Journal of Machine Learning Research 9 (2008) 2607-2633 Submitted 1/07; Revised 8/08; Published 11/08

Model Selection for Regression with Continuous Kernel Functions Using the Modulus of Continuity

Imhoi Koo IMHOI.KOO@KIOM.RE.KR
Department of Medical Research
Korea Institute of Oriental Medicine
Daejeon 305-811, Korea

Rhee Man Kil RMKIL@KAIST.AC.KR
Department of Mathematical Sciences
Korea Advanced Institute of Science and Technology
Daejeon 305-701, Korea

Editors: Isabelle Guyon and Amir Saffari

Abstract

This paper presents a new method of model selection for regression problems using the modulus of continuity. For this purpose, we suggest the prediction risk bounds of regression models using the modulus of continuity which can be interpreted as the complexity of functions. We also present the model selection criterion referred to as the modulus of continuity information criterion (MCIC) which is derived from the suggested prediction risk bounds. The suggested MCIC provides a risk estimate using the modulus of continuity for a trained regression model (or an estimation function) while other model selection criteria such as the AIC and BIC use structural information such as the number of training parameters. As a result, the suggested MCIC is able to discriminate the performances of trained regression models, even with the same structure of training models. To show the effectiveness of the proposed method, the simulation for function approximation using the multilayer perceptrons (MLPs) was conducted. Through the simulation for function approximation, it was demonstrated that the suggested MCIC provides a good selection tool for nonlinear regression models, even with the limited size of data.

Keywords: regression models, multilayer perceptrons, model selection, information criteria, modulus of continuity

1. Introduction

The task of learning from data is to minimize the expected risk (or generalization error) of a regression model (or an estimation function) under the constraint of the absence of *a priori* model of data generation and with the limited size of data. For this learning task, it is necessary to consider nonparametric regression models such as artificial neural networks, as the functional form of the target function is usually unknown. Furthermore, a mechanism to minimize the expected risk from the limited size of data is required. In this context, the model selection is an important issue in the selection of a reasonable network size in order to minimize the expected risk. However, the proper network size (or number of parameters) of a regression model is difficult to choose, as it is possible to obtain the empirical risk only in the case of the limited size of data while the expected risk of a regression model should be measured for the entire data distribution. For the expected risk of a regression model, the loss function of the error square is usually measured, and the expectation of the loss function for the entire data distribution is considered. This expected risk can be decomposed by the bias and variance terms of the regression models. If the number of parameters is increased, the bias term is decreased while the variance term is increased, and the opposite also applies. If the number of parameters is exceedingly small and the performance is thus not optimal due to a large bias term, a situation known as under-fitting of the regression models arises. If the number of parameters is especially large and the performance is thus not optimal due to a large variance term, over-fitting of the regression models arises. Hence, a trade-off exists between the under-fitting and over-fitting of regression models. Here, an important issue is measuring the model complexity associated with the variance term. Related to this issue, the statistical methods of model selection use a penalty term for the measurement of model complexity. Well known criteria using this penalty term are the Akaike information criterion (AIC) (Akaike, 1973), the Bayesian information criterion (BIC) (Schwartz, 1978), the generalized cross-validation (GCV) (Wahba et al., 1979), the minimum description length (MDL) (Rissanen, 1986; Barron et al., 1998), and the risk inflation criterion (RIC) (Foster and George, 1994). These methods can be well fitted with linear regression models when enough samples are available. However, they suffer the difficulty of selecting the optimal structure of the estimation networks in the case of nonlinear regression models and/or a small number of samples. For more general forms of regression models, Vapnik (1998) proposed a model selection method based on the structural risk minimization (SRM) principle. One of the characteristics of this method is that the model complexity is described by structural information such as the VC dimension of the hypothesis space associated with estimation networks, which indicates the number of samples that can be shattered, in other words, which can be completely classified by the given structure of estimation networks. This method can be applied to nonlinear models and also regression models trained for a small number of samples. For this problem, Chapelle et al. (2002), Cherkassky (1999), and Cherkassky and Ma (2003) showed that the SRM-based model selection is able to outperform other statistical methods such as AIC or BIC in regression problems with the limited size of data. On the other hand, these methods require the actual VC dimension of the hypothesis space associated with the estimation functions, which is usually not easy to determine in the case of nonlinear regression models. In this context, we consider the bounds on the expected risks using the modulus of continuity representing a measure of the continuity for the given function. Lorentz (1986) applied the modulus of continuity to function approximation theories. In the proposed method, this measure is applied to determine the bounds on the prediction risk. To be exact, it seeks the expected risk of an estimation function when predicting new observations. To describe these bounds, the modulus of continuity is analyzed for both the target and estimation functions, and the model selection criterion referred to as the modulus of continuity information criterion (MCIC) is derived from the prediction risk bounds in order to select the optimal structure of regression models. One of the characteristics in the suggested MCIC is that it can be estimated directly from the given samples and a trained estimation function. Through the simulation for function approximation using multi-layer perceptrons (MLPs), it is demonstrated that the suggested MCIC is effective for nonlinear model selection problems, even with the limited size of data.

This paper is organized as follows: in Section 2, we introduce the model selection criteria based on statistics such as the AIC and BIC, the model selection criteria based on Shannon's information theory such as the MDL, and the VC dimension based criteria. Section 3 describes the suggested model selection method referred to as the MCIC method starting from the definition of the modulus of continuity for continuous functions. We also describe how we can estimate the modulus of continuity for the regression models with different type of kernel functions. Section 4 describes the simulation results for regression problems for various benchmark data using model selection methods including the suggested MCIC method. Finally, Section 5 presents the conclusion.

2. Model Selection Criteria for Regression Models

For the selection of regression models, the proper criteria for the decision methods are required. Here, various criteria used for the selection of regression models are described. First, let us consider a regression problem of estimating a continuous function f in $C(X,\mathbb{R})$ where $X \subset \mathbb{R}^m$ ($m \geqslant 1$) and $C(X,\mathbb{R})$ is a class of continuous functions. The observed output y for $\mathbf{x} \in X$ can be represented by

$$y(\mathbf{x}) = f(\mathbf{x}) + \varepsilon, \tag{1}$$

where $f(\mathbf{x})$ represents the target function and ε represents random noise with a mean of zero and a variance of σ_ε^2. Here, for regression problems, a data set $D = \{(\mathbf{x}_i, y_i) |\, i = 1, \cdots, N\}$, where $(\mathbf{x}_i, y_i)$ represents the ith pair of input and output samples, is considered. It is assumed that these pairs of input and output samples are randomly generated according to the distribution $P(\mathbf{x})$, $\mathbf{x} \in X$; that is,

$$y_i = f(\mathbf{x}_i) + \varepsilon_i,\ \ \mathbf{x}_i \in X, \tag{2}$$

where ε_i, $i = 1, \cdots, N$ represent independent and identically distributed (*i.i.d.*) random variables having the same distribution with ε. For these samples, our goal of learning is to construct an estimation function $f_n(\mathbf{x}) \in F_n$ (the function space with n parameters) that minimizes the expected risk

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

with respect to the number of parameters n, where $L(y, f_n(\mathbf{x}))$ is a given loss functional, usually the square loss function $L(y, f_n(\mathbf{x})) = (y - f_n(\mathbf{x}))^2$ for regression problems. In general, an estimation function f_n can be constructed as a linear combination of kernel functions; that is,

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

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

To minimize the expected risk (3), it is necessary to identify the distribution $P(\mathbf{x}, y)$; however, this is usually unknown. Rather, we usually find f_n by minimizing the empirical risk $R_{emp}(f_n)$ evaluated by the mean of loss function values for the given samples; that is,

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

Here, if the number of parameters n is increased, the empirical risk of (5) is decreased so that the bias of the estimation function is decreased while the variance of the estimation function is increased, or vice versa. Therefore, a reasonable trade-off must be made between the bias and variance in order to minimize the expected risk. One way of solving this problem is to estimate the expected risks for the given parameters of regression models. In statistical regression models, a popular criterion is the Akaike information criterion (AIC), in which an estimate of the expected risk is given by

$$\mathrm{AIC}(f_n) = R_{emp}(f_n) + 2 \cdot \frac{n}{N}\sigma_\varepsilon^2 \tag{6}$$

under the assumption that the noise term ε has a normal distribution. Here, the noise term $\hat{\sigma}_\varepsilon^2$ can be estimated by

$$\hat{\sigma}_\varepsilon^2 = \frac{RSS}{N - DoF(f_n)}, \tag{7}$$

where RSS represents the sum of the square error over the training samples; that is, $RSS = NR_{emp}(f_n)$, and $DoF(f_n)$ represents the degree of freedom of an estimation function f_n. This criterion is derived in the sense of the maximum-likelihood estimate of the regression parameters. As an alternative to this criterion, the Bayesian approach to model selection referred to as the Bayesian information criterion (BIC) can be considered:

$$\mathrm{BIC}(f_n) = R_{emp}(f_n) + \log N \cdot \frac{n}{N}\sigma_\varepsilon^2. \tag{8}$$

Compared to the AIC, the BIC treats complex models more heavily, giving preference to simpler models, when $N > e^2$, in which e represents the base of natural logarithms. Here, it is important to note that in both criteria, prior knowledge of the variance of noise term σ_ε^2 is needed or estimation of this term using (7) is required. These criteria are good for linear regression models with a large number of samples, as the AIC and BIC formulas hold asymptotically as the number of samples N goes to infinity.

As an alternative to the AIC or BIC, a frequently used model selection criterion is the minimum description length (MDL) criterion. In this method, for the regression model f_n and the data D, the description length $l(f_n, D)$ is described by

$$l(f_n, D) = l(D|f_n) + l(f_n),$$

where $l(f_n)$ represents the length of the regression model and $l(D|f_n)$ represents the length of the data given the regression model. According to Shannon's information theory, the description length in number of bits is then described by

$$l(f_n, D) = -\log_2 p(D|f_n) - \log_2 p(f_n),$$

where $p(f_n|D)$ represents the probability of the output data given the regression model and $p(f_n)$ represents a priori model probability. For a priori model probability, Hinton and Camp (1993) used a zero-mean Gaussian distribution for the neural network parameters. With the additional assumption that the errors of the regression model are $i.i.d.$ with a normal distribution, the description length of the regression model (Cohen and Intrator, 2004) can be described by

$$\mathrm{MDL}(f_n) = \log R_{emp}(f_n) + \frac{n}{N}\left(\log(2\pi) + \log(\frac{1}{n}\sum_{k=1}^{n} w_k^2) + 1\right). \tag{9}$$

This formula for the MDL shows that the description length of the regression model is composed of the empirical risk and the complexity term, which is mainly dependent upon the ratio of the number of parameters to the number of samples and the mean square of weight values. Here, minimizing the description length is equivalent to maximizing the posterior probability of the regression model. Hence the MDL method can be considered as another interpretation of the BIC method. In this context, a regression model that minimizes the description length should be chosen.

A good measure of model complexity in nonlinear models is the VC dimension (Vapnik, 1998) of the hypothesis space associated with estimation networks. The VC dimension can represent the capacity (or complexity) of the estimation network in terms of the number of samples; that is, the maximum number of samples which can be shattered (classified in every possible way) by the estimation network. As the hypothesis space increases, the empirical risk can be decreased but the confidence interval associated with the complexity of the estimation network then increases. From this point of view, it is necessary to make a proper trade-off between the empirical risk and the confidence interval. 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, for the VC dimension h_n measured for the hypothesis space F_n of regression models and the confidence parameter δ (a constant between 0 and 1), the expected risk satisfies the following inequality with a probability of at least $1-\delta$ (Vapnik, 1998; Cherkassky, 1999; Cherkassky and Ma, 2003):

$$R(f_n) \leqslant R_{emp}(f_n)\left(1 - c\sqrt{\frac{h_n(1+\ln(N/h_n)) - \ln\delta}{N}}\right)_+^{-1}, \tag{10}$$

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

If the basis functions $\{\phi_1(\mathbf{x}),\cdots,\phi_n(\mathbf{x})\}$ are orthogonal with respect to the probability measure $P(\mathbf{x})$, the form of (10) can be described in a way that is easier to calculate. For the experimental set up, Chapelle et al. (2002) suggested the following bound with the confidence parameter $\delta = 0.1$:

$$R(f_n) \leqslant R_{emp}(f_n) T_{SEB}(n,N), \tag{11}$$

where

$$T_{SEB}(n,l) = \frac{1+n/(NK)}{1-(n/N)} \quad \text{and}$$

$$K = \left(1 - \sqrt{\frac{n(1+\ln(2N/n))+4}{N}}\right)_+ .$$

These risk estimates of (10) and (11) were successfully applied to the model selection of regression problems with the limited size of data. In these risk estimates, the VC dimension of regression models should be estimated. For the case of nonlinear regression models such as artificial neural networks, the bounds on VC dimensions (Karpinski and Macintyre, 1995; Sakurai, 1995) can be determined. However, in general, it is difficult to estimate the VC dimension of nonlinear regression models accurately.

In this work, we consider a useful method for the selection of nonlinear regression models with the limited size of data. For this problem, the AIC or BIC method may not be effective in view of the fact that the number of samples may not be large enough to apply the AIC or BIC method. Moreover, an estimation of the VC dimension of nonlinear regression models is generally not straightforward. In this context, we consider to use the modulus of continuity representing a measure of continuity for the given function. In the proposed method, this measure is applied to determine the bounds on the prediction risk; that is, the expected risk of the estimation function when predicting new observations. From this result, a model selection criterion referred to as the modulus of continuity information criterion (MCIC) is suggested and it is applied to the selection of nonlinear regression models. The backgrounds and theories related to the suggested method are described in the next section.

3. Model Selection Criteria Based on the Modulus of Continuity

For the description of the bounds on expected risks, the modulus of continuity defined for continuous functions is used. In this section, starting from the definition of the modulus of continuity, the bounds on expected risks are described and the model selection criterion referred to as the MCIC using the described bounds is suggested.

3.1 The Modulus of Continuity for Continuous Functions

The modulus of continuity is a measure of continuity for continuous functions. First, it is assumed that X is a compact subset of Euclidean space $\mathbb{R}^m$; that is, the set X is bounded and closed in Euclidean space $\mathbb{R}^m$. Here, let us consider the case of univariate functions; that is, $m = 1$. Then, the measure of continuity $w(f,h)$ of a function $f \in C(X)$ can be described by the following form (Lorentz, 1986):

$$\omega(f,h) = \max_{x,x+t \in X, |t| \leqslant h} |f(x+t) - f(x)|, \tag{12}$$

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

- $\omega(f,h)$ is continuous at h for each f,
- $\omega(f,h)$ is positive and increases as h increases, 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 positive constants h_1 and h_2.

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

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

$$\omega(f_1+f_2,h) \leqslant \omega(f_1,h) + \omega(f_2,h) \quad \text{for } f_1 \text{ and } f_2 \in C(X).$$

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

$$|f(a_1) - f(a_2)| \leqslant M|a_1 - a_2|^{\alpha}, \ \ a_1, a_2 \in A.$$

In this case, the modulus of continuity is given by

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

In the multi-dimensional input spaces; that is, $X \subset \mathbb{R}^m$ ($m > 1$), there are different definitions of the modulus of continuity for a continuous function f. The following two definitions of the modulus of continuity are considered (Lorentz, 1986; Anastassiou and Gal, 2000):
Definition 1 *Let $m = 2$ and $X \subset \mathbb{R}^m$.*

- *Then, the modulus of continuity for $f \in C(X)$ is defined by*

$$\omega^A(f,h) = \sup\{|f(x_1,y_1) - f(x_2,y_2)|\}$$
$$\text{subject to} \begin{cases} (x_1,y_1),(x_2,y_2) \in X \ \text{ and} \\ \|(x_1,y_1)-(x_2,y_2)\|_2 \leqslant h, \ \text{ for } h > 0. \end{cases}$$

- *Another definition of the modulus of continuity is*

$$\omega^B(f,\alpha,\beta) = \sup\left\{ \begin{array}{l} |f(x_1,y)-f(x_2,y)|, \\ |f(x,y_1)-f(x,y_2)| \end{array} \right\} \tag{13}$$
$$\text{subject to} \left\{ \begin{array}{l} (x_1,y),(x_2,y),(x,y_1),(x,y_2) \in X \text{ and} \\ |x_1-x_2| \leqslant \alpha, |y_1-y_2| \leqslant \beta, \text{for } \alpha,\beta > 0. \end{array} \right.$$

For $f \in C(X)$ on a compact subset $X \subset \mathbb{R}^m$, where $m > 2$, it is possible to define the modulus of continuity by induction.

The main difference in these two definitions of the modulus of continuity is the direction. The first definition measures the variation of all directions at some point $\mathbf{x} \in X$ while the second is dependent upon axis directions only at some point $\mathbf{x} \in X$. The relationship between the two definitions of the modulus of continuity can be described by the following lemma:

Lemma 1 *For $f \in C(X)$, two definitions of the modulus of continuity, $\omega^A(f,h)$ and $\omega^B(f,h,h)$ have the following relationship:*

$$\omega^B(f,h,h) \leqslant \omega^A(f,h) \leqslant 2\omega^B(f,h,h),$$

where h represents a positive constant.

For the proof of this lemma, refer to the Appendix A.1. Furthermore, each definition of the modulus of continuity has the following upper bound:

Lemma 2 *Let $f \in C^1(X)$, the class of continuous functions having continuous 1st derivative on X, a compact subset of $\mathbb{R}^m, m > 1$. Then, for $h > 0$, the modulus of continuity w^A and w^B have the following upper bounds:*

$$\omega^A(f,h) \leqslant h\sqrt{\sum_{i=1}^{m} \left\| \frac{\partial f}{\partial x_i} \right\|_\infty^2} \quad \textit{and}$$

$$\omega^B(f,h,\cdots,h) \leqslant h \max_{1\leqslant i \leqslant m} \left\{ \left\| \frac{\partial f}{\partial x_i} \right\|_\infty \right\},$$

where x_i represents the ith coordinate in the point $\mathbf{x} = (x_1, x_2, \cdots, x_m) \in X$ and $\|\cdot\|_\infty$ represents the supremum norm (or L_∞ norm); that is, for a real- or complex-valued bounded function $g(\mathbf{x})$,

$$\|g\|_\infty = \sup\{|g(\mathbf{x})| \mid \mathbf{x} \in X_g\},$$

where X_g represents the domain of g.

For the proof of this lemma, refer to the Appendix A.2. From this lemma, the second definition of the modulus of continuity $w^B(f,h,h)$ was chosen because it has a smaller upper bound compared to the first modulus of continuity. For our convenience, the notation $w(f,h)$ is used to represent $w^B(f,h,\cdots,h)$ in the remaining sections of this paper.

The computation of the modulus of continuity requires the value of h. First, let us consider the following definition of a density of input samples (Tinman, 1963):

Definition 2 *The density D of an input sample set $\{\mathbf{x}_1, \cdots, \mathbf{x}_N\} \subset X$, a compact subset of $\mathbb{R}^m, m > 1$, is defined by*

$$D(\mathbf{x}_1,\cdots,\mathbf{x}_N) = \sup_{\mathbf{x}\in X} \inf_{1\leqslant i \leqslant N} d(\mathbf{x}_i,\mathbf{x}),$$

where $d(\mathbf{x}_i, \mathbf{x})$ represents the distance between $\mathbf{x}_i$ and $\mathbf{x} \in X$, which is explicitly any metric function such that, for every $\mathbf{x}, \mathbf{y}, \mathbf{z} \in X$, the following properties are satisfied: $d(\mathbf{x}, \mathbf{y}) \geqslant 0$ with the equality if and only if $\mathbf{x} = \mathbf{y}$, $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$, and $d(\mathbf{x}, \mathbf{z}) \leqslant d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z})$.

Let us also consider a point $\mathbf{x}_0 \in X$ such that

$$\mathbf{x}_0 = \arg\max_{\mathbf{x} \in X} |f(\mathbf{x}) - f_n(\mathbf{x})|.$$

Then, the value of h can be bounded by

$$\min_{1 \leqslant i \leqslant N} d(\mathbf{x}_i, \mathbf{x}_0) \leqslant h \leqslant D(\mathbf{x}_1, \cdots, \mathbf{x}_N) \tag{14}$$

to cover the input space X using balls $B(\mathbf{x}_i, h)$ with centers as input samples $\mathbf{x}_i$ and a radius of h: $B(\mathbf{x}_i, h) = \{\mathbf{x} | \, \|\mathbf{x}_i - \mathbf{x}\| < h\}$. This range of h is considered to describe the modulus of continuity for the target and estimation functions.

3.2 Risk Bounds Based on the Modulus of Continuity

In this subsection, the modulus of continuity for the target and estimation functions are investigated, and the manner in which they are related to the expected risks is considered. First, let us consider the loss function for the observed model y and the estimation function $f_n(\mathbf{x})$ with n parameters as $L(y, f_n) = |y - f_n(\mathbf{x})|$. Then, the expected and the empirical risks are defined by the following L_1 measure:

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

$$R_{emp}(f_n)_{L_1} = \frac{1}{N} \sum_{i=1}^{N} |y_i - f_n(\mathbf{x}_i)|.$$

In the first step, let us consider the case of a univariate target function; that is, $f \in C(X)$ with $X \subset \mathbb{R}$. Then, with the definition of the modulus of continuity of (12) and the bound of h as described by (14), the relationship between the expected and empirical risks is described using the modulus of continuity as follows:

Theorem 1 *Let the target function $f \in C^1(X)$ of (1) with X, a compact subset of $\mathbb{R}$, be approximated by the estimation function f_n of (4), that is, a linear combination of weight parameters w_k 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, for the confidence parameters δ (a constant between 0 and 1), the expected risk in the L_1 sense is bounded by the following inequality with a probability of at least $1 - 2\delta$:*

$$\begin{aligned} R(f_n)_{L_1} \leqslant R_{emp}(f_n)_{L_1} &+ \frac{1}{N^2} \sum_{i,j=1}^{N} (|y_i - y_j| + |f_n(x_i) - f_n(x_j)|) \\ &+ (\omega(f_n, h_0) + C) \sqrt{\frac{1}{2N} \ln \frac{2}{\delta}} \quad \textit{and} \end{aligned} \tag{15}$$

$$C = |f_n(x_0) - f_n(x_0^{'})| + 2\|f\|_\infty + 2\sigma_\varepsilon \sqrt{\frac{1}{\delta}} \;\; \textit{for} \; x_0, x_0^{'} \in \{x_1, \cdots, x_N\},$$

where $w(f_n, h_0)$ represents the modulus of continuity of the estimation function f_n and h_0 represents a constant satisfying (14).

For the proof of this theorem, refer to the Appendix A.3. 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 (15) representing the variations of output samples and also the variations of estimation function values for the given input samples, and the third term representing the modulus of continuity for the estimation function $w(f_n,h_0)$ and a constant C associated with target function. Here, let us consider the second term. By investigating this term further, it can be decomposed it into the empirical risk and the term depending on the target function. The next corollary shows the bounds on the expected risks with 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 a probability of at least $1-2\delta$:*

$$R(f_n)_{L_1} \leqslant 3R_{emp}(f_n)_{L_1} + \frac{2}{N}\max\{\lambda_i\} + (\omega(f_n,h_0)+C)\sqrt{\frac{1}{2N}\ln\frac{2}{\delta}}, \tag{16}$$

where λ_i represents the ith eigenvalue of the matrix H_y.

For the proof of this lemma, refer to the Appendix A.4. This corollary states that the dominant terms related to the estimation function f_n in the expected risk are the empirical risk $R_{emp}(f_n)$ and the modulus of continuity $w(f_n,h_0)$, as the eigenvalue λ_i of H_f is not dependent upon f_n and a constant C has little influence on the shape of expected risks as the number of parameters n increases. The bounds on the expected risks of (16) appear to be overestimated, as the empirical risk is multiplied by 3. However, for the purpose of determining the model selection criteria, an 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 problems because only these two terms are mainly dependent upon the estimation function f_n. From this point of view, the following model selection criterion referred to as the modulus of continuity information criterion (MCIC) is suggested:

$$\text{MCIC}(f_n) = R_{emp}(f_n)_{L_1} + \frac{\omega(f_n,h_0)}{3}\sqrt{\frac{1}{2N}\ln\frac{2}{\delta}}. \tag{17}$$

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, as the estimation function f_n becomes a more complex function. Accordingly, it is necessary to make a trade-off between the over-fitting and under-fitting of regression models using the MCIC for the optimization of the regression models.

This far, univariate continuous functions are addressed. At this point, let us consider the case of $X \subset \mathbb{R}^m$ with $m > 1$; that is, the case of multivariate continuous functions. Here, it is possible to show that the prediction risk bounds take a similar form to those of univariate continuous functions. The following theorem of the prediction risk bounds for multivariate continuous functions is suggested using the definition of the modulus of continuity (13):

Theorem 2 *Let $f \in C^1(X)$ with X, a compact subset of $\mathbb{R}^m$ ($m > 1$), and h_0 be a constant satisfying (14). Then, for the confidence parameter δ (a constant between 0 and 1), the expected risk in L_1 sense is bounded by the following inequality with a probability of at least $1-\delta$:*

$$R(f_n)_{L_1} \leqslant R_{emp}(f_n)_{L_1} + \left\{\omega(f-f_n,h_0) + |f(\mathbf{x}_{i_0}) - f_n(\mathbf{x}_{i_0})| + \sigma_\varepsilon\sqrt{\frac{2}{\delta}}\right\}\sqrt{\frac{1}{2N}\ln\frac{2}{\delta}},$$

where $w(f-f_n,h_0)$ represents the modulus of continuity of the function $f-f_n$, h_0 represents a constant satisfying (14), and $\mathbf{x}_{i_0}$ represents an element of an input sample set $\{\mathbf{x}_1,\cdots,\mathbf{x}_N\}$.

For the proof of this theorem, refer to the Appendix B.1. In this theorem, $w(f-f_n,h_0)$ can be replaced with $w(f,h_0)+w(f_n,h_0)$; that is, the sum of the modulus of continuity for the target and estimation functions because the following inequalities always hold:

$$\omega(f_n,h_0)-\omega(f,h_0)\leqslant\omega(f-f_n,h_0)\leqslant\omega(f_n,h_0)+\omega(f,h_0).$$

The suggested theorem states that the expected risk is mainly bounded by the empirical risk $R_{emp}(f_n)$, the modulus of continuity for the target function $w(f,h_0)$, and also the modulus of continuity for the estimation function $w(f_n,h_0)$. As the number of parameters n varies, the empirical risk, the modulus of continuity for the estimation function, and the term $|f(\mathbf{x}_{i_0})-f_n(\mathbf{x}_{i_0})|$ are changed while other terms remain constant. Here, in order to find the optimal model complexity $n=n^*$ that minimizes expected risk $R(f_n)$, these varying terms should be considered. In this case, the effect of the term $|f(\mathbf{x}_{i_0})-f_n(\mathbf{x}_{i_0})|$ is small compared with the other two terms, as the regression model becomes well fitted to the samples as the number of parameters n increases. This implies that the model selection criteria for multivariate estimation functions have the same form as (17) except with a coefficient $1/3$ of $\omega(f_n,h_0)$. In practice, the performance of MCIC for model selection problems is not so sensitive to this coefficient. Summarizing the properties of the suggested MCIC, the distinct characteristics are described as follows:

- The suggested MCIC is dependent upon the modulus of continuity for the trained estimation function.
- The suggested MCIC is also dependent upon the value of h_0 which varies according to the sample distribution.

Considering these characteristics, for model selection problems, the MCIC is a measure sensitive to the trained estimation function using a certain learning algorithm and also sensitive to the distribution of samples while other model selection criteria such as the AIC and BIC depend on structural information such as the number of parameters. For the computation of the suggested MCIC, the modulus of continuity of the trained estimation function should be evaluated, as explained in the next subsection.

3.3 The Modulus of Continuity for Estimation Functions

The modulus of continuity for the estimation function $w(f_n,h)$ is dependent upon the basis function ϕ_k in (4). Here, examples of computing $w(f_n,h)$ according to the type of basis functions are presented:

- A case of the estimation function f_n with algebraic polynomials on $X=[a,b]\subset\mathbb{R}$:

$$\phi_k(x)=x^k\quad\text{for } k=0,1,\cdots,n.$$

Applying the mean value theorem to ϕ_k, we get

$$\begin{aligned}\omega(\phi_k,h)&\leqslant\left\|\phi_k'\right\|_\infty\cdot h\\&\leqslant kh\cdot\max\left\{|a|^{k-1},|b|^{k-1}\right\},\ k=1,\cdots,n.\end{aligned}$$

Therefore, the modulus of continuity for f_n has the following upper bound:

$$\omega(f_n,h) \leqslant \sum_{k=1}^{n} kh|w_k| \cdot \max\left\{|a|^{k-1}, |b|^{k-1}\right\}.$$

- A case of the estimation function f_n with trigonometric polynomials $\phi_k(x)$ on $X \subset \mathbb{R}$:

$$\phi_k(x) = \begin{cases} 1/2 & \text{if } k=0 \\ \sin((k+1)x/2) & \text{if } k = \text{odd number} \\ \cos(kx/2) & \text{if } k = \text{even number} \end{cases}$$

for $k = 1, \cdots, n$. Applying the mean value theorem to ϕ_k, we get

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

Therefore, the modulus of continuity for f_n has the following upper bound:

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

- A case of the estimation function f_n with sigmoid function $\phi_k(\mathbf{x}) = \phi_k(x_1, \cdots, x_m)$ on $X \subset \mathbb{R}^m$:

$$f_n(\mathbf{x}) = \sum_{k=1}^{n} w_k \phi_k(x_1, \cdots, x_m) + w_0,$$

where

$$\phi_k(x_1, \cdots, x_m) = \tanh\left(\sum_{j=1}^{m} v_{kj} x_j + v_{k0}\right).$$

Applying the mean value theorem to ϕ_k with respect to each coordinate $x_1, \cdots, x_m$, we get

$$|f_n(\cdots, x_j, \cdots) - f_n(\cdots, x_j - h, \cdots)| \leqslant h \cdot \left\| \frac{\partial f_n}{\partial x_j} \right\|_\infty \text{ for } j = 1, \cdots, m.$$

Therefore, the modulus of continuity for f_n has the following upper bound:

$$\begin{aligned} \omega(f_n,h) &\leqslant h \cdot \max_{1 \leqslant j \leqslant m} \left\| \frac{\partial f_n}{\partial x_j} \right\|_\infty \\ &\leqslant h \cdot \max_{1 \leqslant j \leqslant m} \left\| \sum_{k=1}^{n} w_k v_{kj} \cdot \left(1 - \tanh^2\left(\sum_{i=1}^{m} v_{ki} x_i + v_{k0}\right)\right) \right\|_\infty \\ &\leqslant h \cdot \max_{1 \leqslant j \leqslant m} \sum_{k=1}^{n} \left| w_k v_{kj} \right|. \end{aligned} \tag{18}$$

As shown in these examples, the modulus of continuity for the estimation function f_n is dependent upon the trained parameter values associated with f_n and h_0 whose range is given by (14). However, the proper value of h_0 satisfying (14) requires the intensive search of the input space. From this point of view, in practice, the value of h_0 is considered as the half of the average distance between two adjacent samples. Assuming a uniform distribution of input samples, the value of h_0 can be determined from the range of data values in each coordinate. For example, for m dimensional input patterns, h_0 can be determined by

$$h_0 = \frac{1}{2}\left(\frac{1}{m}\sum_{i=1}^{m}\frac{\max_i - \min_i}{N-1}\right)^{1/m}, \tag{19}$$

where $\max_i$ and $\min_i$ represent the maximum and minimum values of the samples in the ith coordinate.

After the value of h_0 is determined, the computation of the modulus of continuity requires access to all the parameter values of f_n which are obtained after the learning of training samples. In this context, the computational complexity of the modulus of continuity is proportional to the number of parameters n in the estimation function, that is, in big-O notation, $O(n)$. This computational complexity is not so heavy compared to the calculation of the empirical risk term, as it requires the computational complexity of $O(N)$ and in general, $N \gg n$. Hence, in total, the computational complexity of MCIC is described by $O(N)$ which is equivalent to the computational complexity of the AIC or BIC.

Once the modulus of continuity is determined, the MCIC can be determined by (17). Then, the model with the smallest value of MCIC will then be selected. Here, $\widehat{n}$ is selected such that

$$\widehat{n} = \arg\min_n \mathrm{MCIC}(f_n).$$

The validity of the suggested MCIC is shown in the next section through the simulation for nonlinear model selection problems.

4. Simulation

The simulation for function approximation was performed using the multilayer perceptrons (MLPs) composed of the input, hidden, and output layers. For this simulation, the number of sigmoid units n in the hidden layer was increased from 1 to 50. Here, f_n was denoted as the MLP with a hidden layer including n sigmoid units with the m dimensional input. The functional form of the estimation function is given by

$$f_n(x) = \sum_{k=1}^{n} w_k \tanh(\sum_{j=1}^{m} v_{kj}x_j + v_{k0}) + w_0,$$

where v_{kj} and w_k represent the input and output weights, respectively, that are associated with the kth sigmoid unit, and v_{k0} and w_0 represent the bias terms of the kth sigmoid unit and of the estimation function, respectively. In this regression model, the conjugate gradient method was used for the training of the input and output weights of the MLPs. As for the different type of kernel functions, we presented the model selection method using the suggested MCIC for the regression model with trigonometric polynomials (Koo and Kil, 2006) and showed the effectiveness of the MCIC compared to other model selection criteria.

As the benchmark data for this simulation of function approximation, the target functions given by Donoho and Johnstone (1995) were used: they are Blocks, Bumps, Heavysine, and Doppler functions, as illustrated in Figure 1. To generate the data for each target function from D-J (Donoho and Johnstone), the input values x_i, $i = 1, \cdots, N$ were generated from a uniform distribution within the interval of $[0, 2\pi]$. Here, for the normalization of D-J data, the outputs were adjusted to have the mean square value of 1 within the interval of $[0, 2\pi]$. The noise terms were also generated from a normal distribution with a mean of zero and a standard deviation of $\sigma_\varepsilon = 0.2$ or 0.4. They were then added to the target values computed from the randomly generated input values. For these target functions, 100 sets of N (= 200) training samples were generated randomly to train the MLP. In addition to the training samples, 1000 test samples were also generated separately according to identical input and noise distributions.

As for another application of the MCIC, simulations for the target functions with binary output values were considered using the benchmark data suggested by Hastie et al. (2003): the target value is defined by

$$y(\mathbf{x}) = \begin{cases} 1 & \text{if} \sum_{j=1}^{10} x_j > 5 \\ 0 & \text{otherwise,} \end{cases}$$

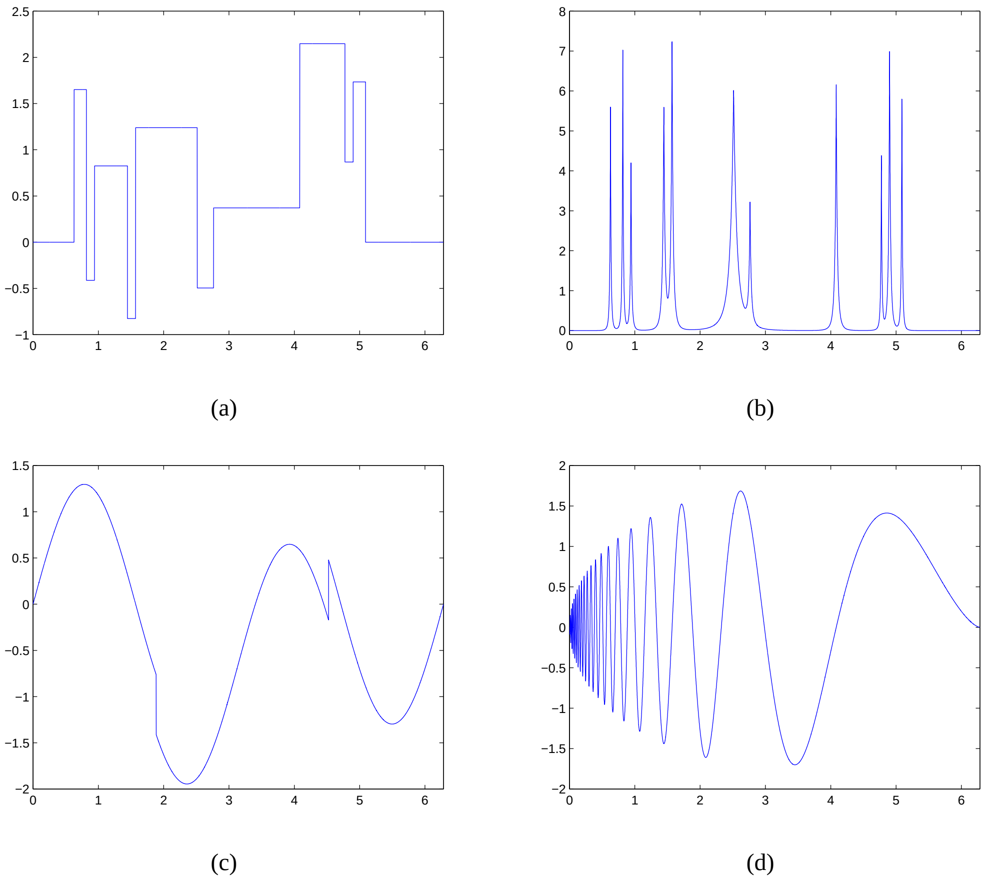
where $\mathbf{x}$ is uniformly generated in $[0, 1]^{20}$. For the training of this target function, 100 sets of 50 samples were generated. In addition to the training samples, 500 test samples were also generated separately. The noise terms were also generated from a normal distribution with a mean of 0 and a standard deviation of $\sigma_\varepsilon = 0.0$ or 0.2. They were then added to the target values computed from the randomly generated input values.

For the simulation of selecting regression models with multi-dimensional input data, the benchmark data suggested by Chang and Lin (2001) were also used: they are Abalone, CPU_Small, MG, and Space_GA data sets as described in Table 1. For each data set, 30 sets of 500 samples were randomly generated as the training samples and the remaining samples in each set were used as the test samples; that is, 30 sets of test samples were also generated. For these data sets, the range of input and output was normalized between -1 and 1.

Data Set	Description	No. of Features	No. of Data
Abalone	predicting the age of abalone	8	4177
CPU_Small	predicting a computer system activity	12	8192
MG	predicting the Mackey-Glass time series	6	1385
Space_GA	election data on 3107 US counties	6	3107

Table 1: The benchmark data sets for regression problems

For our experiments, various model selection methods such as the AIC, BIC, MDL, and the suggested MCIC-based methods were tested. Once the MLP was trained, the empirical risk $R_{emp}(f_n)$ evaluated by the training samples was obtained, and the estimated risk $\widehat{R}(f_n)$ value could then be determined by the AIC, BIC, MDL, and MCIC-based methods. In the cases of the AIC and BIC methods, we selected the estimation function $f_{\widehat{n}}$ which gives the smallest value of information criterion described as (6) and (8), respectively. In these criteria, we assume that the noise variance σ_ε^2 value was known. In the case of MDL, we selected the estimation function $f_{\widehat{n}}$ which gives the smallest value of the description length of (9). In the suggested method, we used the following form



(a)

(b)

(c)

(d)

Figure 1: Target functions from Donoho and Johnstone (1995): (a), (b), (c), and (d) represent the Blocks, Bumps, Heavysine, and Doppler functions respectively.

of MCIC for MLPs using the modulus of continuity described as (18):

$$\mathrm{MCIC}(f_n) = R_{emp}(f_n)_{L_1} + \frac{h_0}{3} \max_{1\leqslant j\leqslant m} \sum_{k=1}^{n} \left|w_k v_{kj}\right| \sqrt{\frac{1}{2N}\ln\frac{2}{\delta}}, \tag{20}$$

where h_0 was set to the half of the average distance between two adjacent samples using (19) and δ was set to 0.05. In our case, we selected $f_{\hat{n}}$ which gave the smallest value of (20).

To compare the performance of the model selection methods, the risks for the selected $f_{\hat{n}}$ were evaluated by the test samples and the results were compared with the minimum risk among all risks for f_n, $n = 1, \cdots, 50$. Quantitatively, the log ratio r_R of two risks $R_{test}(f_n)$ and $\min_n R(f_n)$ were computed:

$$r_R = \log \frac{R_{test}(f_{\hat{n}})}{\min_n R_{test}(f_n)}, \tag{21}$$

where R_{test} represents the empirical risk for the squared error loss function evaluated by the test samples. This risk ratio represents the quality of the estimated distance between the optimal and the estimated optimal risks.

Target	AIC		BIC		MDL		MCIC	
Functions	mean	s. dev.	mean	s. dev.	mean	s. dev.	mean	s. dev.
Blocks	**0.0255**	0.0262	0.0416	0.0411	0.1459	0.0763	0.0371	0.0395
Bumps	**0.0393**	0.0697	0.0507	0.0718	0.1646	0.1294	0.0458	0.0426
Heavysine	0.0420	0.0537	0.1059	0.1147	0.1071	0.1225	**0.0107**	0.0160
Doppler	0.0343	0.0523	0.0932	0.0973	0.2318	0.1345	**0.0222**	0.0359

Table 2: Risk ratios for the regression of the four D-J target functions with $\sigma_\varepsilon = 0.2$.

Target	AIC		BIC		MDL		MCIC	
Functions	mean	s. dev.	mean	s. dev.	mean	s. dev.	mean	s. dev.
Blocks	0.0441	0.0350	0.0853	0.0509	0.1146	0.0563	**0.0262**	0.0302
Bumps	**0.0511**	0.0572	0.0804	0.0699	0.1451	0.0810	0.0516	0.0433
Heavysine	0.0846	0.0616	0.1483	0.0773	0.1458	0.0762	**0.0130**	0.0151
Doppler	0.0801	0.0728	0.1421	0.0860	0.2225	0.1164	**0.0218**	0.0311

Table 3: Risk ratios for the regression of the four D-J target functions with $\sigma_\varepsilon = 0.4$.

After all experiments had been repeated for the given number of training sample sets, the means and standard deviations of the risk ratios of (21) for each target function were presented. First, in the case of D-J data sets, the simulation results of the model selection using the AIC, BIC, MDL, and MCIC based methods are presented in Tables 2 and 3. These simulation results showed that the suggested MCIC method provided the top level performances in all cases except the blocks and bumps target functions when $\sigma_\varepsilon = 0.2$ in which the AIC method showed the best performances. This was mainly due to the fact that the known noise standard deviation of σ_ε was used in the AIC method. To clarify this fact, another simulation for these target functions in which the AIC and BIC methods with the estimation of noise variances using (7) were used. These simulation results are presented in Table 4. In this simulation, as we expected, the suggested MCIC method showed the best performance.

We also observed the dependency of the number of samples during the selection of regression models. For this simulation, the numbers of samples were set to $N = 100$, 200, 400, and 800 for the regression of the Doppler target function with a noise standard deviation of $\sigma_\varepsilon = 0.4$. The simulation results are presented in Table 5. Here, note that the complexity of the doppler target function increases as the input value decreases. These results showed that the performances of the AIC, BIC, MDL, and MCIC methods were improved as the number of samples becomes larger as shown in Table 5. Among these model selection methods, the MCIC method always showed the better performances compared to other model selection methods, even in the smaller numbers of samples. This is mainly due to the fact that in the MCIC method, the modulus of continuity, which can be interpreted as the complexity of the estimation function was computed for each trained estimation function directly.

In the case of Hastie et al.'s benchmark data, the MDL model selection methods showed some merits in performances compared to other model selection methods as shown in Table 6. One of the reasons why the MCIC method does not show the better performances in this target function compared to other model selection methods is that this target function can be properly solved by the classification problem (not regression problem) in which the discriminant function is linear.

Target Functions	AIC		BIC		MDL		MCIC	
	mean	s. dev.	mean	s. dev.	mean	s. dev.	mean	s. dev.
Blocks	0.0718	0.0612	0.1117	0.0744	0.1459	0.0763	**0.0371**	0.0395
Bumps	0.1107	0.0933	0.1558	0.1395	0.1646	0.1294	**0.0458**	0.0426

Table 4: Risk ratios for the regression of blocks and bumps functions with $\sigma_\varepsilon = 0.2$ using the AIC and BIC methods with the estimation of noise variances using (7), and the MDL and MCIC methods.

	AIC		BIC		MDL		MCIC	
N	mean	s. dev.	mean	s. dev.	mean	s. dev.	mean	s. dev.
100	0.1072	0.0820	0.1552	0.1119	0.3205	0.1710	**0.0794**	0.0696
200	0.0801	0.0728	0.1421	0.0860	0.2225	0.1164	**0.0218**	0.0311
400	0.0477	0.0478	0.1028	0.0670	0.1588	0.0759	**0.0077**	0.0135
800	0.0174	0.0275	0.0610	0.0565	0.0860	0.0774	**0.0035**	0.0107

Table 5: The variation of risk ratios for the regression of Doppler function with $\sigma_\varepsilon = 0.4$ using the AIC, BIC, MDL, and MCIC methods according to the number of samples $N = 100$, 200, 400, and 800.

However, even in this case, if the sample size is reduced, the proposed method can have the merits in performances since the MCIC includes the complexity term of the estimation function using the modulus of continuity and for smaller number of samples, this complexity term has the high influence on the bounds on the expected risk. To clarify this fact, we made another simulation results for the number of samples reduced by half; that is, $N = 25$ and compared with the previous simulation results as shown in Table 7. These simulation results showed that the MCIC method demonstrated the better performances compared to other model selection methods by reducing the sample size.

	AIC		BIC		MDL		MCIC	
σ_ε	mean	s. dev.	mean	s. dev.	mean	s. dev.	mean	s. dev.
0.0	0.3161	0.2197	0.3161	0.2197	**0.3100**	0.2273	0.4307	0.2624
0.2	0.3400	0.5028	0.3115	0.4658	**0.1881**	0.1175	0.3034	0.1503

Table 6: Risk ratios for the regression of the binary target function using the AIC, BIC, MDL, and MCIC methods when the number of samples N is 50.

The simulation results for the selection of regression models with multi-dimensional input data are summarized in Table 8. These simulation results showed that the suggested MCIC method achieved top or second level performances compared to other model selection methods. As shown in the previous case; that is, the regression problem of Hastie et al.'s benchmark data, the MCIC method is more effective when the sample size is small. To see the effect on smaller number of samples, we also made another simulation results for the number of samples reduced by half; that is,

	AIC		BIC		MDL		MCIC	
σ_ε	mean	s. dev.	mean	s. dev.	mean	s. dev.	mean	s. dev.
0.0	0.3620	0.1977	0.3620	0.1977	0.3430	0.1855	**0.2652**	0.1589
0.2	1.7428	0.9582	1.7428	0.9582	0.3169	0.1816	**0.2716**	0.1743

Table 7: The variation of risk ratios for the regression of the binary target function when the number of samples is reduced by half; that is, $N = 25$.

$N = 250$ and compared with the previous simulation results as shown in Table 9. These simulation results showed that the MCIC method demonstrated the top level performances compared to other model selection methods. All of these observations support that the suggested MCIC method is quite effective for nonlinear regression models especially for smaller number of samples. This is mainly due to the fact that the complexity term as a form of the modulus of continuity of the trained regression model provides high influence on selecting the regression model especially for smaller number of samples. This can be explained by the following observations:

- Once the estimation function is trained, the estimation function provides accurate values for the training samples. In this estimation function, the variation of the predicted values for the unobserved data with respect to the function values for the known data (or training samples) can be described by the modulus of continuity, as presented in the definition of the modulus of continuity.

- If the number of samples decreases, the density of input space becomes low and it makes a big value of h. Then, in the suggested MCIC, this makes high influence of the modulus of continuity compared to the empirical risk which usually has a small value.

- If there are enough number of samples for the target function, the opposite phenomenon of the above case happens.

In summary, through the simulation for function approximation using the MLPs, we have shown that the suggested MCIC provides performance advantages for the selection of regression models compared to other model selection methods in various situations of benchmark data. Compared to other model selection methods, the MCIC methods provides the considerable merits in performances especially when no knowledge of noise variances for the given samples is available and also when not enough number of samples considering the complexity of target function is available.

5. Conclusion

We have suggested a new method of model selection in regression problems based on the modulus of continuity. The prediction risk bounds are investigated from a view point of the modulus of continuity for the target and estimation functions. We also present the model selection criterion referred to as the MCIC which is derived from the suggested prediction risk bounds. The suggested MCIC is sensitive to the trained regression model (or estimation function) obtained from a specific learning algorithm and is also sensitive to the distribution of samples. As a result, the suggested MCIC is able to discriminate the performances of the trained regression models, even with the same structure of regression models. To verify the validity of the suggested criterion, the selection

	AIC		BIC		MDL		MCIC	
Data Set	mean	s. dev.	mean	s. dev.	mean	s. dev.	mean	s. dev.
Abalone	0.0428	0.0586	0.0496	0.0410	0.0477	0.0493	**0.0324**	0.0303
CPU_Small	0.1646	0.1578	0.1212	0.1158	**0.0940**	0.0941	0.0941	0.1076
MG	0.0665	0.0442	0.0649	0.0371	0.0523	0.0470	**0.0449**	0.0343
Space_GA	**0.0851**	0.0612	0.1597	0.0869	0.1039	0.0626	0.0870	0.0526

Table 8: The variation of risk ratios for the regression of the benchmark data sets using the AIC, BIC, MDL, and MCIC methods when the number of samples N is 500.

	AIC		BIC		MDL		MCIC	
Data Set	mean	s. dev.	mean	s. dev.	mean	s. dev.	mean	s. dev.
Abalone	0.1385	0.1947	0.0668	0.1019	0.0618	0.0966	**0.0307**	0.0418
CPU_Small	0.2832	0.2906	0.2864	0.2646	0.2828	0.2929	**0.1942**	0.2219
MG	0.1451	0.1037	0.0816	0.0947	0.0887	0.0934	**0.0456**	0.0508
Space_GA	0.0768	0.0618	0.1466	0.0872	0.0801	0.0651	**0.0659**	0.0518

Table 9: The variation of risk ratios for the regression of the benchmark data sets when the number of samples is reduced by half; that is, $N = 250$.

of regression models using the MLPs that were applied to function approximation problems was performed. Through the simulation for function approximation using the MLPs, it was shown that the model selection method using the suggested MCIC has the advantages of risk ratio performances over other model selection methods such as the AIC, BIC, and MDL methods in various situations of benchmark data. Compared to other model selection methods, this merit of regression performances is significant especially when not enough number of samples considering the complexity of target function is available. Furthermore, the suggested MCIC method does not require any knowledge of a noise variance of samples which is usually given or estimated in other model selection methods. For regression models with other types of estimation functions that have some smoothness constraints, the suggested MCIC method can be easily extended to the given regression models by evaluating the modulus of continuity for the corresponding estimation functions.

Acknowledgments

The authors would like to thank to the editors and the anonymous reviewers for their helpful comments. This work was partially supported by the Korea Science and Engineering Foundation (KOSEF) grant (M10643020004-07N4302-00400) funded by the Korean Ministry of Education, Science, and Technology.

Appendix A.

In this appendix, we prove the lemmas 1 and 2 in Section 3.1. We also prove the theorem 1 and corollary 1 in Section 3.2; that is, the case of univariate target functions.

A.1 Proof of Lemma 1

Since $\omega^A(f,h)$ are considered all directions on h-ball on X, the following inequality always holds:

$$\omega^B(f,h,h) \leqslant \omega^A(f,h).$$

From the triangular inequality, the following inequality holds:

$$|f(x_1,y_1) - f(x_2,y_2)| \leqslant |f(x_1,y_1) - f(x_1,y_2)| + |f(x_1,y_2) - f(x_2,y_2)|.$$

Let $\|(x_1,y_1) - (x_2,y_2)\| \leqslant h$. Then, $|x_1 - x_2| \leqslant h$ and $|y_1 - y_2| \leqslant h$. Therefore, from the definition of the modulus of continuity, we obtain

$$\omega^A(f,h) \leqslant 2\omega^B(f,h,h).$$

□

A.2 Proof of Lemma 2

- Upper bound of $w^A(f,h)$: Let $\mathbf{x} \in X$ and $\mathbf{x} - \mathbf{h} \in X$ satisfying $\|\mathbf{h}\| \leqslant h$. Then,

$$\begin{aligned} |f(\mathbf{x}) - f(\mathbf{x}-\mathbf{h})| &\leqslant |\nabla f(\mathbf{x} - \xi\mathbf{h}) \cdot \mathbf{h}| \text{ for some } \xi \in (0,1) \\ &\leqslant \|\mathbf{h}\|\,\|\nabla f(\mathbf{x} - \xi\mathbf{h})\| \\ &\quad \text{(because of Cauchy-Schwartz inequality)} \\ &= \|\mathbf{h}\| \sqrt{\sum_{i=1}^{m} \left|\frac{\partial f}{\partial x_i}(\mathbf{x} - \xi\mathbf{h})\right|^2} \\ &\leqslant \|\mathbf{h}\| \sqrt{\sum_{i=1}^{m} \left\|\frac{\partial f}{\partial x_i}\right\|_\infty^2}. \end{aligned}$$

  Since the last term of the above equation is independent of $\mathbf{x} \in X$, we can conclude that

$$\omega^A(f,h) \leqslant h\sqrt{\sum_{i=1}^{m} \left\|\frac{\partial f}{\partial x_i}\right\|_\infty^2}.$$

- Upper bound of $w^B(f,h,h)$: Let $\alpha \in \mathbb{R}$ and $|\alpha| \leqslant h$. Here, let us define $\mathbf{e}_i$ as a vector on $\mathbb{R}^m$ whose i-th coordinate is 1 and the others are 0. Then, there exists $\xi \in (0,1)$ such that

$$|f(\mathbf{x}) - f(\mathbf{x} - h\mathbf{e}_i)| \leqslant \left|\frac{\partial f}{\partial x_i}(\mathbf{x} - \xi\alpha\mathbf{e}_i)\right| |h| \quad \text{for } i = 1, \cdots, m.$$

  This implies that

$$\max_i \{|f(\mathbf{x}) - f(\mathbf{x} - \alpha\mathbf{e}_i)|\} \leqslant |\alpha| \max_{1\leqslant i\leqslant m} \left\{\left|\frac{\partial f}{\partial x_i}(\mathbf{x} - \xi\alpha\mathbf{e}_i)\right|\right\}.$$

  Therefore, we can conclude that

$$\omega^B(f,h,\cdots,h) \leqslant h \max_{1\leqslant i\leqslant m} \left\{\left\|\frac{\partial f}{\partial x_i}\right\|_\infty\right\}.$$

□

A.3 Proof of Theorem 1

Before the description of main proof, let us introduce the Hoeffding inequality (Hoeffding, 1963): 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 $\varepsilon > 0$, we have

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

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

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

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

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

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

$$\omega(f_n, h_i) + |f_n(x_i) - f_n(x_j)| \leqslant \omega(f_n, h_0) + |f_n(x_0) - f_n(x'_0)| \quad \text{for } i,j = 1,\cdots,N.$$

For the illustration of this concept, refer to Figure 2.

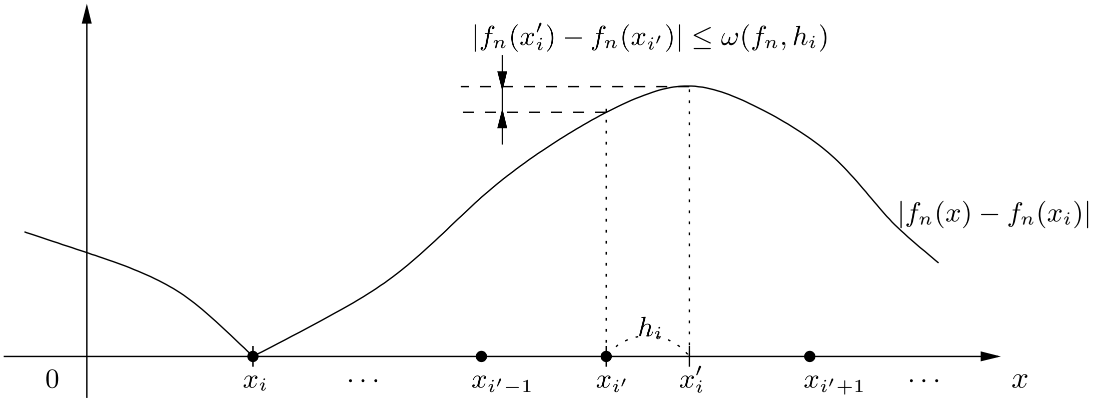


Figure 2: The plot of $|f_n(x) - f_n(x_i)|$ versus x: the value of $|f_n(x) - f_n(x_i)|$ is maximum at x_i' and this maximum value is decomposed by two factors: one is the value of $|f_n(x) - f_n(x_i)|$ at a sample point $x_{i'}$ and another is the modulus of continuity $\omega(f_n, h_i)$ with respect to h_i. The value h_i is chosen by the distance $d(x_i', x_{i'})$.

Thus, the probability of an event A is bounded by

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

Here, let us set

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

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

$$\frac{1}{N}\sum_{i=1}^{N}\int_X |f_n(x) - f_n(x_i)| dP(x) \leqslant \frac{1}{N^2}\sum_{i,j=1}^{N} |f_n(x_i) - f_n(x_j)| + \sqrt{\frac{1}{2N}\ln\frac{2}{\delta_1}}\left(\omega(f_n, h_0) + |f_n(x_0) - f_n(x_0')|\right). \quad (22)$$

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:

$$\frac{1}{N}\sum_{i=1}^{N}\int_X |f(x) - f(x_i)| dP(x) \leqslant \frac{1}{N^2}\sum_{i,j=1}^{N} |f(x_i) - f(x_j)| + 2\|f\|_\infty\sqrt{\frac{1}{2N}\ln\frac{2}{\delta_1}}. \quad (23)$$

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 (22) and (23), the difference between the true and empirical risks is bounded by the following inequality with a probability of at least $1-\delta_1$:

$$\begin{aligned}
&\int_X |f(x)-f_n(x)|dP(x) - \frac{1}{N}\sum_{i=1}^{N}|f(x_i)-f_n(x_i)| \\
&\qquad \leqslant \frac{1}{N^2}\sum_{i,j=1}^{N}(|f(x_i)-f(x_j)|+|f_n(x_i)-f_n(x_j)|) \\
&\qquad\qquad + \sqrt{\frac{1}{2N}\ln\frac{2}{\delta_1}}\,(\omega(f_n,h_0)+|f_n(x_0)-f_n(y_0)|+2\|f\|_\infty). \quad (24)
\end{aligned}$$

Second, let us consider the noisy condition; that is, $y=f(x)+\varepsilon$. Here, we assume that for the output samples $y_1,\cdots,y_N$, the noise terms $\varepsilon_1,\ldots,\varepsilon_N$ are *i.i.d.* random variables with a mean of 0 and a variance of σ_ε^2. We will define the event B as

$$|\varepsilon| \geqslant a, \quad (25)$$

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

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

Let us set

$$\delta_2 = \frac{\sigma_\varepsilon^2}{a^2}.$$

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

$$|\varepsilon| \leqslant \sigma_\varepsilon\sqrt{\frac{1}{\delta_2}}.$$

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

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

Here, let us define the event E as

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

Then, from the Hoeffding inequality, we obtain

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

Let us set

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

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

$$\frac{1}{N}\sum_{i=1}^{N}\int|y-y_i|dP(y)-\frac{1}{N^2}\sum_{i,j=1}^{N}|y_i-y_j|\leqslant\sqrt{\frac{2}{N}\ln\frac{2}{\delta_1}}\left(\|f\|_\infty+\sigma_\varepsilon\sqrt{\frac{1}{\delta_2}}\right) \tag{26}$$

since

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

Similar to (24), the difference between the expected and empirical risks of $|y-f_n(x)|$ is bounded by

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

Here, let us set $\delta_1=\delta_2=\delta$. This is possible by controlling the value of a in (25). Then, finally, from (22) and (26), with a probability of at least $1-2\delta$

$$\begin{aligned}\int_{X\times\mathbb{R}}|y-f_n(x)|dP(x,y) &- \frac{1}{N}\sum_{i=1}^{N}|y_i-f_n(x_i)|\\ &\leqslant \frac{1}{N^2}\sum_{i,j=1}^{N}(|y_j-y_i|+|f_n(x_j)-f_n(x_i)|)\\ &\quad+(\omega(f_n,h_0)+C)\sqrt{\frac{1}{2N}\ln\frac{2}{\delta}},\end{aligned} \tag{27}$$

where $C=|f_n(x_0)-f_n(y_0)|+2\|f\|_\infty+2\sigma_\varepsilon\sqrt{1/\delta}$. □

A.4 Proof of Corollary 1

Let 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,

$$\frac{1}{N}\sum_{i,j=1}^{N} |y_i - y_j| = \mathbf{a}^T H_y \mathbf{a}. \tag{28}$$

Here, the matrix H_y can be decomposed by

$$H_y = E\Lambda E^T = \sum_{i=1}^{N} \lambda_i \mathbf{e}_i \mathbf{e}_i^T, \tag{29}$$

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. Then, from (28) and (29),

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

Now, let us consider the following inequality:

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

Therefore, from the above inequality and (27), the following inequality holds with a probability of at least $1 - 2\delta$:

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

□

Appendix B.

In this appendix, we prove the theorem 2 in Section 3.2; that is, the case of multivariate target functions.

B.1 Proof of Theorem 2

First, let us consider noise free target function; that is,

$$y = f(\mathbf{x}).$$

The probability that the difference between the expected and empirical risks is larger than a positive constant ε can be described by

$$\Pr\{R(f_n)_{L_1} - R_{emp}(f_n)_{L_1} > \varepsilon\} \leqslant \exp\left\{\frac{-2\varepsilon^2 N}{(\max_{\mathbf{x}\in X}|y - f_n(\mathbf{x})|)^2}\right\} \tag{30}$$

from the Hoeffding inequality (Hoeffding, 1963). Here, there exist $\mathbf{x}_0 \in X$ and $\mathbf{x}_{i_0} \in \{\mathbf{x}_0, \cdots, \mathbf{x}_N\}$ such that

$$\mathbf{x}_0 = \arg\max_{\mathbf{x}\in X}|f(\mathbf{x}) - f_n(\mathbf{x})| \text{ and } d(\mathbf{x}_{i_0}, \mathbf{x}_0) \leqslant h_0$$

because $f - f_n \in C(X)$ and X is a compact subset of $\mathbb{R}^m$. Thus, from the dominator term of the righthand side of (30), we have

$$\begin{aligned}\max_{\mathbf{x}\in X}|f(\mathbf{x}) - f_n(\mathbf{x})| &\leqslant |f(\mathbf{x}_0) - f_n(\mathbf{x}_0) - f(\mathbf{x}_{i_0}) + f_n(\mathbf{x}_{i_0})| + |f(\mathbf{x}_{i_0}) - f_n(\mathbf{x}_{i_0})| \\ &\leqslant \omega(f - f_n, h_0) + |f(\mathbf{x}_{i_0}) - f_n(\mathbf{x}_{i_0})|. \end{aligned} \tag{31}$$

Here, we set the bound on the probability of (30) as

$$\begin{aligned}\exp\left\{\frac{-2\varepsilon^2 N}{(\max_{\mathbf{x}\in X}|f(\mathbf{x}) - f_n(\mathbf{x})|)^2}\right\} &\leqslant \exp\left\{\frac{-2\varepsilon^2 N}{(\omega(f - f_n, h_0) + |f(\mathbf{x}_{i_0}) - f_n(\mathbf{x}_{i_0})|)^2}\right\} \\ &\leqslant \frac{\delta}{2}. \end{aligned} \tag{32}$$

Therefore, from (30), (31), and (32), the following inequality holds with a probability of at least $1 - \delta/2$:

$$R(f_n)_{L_1} \leqslant R_{emp}(f_n)_{L_1} + \{\omega(f - f_n, h_0) + |f(\mathbf{x}_{i_0}) - f_n(\mathbf{x}_{i_0})|\}\sqrt{\frac{1}{2N}\ln\frac{2}{\delta}}. \tag{33}$$

Second, let us consider the noisy target function; that is,

$$y = f(\mathbf{x}) + \varepsilon.$$

From Chebyshev inequality, the following inequality always holds:

$$\Pr\{|\varepsilon| > a\} \leqslant \frac{\sigma_\varepsilon^2}{a^2}, \tag{34}$$

where a represents a positive constant. In this case, from the triangular inequality, $|y - f_n(\mathbf{x})|$ has the following upper bound:

$$\max_{\mathbf{x}\in X}|y - f_n(\mathbf{x})| \leqslant \max_{\mathbf{x}\in X}|f(\mathbf{x}) - f_n(\mathbf{x})| + |\varepsilon|. \tag{35}$$

Let us set the bound on the probability of (34) as

$$\frac{\sigma_\varepsilon^2}{a^2} = \frac{\delta}{2}. \tag{36}$$

Then, from (31), (35), and (36), the following inequality holds with a probability of at least $1-\delta/2$:

$$\begin{aligned}\max_{\mathbf{x}\in X}|y-f_n(\mathbf{x})| &\leqslant \max_{\mathbf{x}\in X}|f(\mathbf{x})-f_n(\mathbf{x})|+\sigma_\varepsilon\sqrt{\frac{2}{\delta}}\\ &\leqslant \omega(f-f_n,h_0)+|f(\mathbf{x}_{i_0})-f_n(\mathbf{x}_{i_0})|+\sigma_\varepsilon\sqrt{\frac{2}{\delta}}. \qquad (37)\end{aligned}$$

Therefore, from (33) and (37), the following inequality holds with a probability of at least $1-\delta$:

$$R(f_n)_{L_1} \leqslant R_{emp}(f_n)_{L_1} + \left\{\omega(f-f_n,h_0)+|f(\mathbf{x}_{i_0})-f_n(\mathbf{x}_{i_0})|+\sigma_\varepsilon\sqrt{\frac{2}{\delta}}\right\}\sqrt{\frac{1}{2N}\ln\frac{2}{\delta}}.$$

□

References

H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Proceedings of the Second International Symposium on Information Theory*, pages 267–281, 1973.

G. Anastassiou and S. Gal. *Approximation Theory: Moduli of Continuity and Global Smoothness Preservation*. Birkhäuser, Boston, 2000.

A. Barron, J. Rissanen, and B. Yu. The minimum description length principle in coding and modeling. *IEEE Transactions on Information Theory*, 44:2743–2760, 1998.

C. Chang and C. Lin. LIBSVM: a library for support vector machines, 2001. Software available at `http://www.csie.ntu.edu.tw/˜cjlin/libsvm.`

O. Chapelle, V. Vapnik, and Y. Bengio. Model selection for small sample regression. *Machine Learning*, 48:315-333, 2002.

V. Cherkassky and Y. Ma. Comparison of model selection for regression. *Neural Computation*, 15:1691–1714, 2003.

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

S. Cohen and N. Intrator. On different model selection criteria in a forward and backward regression hybrid network. *International Journal of Pattern Recognition and Artificial Intelligence*, 18:847–865, 2004.

D. Donoho and I. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, 90:1200–1224, 1995.

D. Foster and E. George. The risk inflation criterion for multiple regression. *Annals of Statistics*, 22:1947–1975, 1994.

T. Hastie, R. Tibshirani, and J. Friedman. Note on "comparison of model selection for regression" by V. Cherkassky and Y. Ma. *Neural Computation*, 15:1477–1480, 2003.

G. Hinton and D. van Camp. Keeping neural networks simple by minimizing the description length of the weights. In *Proceedings of the Sixth Annual ACM Conference on Computational Learning Theory*, pages 5–13, 1993.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13-30, 1963.

M. Karpinski and A. Macintyre. Polynomial bounds for VC dimension of sigmoidal neural networks. In *Proceedings of the Twenty-Seventh Annual ACM Symposium on Theory of Computing*, pages 200–208, 1995.

I. Koo and R. Kil. Nonlinear model selection based on the modulus of continuity. In *Proceedings of World Congress on Computational Intelligence*, pages 3552–3559, 2006.

G. Lorentz. *Approximation of Functions*. Chelsea Publishing Company, New York, 1986.

J. Rissanen. Stochastic complexity and modeling. *Annals of Statistics*, 14:1080–1100, 1986.

A. Sakurai. Polynomial bounds for the VC dimension of sigmoidal, radial basis function, and sigma-pi networks. In *Proceedings of the World Congress on Neural Networks*, pages 58–63, 1995.

G. Schwartz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

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.

V. Vapnik. *Statistical Learning Theory*. J. Wiley, 1998.

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