4) and the combined function defined by

$$f(x) = 2x \exp(-\frac{x^2}{2}) \cos(2\pi x). \tag{49}$$

The target functions are illustrated in Figure 3. The first target function was discontinuous while the other functions were continuous. The second and third functions showed a similar complexity while the fourth function was more complex than the other functions. For these target functions, we generated N $(= 50)$ training samples randomly to train the estimation network. We also generated 300 test samples separately according to same input and noise distributions. For the estimation network, the basis functions of trigonometric polynomial network were given by

$$\phi_0(x) = \frac{1}{2},\ \phi_{2j-1}(x) = \sin jx, \text{ and } \phi_{2j}(x) = \cos jx,$$

where $2\pi/j$ represents the period of the sinusoidal function. Here, the estimation function $f_n(x)$ was given by

$$f_n(x) = \sum_{k=0}^{n} w_k \phi_k(x). \tag{50}$$

For N samples, the observation vector defined by $\mathbf{y} = (y_1, \cdots, y_N)^T$ can be approximated by the following vector form:

$$\mathbf{y} = \Phi_n \mathbf{w} \tag{51}$$

where Φ_n was a matrix in which the ij-th element was given by $\phi_j(x_i)$ and $\mathbf{w}$ was a weight vector defined by $\mathbf{w} = (w_0, \cdots, w_n)^T$. From the empirical risk minimization of the square loss function, the estimated weight vector $\widehat{\mathbf{w}}$ could be determined by

$$\widehat{\mathbf{w}} = (\Phi_n^T \Phi_n)^{-1} \Phi_n^T \mathbf{y}. \tag{52}$$

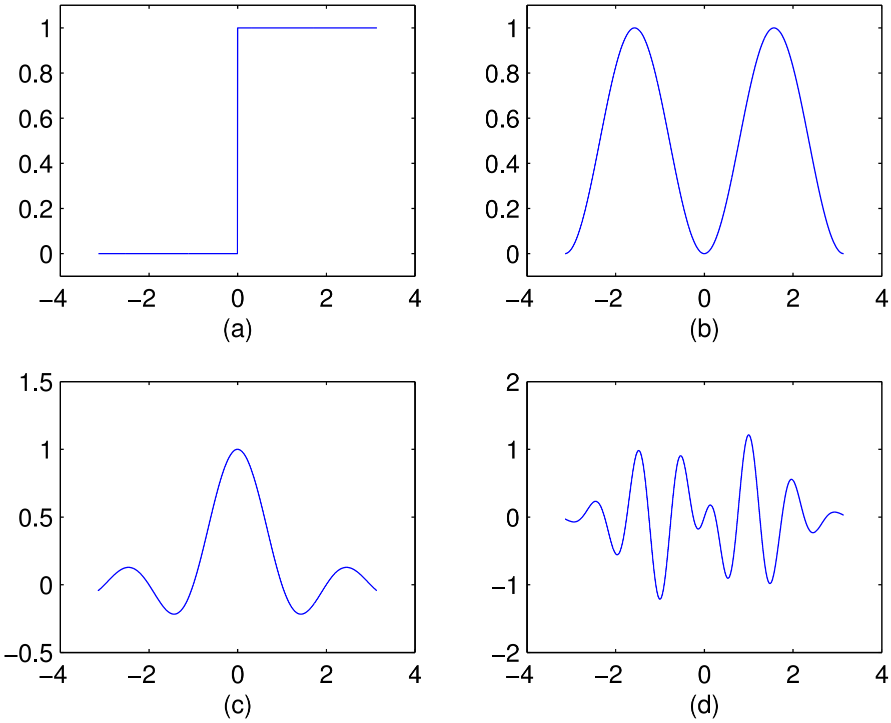


Figure 3: The target functions for the simulation of model selection:(a), (b), (c), and (d) represent the step, sin-square, sinc, and combined functions respectively.

By substituting the estimated weight vector for (51), we obtained the empirical risk $R_{emp}(f_n)$ evaluated by the training samples, and the estimated risk $\widehat{R}(f_n)$ could be determined by the AIC, BIC, SEB, and MCIC based methods. In the case of the MCIC method, only the terms of $R_{emp}(f_n)$ and the modulus of continuity for the estimation function f_n were considered to select the optimal number of nodes as indicated in the model selection criteria of (30). Here, the model selection criteria based on the modulus of continuity for the trigonometric polynomial function was determined by

$$MCIC(n) = R_{emp}(f_n) + \frac{h_0}{3}\sqrt{\frac{1}{2N}\ln\frac{2}{\delta}}\sum_{i=0}^{n}|\widehat{w}_i|\left\lfloor\frac{i}{2}\right\rfloor \tag{53}$$

from (30) and (37). Here, we set h_0 as $\pi/(N+1)$ (assuming uniform distribution) and δ as 0.05.

To compare the performance of model selection methods, the estimated optimal number of nodes (or basis functions) was determined by

$$\widehat{n} = \arg\min_{n}\widehat{R}(f_n). \tag{54}$$

The expected risks obtained for the estimated optimal number of nodes $\widehat{n}$ were compared with the expected risks for the minimum number of nodes obtained from the test samples, that is, we computed the log ratio of two risks [8]

$$r_R = \log\frac{R(f_{\widehat{n}})}{\min_n R(f_n)} \tag{55}$$

where $R(f_n)$ represented the expected risk for the squared error loss function $L(y, f_n) = (y - f_n(x))^2$ evaluated by the test samples. This risk ratio represented the quality of distance between the optimal and the estimated optimal risks. We also computed the log ratio of the estimated optimal number of nodes $\widehat{n}$ to the minimum number of nodes obtained from the test samples, that is,

$$r_n = \log\frac{\widehat{n}}{\arg\min_n R(f_n)}. \tag{56}$$

This node ratio represented the quality of distance between the optimal and the estimated optimal number of nodes. After all experiments had been repeated 1000 times, the risk ratios of (55) and the node ratios of (56) were plotted using the box-plot method. The simulation results of model selection using the AIC, BIC, SEB and MCIC based methods are illustrated in Figures 4 through 7. These simulation results showed us that 1) the SEB based method outperformed the AIC and BIC based methods from the view point of risk ratios in the case of the first, second, and third target functions but not in the more complicated fourth function, 2) while the MCIC based method demonstrated the top level performance from the view points of risk and node ratios for all four target functions. In general, the SEB method showed good performance when the ratio of the optimal number of nodes to the number of samples n^*/l was small as illustrated in Figure 8. Note that the fourth target function required high n^*/l compared to other target functions due to the complexity of function as shown in Figure 3. In this case, the SEB method did not show good performance as shown in Figure 7.

To see the risk prediction of the model selection methods, we plotted the estimated error versus the number of nodes n for the AIC, BIC, SEB, and MCIC methods. The predicted results were compared with test errors with respect to the L_1 or L_2 sense as shown in Figures 9 and 10. These figures showed that the trends of the error curves in the L_1 and L_2 senses were similar although the exact comparison of the two L_1 and L_2 error values was not possible. These results showed that 1) the risk prediction using the AIC and BIC methods fit well except for the sudden change in risk functions at large n, 2) the risk prediction using the SEB method had a tendency to fit well when the number of nodes n was small, and 3) the risk prediction using the MCIC method was well suited with the test errors in overall range of the number of nodes. In these predicted results, it is interesting to note that the MCIC method was able to predict the sudden change of test errors while the other methods weren't.

In summary, the performance of the MCIC method showed the better performance for various types of target functions from the view points of risk and node ratios compared to other methods. We also demonstrated that the risk prediction using the MCIC method was able to predict the trend of test errors. They are mainly due to the fact that the MCIC represent the risk estimates for trained regression models (not the structure of regression models).

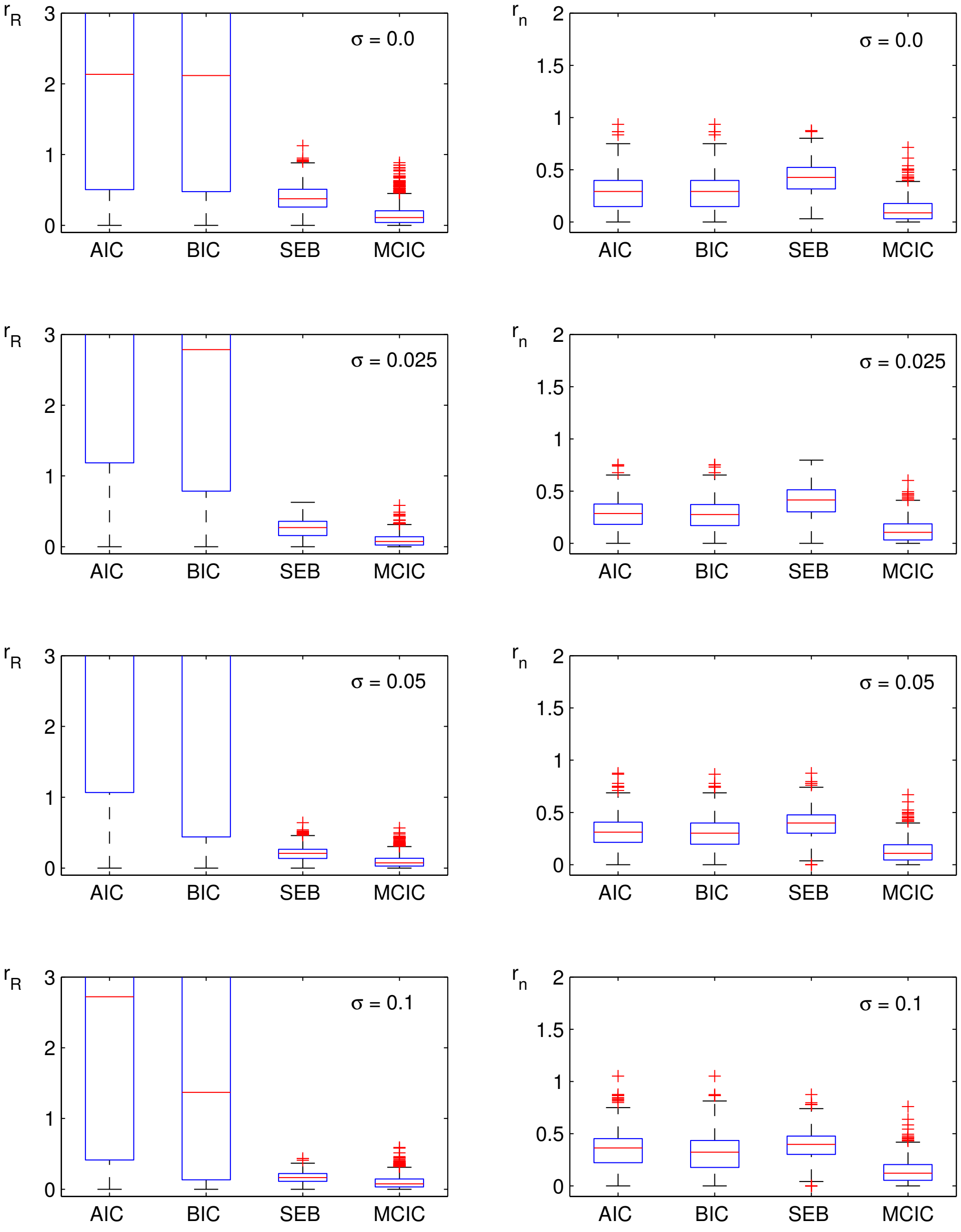


Figure 4: The left box-plots of risk ratios r_R and the right box-plot of node ratios r_n for the regression of step function.

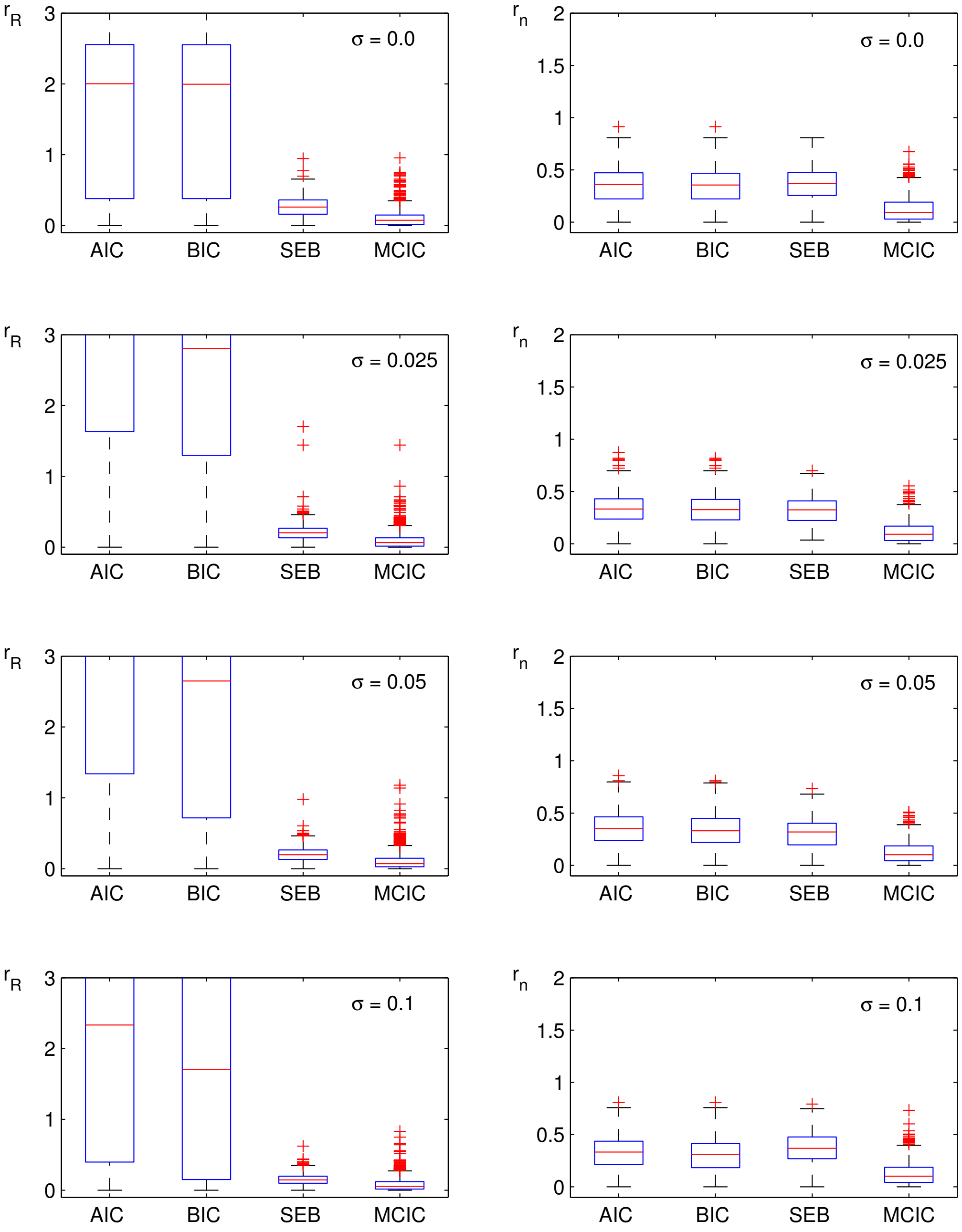


Figure 5: The left box-plots of risk ratios r_R and the right box-plot of node ratios r_n for the regression of sine-square function.

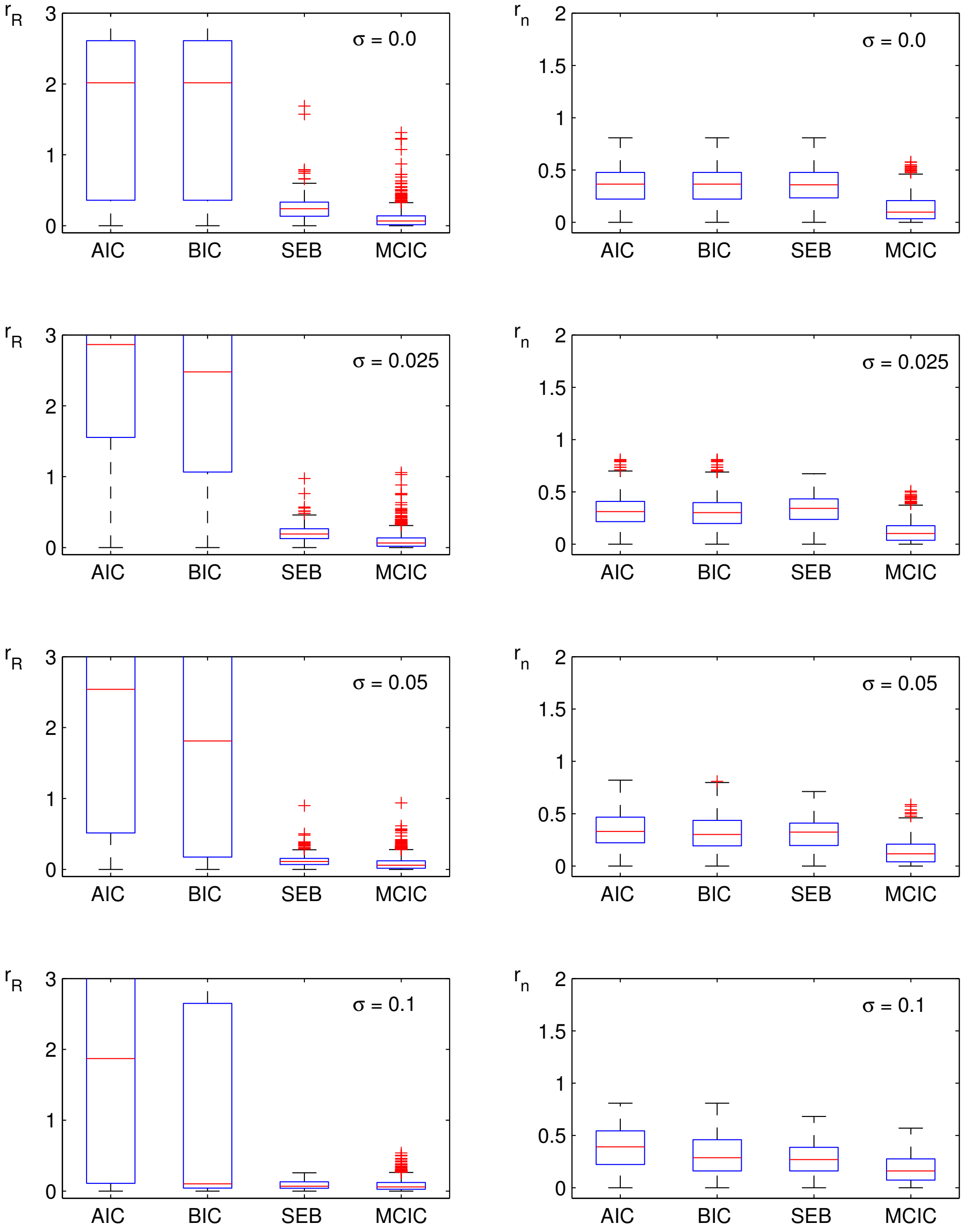


Figure 6: The left box-plots of risk ratios r_R and the right box-plot of node ratios r_n for the regression of sinc function.

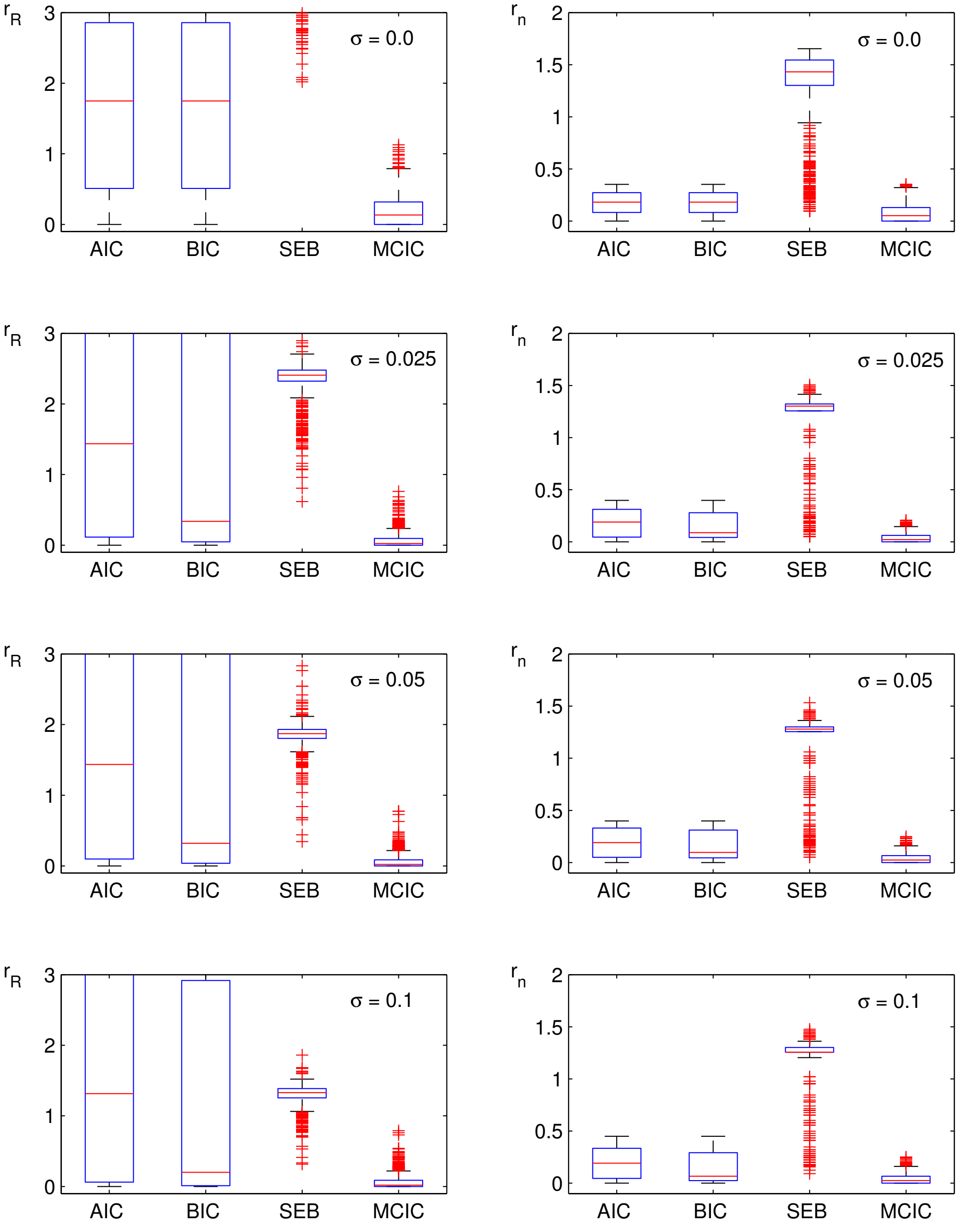


Figure 7: The left box-plots of risk ratios r_R and the right box-plot of node ratios r_n for the regression of combined function.

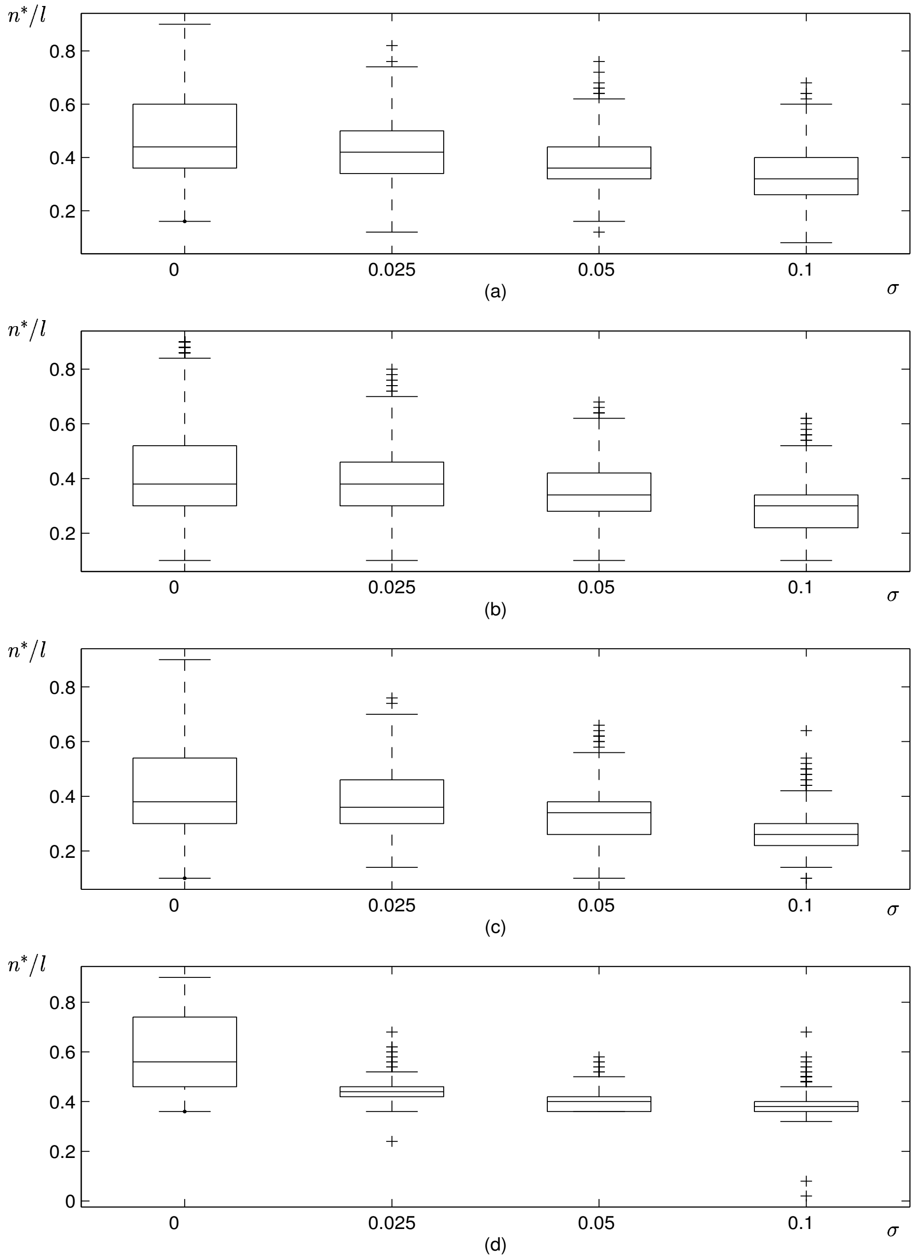
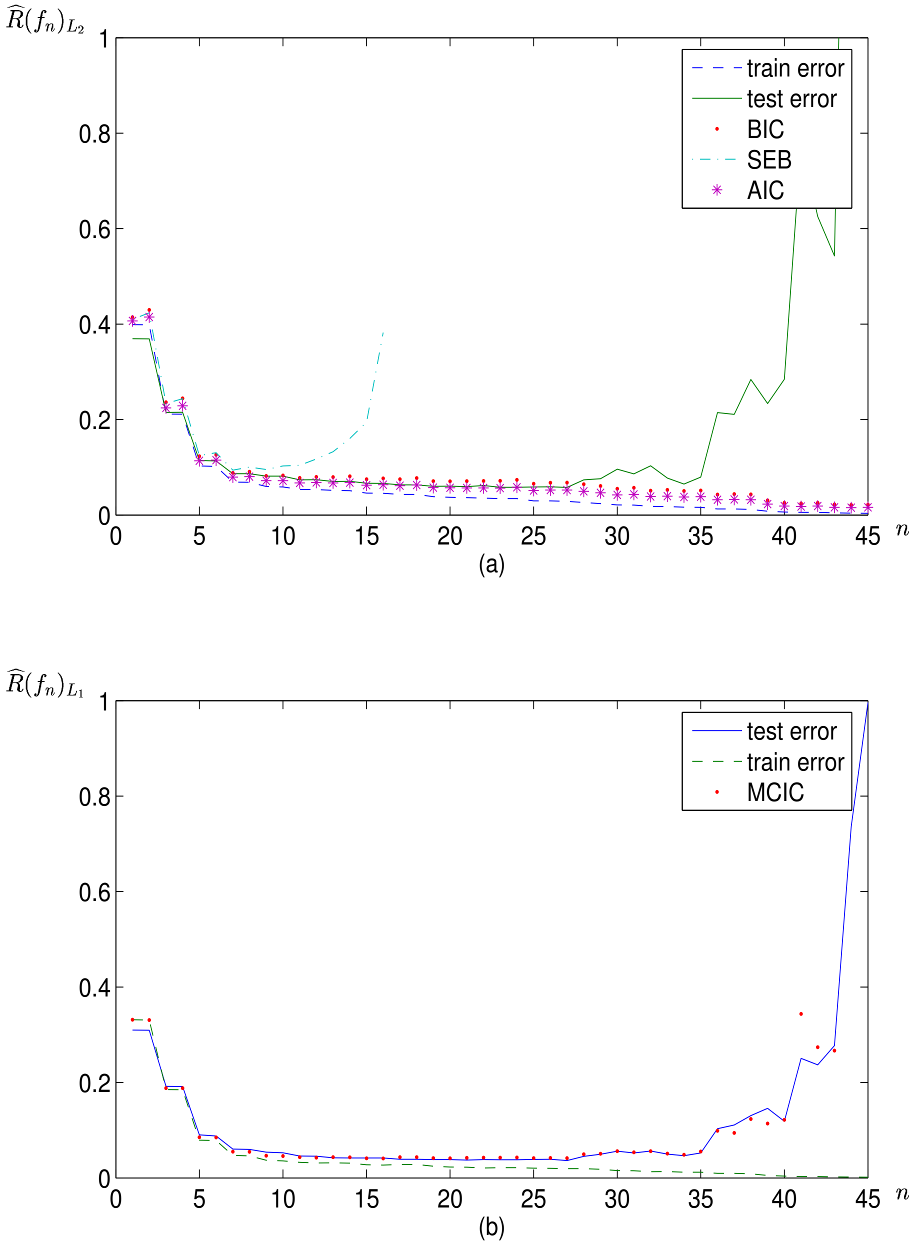


Figure 8: The box-plots for the optimal number of nodes to the number of samples n^*/l: (a), (b), (c), and (d) represent the box-plots of n^*/l in the regression of step, sine-square, sinc, and combined functions respectively.



(a)

(b)

Figure 9: Performance predictions of model selection methods in the case of the sinc function: (a) represents the estimated error in the L_2 sense versus the number of nodes using the AIC, BIC, and SEB methods and (b) represents the estimated error in the L_1 sense versus the number of nodes using the MC method.

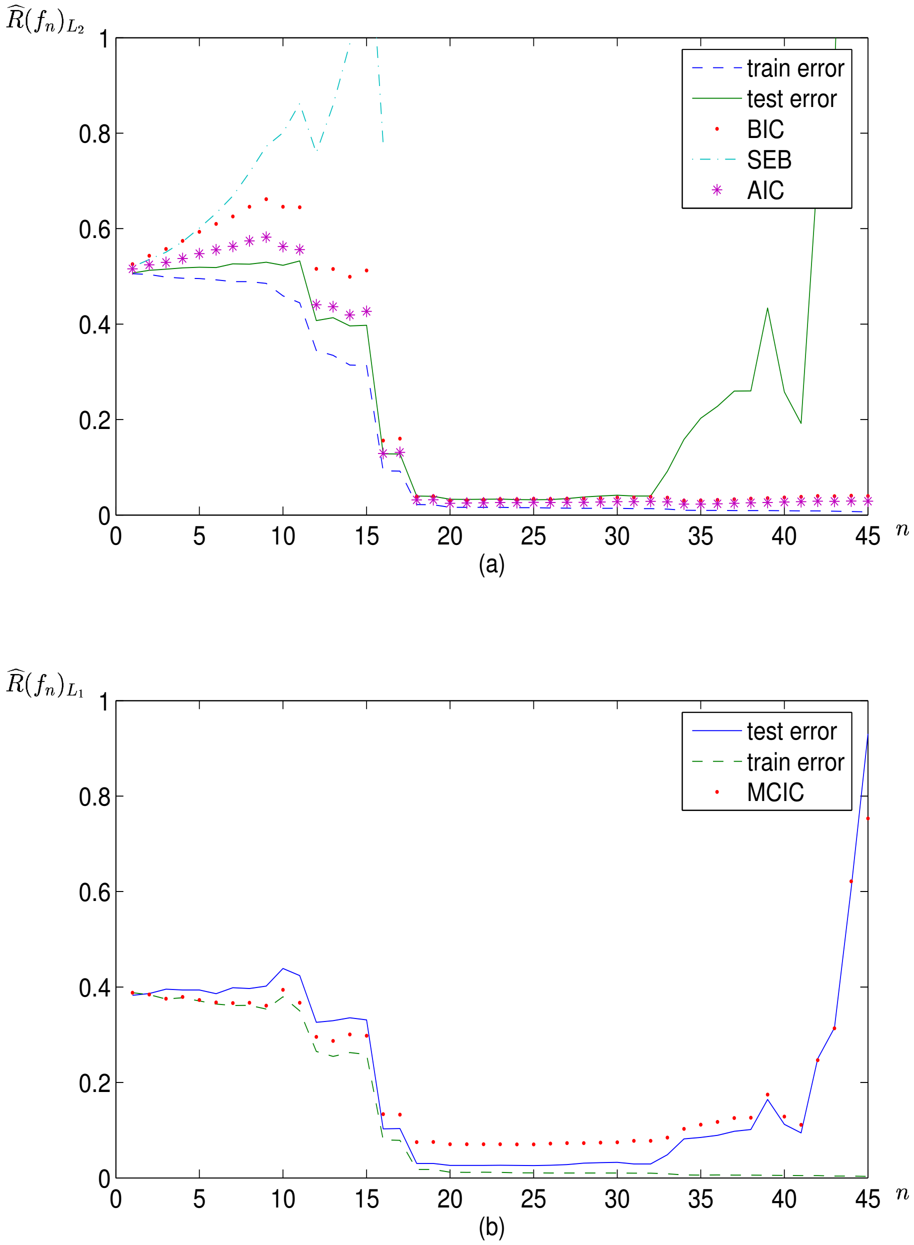


Figure 10: Performance predictions of model selection methods in the case of the combined function: (a) represents the estimated error in the L_2 sense versus the number of nodes using the AIC, BIC, and SEB methods and (b) represents the estimated error in the L_1 sense versus the number of nodes using the MC method.

7 Conclusion

We have suggested a new method of model selection in regression problems based on the modulus of continuity. The risk function bounds are investigated from the view point of the modulus of continuity for both the target and estimation functions. The final form of the risk function bounds incorporate the information of learned results such as the parameter values of regression models. To verify the validity of the suggested bound, the model selection in regression with the trigonometric polynomials was applied to function approximation problems. As a result, the suggested method showed a better performance for various types of target functions from the view points of risk and node ratios compared to other model selection methods such as AIC, BIC, and SEB. We also demonstrated that the risk prediction using the MCIC method was able to predict the trend of test errors. They are mainly due to the fact that the MCIC represent the risk estimates for trained regression models (not the structure of regression models). Furthermore, the suggested MCIC method can be easily extended to various types of regression models with nonlinear kernel functions which have some smoothness constraints.

For the extension of this work, we are considering the derivation of the bounds on expected risks using the MC in the regression of multi-dimensional estimation functions. To apply the MC to the multi-dimensional input space $X \subset \mathbb{R}^d$ $(d > 1)$, we have to consider two important issues: the first one is selecting the proper definition of the MC from various definitions of the MC associated with the input space X[15] so that the better prediction of optimal models is possible, and the second one is to determine h_0 which makes good relationship between the MC and expected risk. This will be remained as our future work.

References

[1] H. Akaike, Information theory and an extansion of the maximum likelihood principle. *In B. Petrov and F. Csaki, editors, 2nd Interanational Symposium on Information Theory, Budapest*, 22 (1973), 267–281.

[2] G. Schwartz, Estimating the dimension of a model. *Annals of Statistics*, 6 (1978), 461–464.

[3] G. Wahba, G.Golub, and M. Heath, Generalized cross-validation as a method for choosing a good ridge parameter, *Technometrics*, 21 (1979), 215–223.

[4] J. Rissanen, Stochastic complexity and modeling, *Annals of Statistics*, 14 (1986), 1080–1100.

[5] A. Barron, J. Rissanen, and B. Yu, The minimum description length principle in coding and modeling, *IEEE Transactions on Infomation Theory*, 44 (1998), 2743–2760.

[6] D. Foster and E. George, The risk inflation criterion for multiple regression, *Annals of Statistics*, 22 (1994), 1947–1975.

[7] V. Vapnik, *Statistical Learning Theory*, J. Wiley, 1998.

[8] O. Chapelle, V. Vapnik, and Y. Bengio, Model selection for small sample regression, *Machine Learning*, 48 (2002), 315-333.

[9] V. Cherkassky, X. Shao, F. Mulier, and V. Vapnik, Model complexity control for regression using VC generalization bounds, *IEEE transactions on Neural Networks*, 10 (1999),1075–1089.

[10] G. Lorentz, *Approximation of functions*, Chelsea Publishing Company, New York, 1986.

[11] D. Jackson, On the approximation by trigonometric sums and polynomials, *Transactions of the American Mathematical society*, 13 (1912), 491–515.

[12] D. Jackson, The Theory of Approximation, *Amererican Mathematical Society Colloquium Publications*, XI, New York, 1930.

[13] A. Timan, *Theory of Approximation of Functions of a Real Variable*, English translation 1963, Pergaman Press, Russian original published in Moscow by Fizmatgiz in 1960.

[14] W. Hoeffding, Probability Inequalities for Sums of Bounded Random Variables, *Journal of the American Statistical Association*, 58 (1963), 13-30.

[15] G. Anastassiou and S. Gal, *Approximation Theory: Moduli of continuity and global smoothness preservation*, Birkhäuser, 2000.