# Model Selection Based on the Modulus of Continuity

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
Email: rmkil@kaist.ac.kr

**Abstract.** This paper presents a new method of model selection in regression problems based on the modulus of continuity. For this purpose, the risk function bounds are suggested using the modulus of continuity. The suggested bounds incorporate the information of learned results as well as the structural information of the regression model. As a result, the suggested bounds can fit the risk functions of the trained regression models in more realistic sense. Through the simulation for function approximation, we have shown that the model selection based on the suggested bounds can provide the better performance than other model selection methods from the view points of the risk functions and the number of parameters.

## 1 Introduction

Model selection is an important issue of selecting the reasonable network size to get the optimal performance of regression models. The regression of real-valued functions is performed for the 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 whole distribution of samples (not just given samples) should be optimized. The performance of regression models can be decomposed by the bias and variance terms. 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 too small so that the performance is not optimal due to large bias term 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 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 is to use a penalty (correction) factor as the measure of model complexity. The well known criteria are Akaike information criteria (AIC)[1], Bayesian information criteria (BIC)[2], generalized cross-validation (GCV)[3], minimum description length (MDL)[4, 5], and risk inflation criteria (RIC)[6]. These methods can be easily fit to linear regression models when the 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 the 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 the nonlinear models and also the regression models using 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. 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.

In this paper, we present the risk function bounds in the sense of the modulus of continuity (MC) representing a measure of continuity for the given function. Lorentz[10] applied the MC to the function approximation theories. In our method, this measure is applied to determine the risk function bounds in the $L_1$ sense. For the estimation of these bounds, the MC is analyzed for both the target and estimation functions. As a result, the suggested bounds include both the structural information such as the number of parameters and the learned results such as the weight values of the estimation network. Through the simulation for function approximation, we have shown that the MC based method can provide the better

performance than other model selection methods from the view points of the risk functions and the number of parameters.

In section 2, we introduce the various model selection methods based from classical statistic and VC approach. Section 3 describes the model selection method based on the modulus of continuity for continuous or differentiable functions. Section 4 describes simulation results for the regression of real-valued functions using the estimation model of trigonometric polynomials. Finally, section 5 makes the conclusion.

## 2 Risk Function Bounds in 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 $\mathbb{R}^n$ and $C^k(X, \mathbb{R})$ is a class of functions having continuous $k$th derivative on $X$. Let the observed output $y$ be represented by

$$y = f(x) + \epsilon \tag{1}$$

where $f(x)$ is the target function and $\epsilon$ is a random noise with mean zero and variance $\sigma^2$. Here, the input and output samples $(x_i, y_i)$, $i = 1, \cdots, N$ are randomly generated by a distribution $P$ where $x_i \in X$ and $y_i = f(x_i) + \epsilon$. For these samples, our goal is to construct an estimation network (or function) $f_n(x)$ which minimizes the expected risk (or true risk)

$$R(f_n) = \int_{X \times \mathbb{R}} L(y, f_n(x)) dP(x, \epsilon) \tag{2}$$

with respect to the 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. To minimize the expected risk (16), we have to identify the distribution $P(x, \epsilon)$ but it is usually unknown. Rather, the parameters to minimize the empirical risk

$$R_{emp}(f_n) = \frac{1}{l} \sum_{i=1}^{l} L(y_i, f_n(x_i)) \tag{3}$$

are obtained for the given samples.

If we increase the number of parameters, the empirical risk of (3) 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 problems is referred to as the model selection, that is, we have to determine the optimal number of parameters (or complexity), that is, 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 risk functions for the given parameters in regression models. Here, let us represent the form of the estimate of risk functions as

$$\hat{R}(f_n) = R_{emp}(f_n)T(n,l) \tag{4}$$

where $T(n,l)$ is a penalty (correction) factor, $n$ 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 (4), they are

$$T_{AIC}(n,l) = \frac{1+\frac{n}{l}}{1-\frac{n}{l}} \text{ and} \tag{5}$$

$$T_{BIC}(n,l) = 1 + \frac{\ln l}{2}\frac{\frac{n}{l}}{\left(1-\frac{n}{l}\right)} \tag{6}$$

where $T_{AIC}$ and $T_{BIC}$ represent the penalty factors of the AIC and BIC criteria respectively. These model selection criteria come form the asymptotic analysis for linear models using $L_2$ 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 points of the number of samples. As the hypothesis space is increased, the empirical risk can be decreased but the confidence interval, which is proportional to the complexity of the estimation network, is increased. In this

sense, we need to make the proper trade-off between the empirical risk and the complexity of the estimation network. The 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 risk functions with the VC dimension $d$ associated with the hypothesis space defined by $f_n$, holds the following inequality with as least probability $1-\delta$[7, 9]:

$$\hat{R}(f_n) \leqslant R_{emp}(f_n) \left(1 - c\sqrt{\frac{d\left(1+\ln\frac{l}{d}\right) - \ln\delta}{l}}\right)_+^{-1} \tag{7}$$

where $c$ is a constant dependent on 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 (7) can be described in easier form to calculate. For instance, let the estimation network $f_n(x)$ be described by

$$f_n(x) = \sum_{k=1}^{n} w_k \phi_k(x) \tag{8}$$

where $\phi_k(x)$ represents the $k$th orthogonal polynomial and a continuous target function $f(x)$ in some Hilbert space with respect to the basis functions $\{\phi_1(x), \cdots, \phi_n(x)\}$ be given to be approximated. Then, the estimated expected risk[8] is given by

$$E[R(f_n)] = E[R_{emp}(f_n)] \left(1 + \frac{E[\sum_{i=1}^{n} 1/\lambda_i]}{l}\right) \left(1 - \frac{d}{l}\right)^{-1} \tag{9}$$

where $\{\lambda_1, \cdots, \lambda_n\}$ are the eigenvalues of the $n \times n$ covariant matrix $C$ in which the $pq$th 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, Chepelle et al.[8] suggest the following bound with the confidence $\delta = 0.1$:

$$E[R(f_n)] \leqslant E[R_{emp}(f_n)] T_{SEB}(n, l) \tag{10}$$

where

$$T_{SEB}(n,l) = \frac{1+\frac{n}{lk}}{1-\frac{n}{l}} \text{ and} \tag{11}$$

$$k = \left(1 - \sqrt{\frac{n\left(1+\ln\frac{2l}{n}\right)+4}{l}}\right)_+ . \tag{12}$$

The risk function of (10) 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 Risk Function Bounds Based on the Modulus of Continuity

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

$$\omega(f,h) = \max_{x,t\in X,|t|\leqslant h} |f(x+t) - f(x)| \tag{13}$$

where $h$ a positive constant. Here, the function $\omega(f,h)$ is called the modulus of continuity of $f$. The modulus of continuity of $f$ has the following properties: (1) $\omega(f,h)$ is continuous at $h=0$ for each $f$, (2) $\omega(f,h)$ is positive and increasing for $h>0$, (3) $\omega(f,h)$ is sub-additive, that is, $\omega(f,h_1+h_2) \leqslant \omega(f,h)+\omega(f,h_2)$ for each $f$, and (4) $\omega(f,h)$ is continuous for $h>0$.

Assume that the input data sample $x_i$ $(\in X)$ is randomly generated by a probability $P$ with unknown distribution function $F$. Then, we can say that there exists a partition $X_i \subset X$ satisfying

$$x_i \in X_i \text{ and } P(X_i) = \frac{1}{N} \text{ for } i = 1, \cdots, N. \tag{14}$$

Since $X$ is a compact subset of $\mathbb{R}$, we can take $h>0$ such that

$$\max_{1\leqslant i\leqslant N} \operatorname{diam}(X_i) \leqslant h. \tag{15}$$

where $\mathrm{diam}(X_i) = \sup_{x,y \in X_i} |x - y|$.

For the selection of $f_n(x)$, 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 risk function and the empirical risk are defined by the following $L_1$ measure:

$$R(f_n)_{L_1} = \int_{X \times \mathbf{R}} |y - f_n(x)| dP(x, y) \text{ and} \tag{16}$$

$$R_{emp}(f_n)_{L_1} = \frac{1}{N} \sum_{i=1}^{N} |y_i - f_n(x_i)|. \tag{17}$$

For these risk functions, we can suggest the following theorem based on the modulus continuity:

**Theorem 1** *Let the samples $x_i \in X$ and target values $y_i$, $i = 1, \cdots, N$, be generated by (1) and $h$ be a constant satisfying condition (15). Let us also assume that the estimation function $f_n(x)$ is described by the linear combination of basis function $\{\phi_k(x)\}_{k=1}^n$, that is,*

$$f_n(x) = \sum_{k=1}^{n} w_k \phi_k(x) \tag{18}$$

*where $w_k$ represents the weight value associated with the $k$th basis function. Then, the risk function for a continuous function $f$ on $X$ is bounded by the following inequality with the probability at least $1 - \delta$:*

$$R(f_n)_{L_1} \leqslant R_{emp}(f_n)_{L_1} + \omega(f, h) + \omega(f_n, h) + \sigma \sqrt{\frac{2}{\delta}} \tag{19}$$

*where*

$$\omega(f_n, h) \leqslant \max\{\omega(\phi_k, h)\}_{k=1}^n \sum_{k=1}^{n} |w_k|. \tag{20}$$

For the proof of this theorem, refer to the appendix A-1. This theorem states that the risk function $R(f_n)_{L_1}$ is bounded by the empirical risk $R_{emp}(f_n)_{L_1}$, the modulus of continuity for target function $\omega(f, h)$, and also for the estimation function $\omega(f_n, h)$. For the fixed $h > 0$, the modulus of continuity for the target function $w(f, h)$ can be considered as a constant but the modulus of continuity for the estimation function $\omega(f_n, h)$ is proportional to the complexity measurement which is described by the modulus of continuity of basis functions

and the sum of absolute value of weights. If the number of parameters $n$ is increasing, the empirical risk $R_{emp}(f_n)_{L_1}$ is decreasing, but the modulus of continuity of estimation function $w(f_n, h)$ is increasing in general. So we need the trade-off between the empirical risk and the modulus of continuity of target function to obtain the optimal estimation network.

The modulus of continuity for target function $w(f, h)$ is related with the degree of approximation. In the classical function approximation theory, the modulus of continuity plays important role for the degree of approximation of $f$[10]. For example, if $f \in C([0, 2\pi])$ and $f_n$ is the trigonometric polynomial with degree $n$ on $[0, 2\pi]$, the following approximation bound can be satisfied:

$$\sup_{f_n \in \Phi} \|f - f_n\|_\infty \leqslant M\omega(f, \frac{1}{n}) \tag{21}$$

where $\Phi$ represents the linear combination of $\phi_1, \cdots, \phi_n$ and $M$ represents a constant dependent upon the target function $f$. In general, we cannot compute the modulus of continuity in advance because we don't exactly know the target function $f$. Assuming that the first or second derivatives of $f$ is bounded, that is, $\|f'\|_\infty \leqslant M_1$ or $\|f''\|_\infty \leqslant M_2$ for some $M_1, M_2 > 0$, we can estimate the modulus of continuity as follows:

1) if $f \in C^1(X, \mathbb{R})$,

$$\omega(f, h) \leqslant O\left(\|f'\|_\infty h \log \frac{l}{h} + \frac{h}{l}\|f\|_\infty\right), \tag{22}$$

2) and if $f \in C^2(X, \mathbb{R})$,

$$\omega(f, h) \leqslant O\left(\|f''\|_\infty h(l - h) + \frac{h}{l}\|f\|_\infty\right) \tag{23}$$

where $l$ represents a measure of domain $X$. These inequalities show that the modulus of continuity for the target function $f$ is bounded by the terms of the sup-norm of $f$, the maximum distance $h$ between the sample data, and the measure of $X$. Note that as the number of samples increases, $h$ decreases and it makes the modulus of continuity decreases.

For the modulus of continuity of the estimation function $\omega(f_n, h)$, let us consider the regression models with the following basis functions:

**Example:** Consider the trigonometric polynomials $\phi_0(x) = 1/2$, $\phi_{2j-1}(x) = \sin jx$, and $\phi_{2j}(x) = \cos jx$ for $j = 1, 2, \cdots, n$. Applying the mean value theorem to $\phi_j$, we get

$$\begin{aligned}\omega(\phi_k, h) &\leqslant \|\phi_k'\|_\infty h \\ &\leqslant n \cdot h \quad \text{for} \quad k = 0, \cdots, 2n.\end{aligned}$$

That is,

$$\max_{0 \leqslant k \leqslant 2n} \{\omega(\phi_k, h)\} \leqslant n \cdot h.$$

Therefore, the upper bounds of risk functions can be determined by

$$R(f_n)_{L_1} \leqslant R_{emp}(f_n)_{L_1} + \omega(f, h) + n \cdot h \sum_{k=0}^{2n} |w_k| + \sigma\sqrt{\frac{2}{\delta}}. \tag{24}$$

Similarly, we can derive the upper bounds of risk functions for the regression model using algebraic polynomials, that is, $f_n(x) = \sum_{i=0}^{n} w_i x^i$.

**Example:** Consider the following form of 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 $\sigma_k^2$ represent the center and width of the basis function $\phi_k$ respectively. Applying the mean value theorem to $\phi_k$, 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 $\phi_k'$ has the maximum and minimum at $x = t_i - \sigma_k$ and $x = t_i + \sigma_k$ respectively. This implies that

$$\max_{1 \leqslant k \leqslant n} \{\omega(\phi_k, h)\} \leqslant \frac{h}{\sigma'} \exp\left(-\frac{1}{2}\right)$$

where

$$\sigma' = \min_{1 \leqslant k \leqslant n} \sigma_k.$$

Therefore, the upper bound of risk function can be determined by

$$R(f_n)_{L_1} \leqslant R_{emp}(f_n)_{L_1} + \omega(f,h) + \frac{h}{\sigma'}\exp\left(-\frac{1}{2}\right)\sum_{k=1}^{n}|w_k| + \sigma\sqrt{\frac{2}{\delta}}. \tag{25}$$

**Example:** Consider the following form of sigmoid functions:

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

where $a_k$ and $b_k$ represent the adaptable parameters of the basis function $\phi_k$. Applying the mean value theorem to $\phi_k$, we get

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

since $\phi_k'$ has the maximum at $x = b_k/a_k$. That is,

$$\max_{1\leqslant k\leqslant n}\{\omega(\phi_k, h)\} \leqslant \frac{h}{4}.$$

Therefore, the upper bound of the risk functions can be determined by

$$R(f_n)_{L_1} \leqslant R_{emp}(f_n)_{L_1} + \omega(f,h) + \frac{h}{4}\sum_{k=1}^{n}|w_k| + \sigma\sqrt{\frac{2}{\delta}}. \tag{26}$$

The suggested theorem of true risk bounds is derived in the sense of the $L_1$ measure, that is, the risk function described by (16). Other model selection criteria such as AIC and SEB are derived from the risk function using the quadratic loss function $L(y, f_n(x)) = (y - f_n(x))^2$. In this context, we can show that the optimal function $f_{n^*}(x)$ determined from the risk function of the $L_1$ measure is also optimal in the sense of the risk function using the quadratic loss function under the assumption that the estimation function is unbiased and the error term $y - f_n(x)$ has normal distribution.

## 4 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 MC 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 the linear regression model so that the VC dimension of the network was easy to find out and the optimization of network parameters could be easily solved. This makes 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 $x_i$ were uniformly distributed within the interval of $[-\pi, \pi]$. As the target functions, the following functions were used:

1) the step function defined by

$$f(x) = \begin{cases} 1 & \text{if } x \geqslant 0 \\ 0 & \text{otherwise,} \end{cases} \tag{27}$$

2) the sine square function defined by

$$f(x) = \sin^2(x), \tag{28}$$

3) the sinc function defined by

$$f(x) = \frac{\sin(\pi x)}{\pi x}, \tag{29}$$

4) and the combined function defined by

$$f(x) = 2x \exp(-\frac{x^2}{2}) \cos(2\pi x). \tag{30}$$

The target functions were illustrated in Figure 1. The first target function was discontinuous while other functions were continuous. The second and third functions showed the

similar complexity while the fourth function was more complex than other functions. For these target function, we generated $N$ $(= 50)$ training samples randomly according to the target function to train the estimation network. We also generated 300 test samples separately. 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 $j$ represents the period of sinusoidal function. Here, the estimation function $f_n(x)$ was given by

$$f_n(x) = \sum_{k=0}^{n} w_k \phi_k(x). \tag{31}$$

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

$$\mathbf{y} = \Phi_n \mathbf{w} \tag{32}$$

where $\Phi_n$ 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 square loss function, the estimated weight vector $\widehat{\mathbf{w}}$ could be determined by

$$\widehat{\mathbf{w}} = (\Phi_n^T \Phi_n)^{-1} \Phi_n^T \mathbf{y}. \tag{33}$$

By substituting the estimated weight vector to (32), 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 MC based methods. Here, the estimated optimal number of nodes was determined by

$$\widehat{n} = \arg\min_n \widehat{R}(f_n). \tag{34}$$

Note that in the MC method, only the terms of $R_{emp}(f_n)$ and the modulus of continuity for the estimation function were considered to select the optimal number of nodes since other terms in (19) were constant once the training samples were given. To compare the risk

functions obtained for the estimated optimal number of nodes $\widehat{n}$ with the risk functions for the minimum number of nodes obtained from the test samples, we computed the log ratio of two risks, that is,

$$r_R = \log \frac{R(f_{\widehat{n}})}{\min_n R(f_n)} \tag{35}$$

where $R(f_n)$ represented the risk function for the squared error loss function $L(y, f_n) = (y - f_n(x))^2$. 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 patterns, that is,

$$r_n = \log \frac{\widehat{n}}{\arg\min_n R(f_n)}. \tag{36}$$

This node ratio represented the quality of distance between the optimal and the estimated optimal complexity of the network. After all experiments have been repeated 1000 times, the risk ratios of (35) and the node ratios of (36) were plotted using the box-plot method. The simulation results of model selection using the AIC, BIC, SEB and MC based methods were illustrated in Figures 2 through 9. 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 case of the fourth function, that is, more complicated function, 2) while the MC 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 since the true risk bounds based on the VC dimension were derived in the sense of uniform convergence, that is, the worst case in the hypothesis space. As an example, we illustrated $n^*/l$ for four target functions in Figure 10. As we expected, the fourth target function required high $n^*/l$ compared to other target functions due to the complexity of function as shown in Figure 1. In this case, the SEB method did not show good performance.

To see the risk prediction of the model selection methods, we plotted the estimated error versus the number of nodes for the AIC, BIC, and SEB methods in the $L_2$ sense of loss function, that is, the square root of risk function, and also for the MC method in the $L_1$ sense of loss function defined by (16). The predicted results were compared with test errors as shown in Figures 11 and 12. Theses results showed that 1) the risk prediction using the AIC and BIC methods fit well except the sudden change in risk functions, 2) the risk prediction using the SEB method had the tendency to fit well when the ratio of the number of nodes to the number of samples $n/l$ was small, and 3) the risk prediction using the MC method was well suited with the test errors in overall range of the number of nodes. In these predicted results, it was interesting that the MC method was able to catch the sudden change of test errors while other methods didn't.

In summary, the performance of the MC based 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 MC method was able to catch the trend of test errors. This was mainly due to the fact that the MC based method was performed using the risk function bounds incorporating the information of learned results such as the sum of absolute weights as well as the structural information of the estimation network. Furthermore, the suggested MC method can be easily extended to various types of regression models with nonlinear kernel functions which have some smoothness constraints.

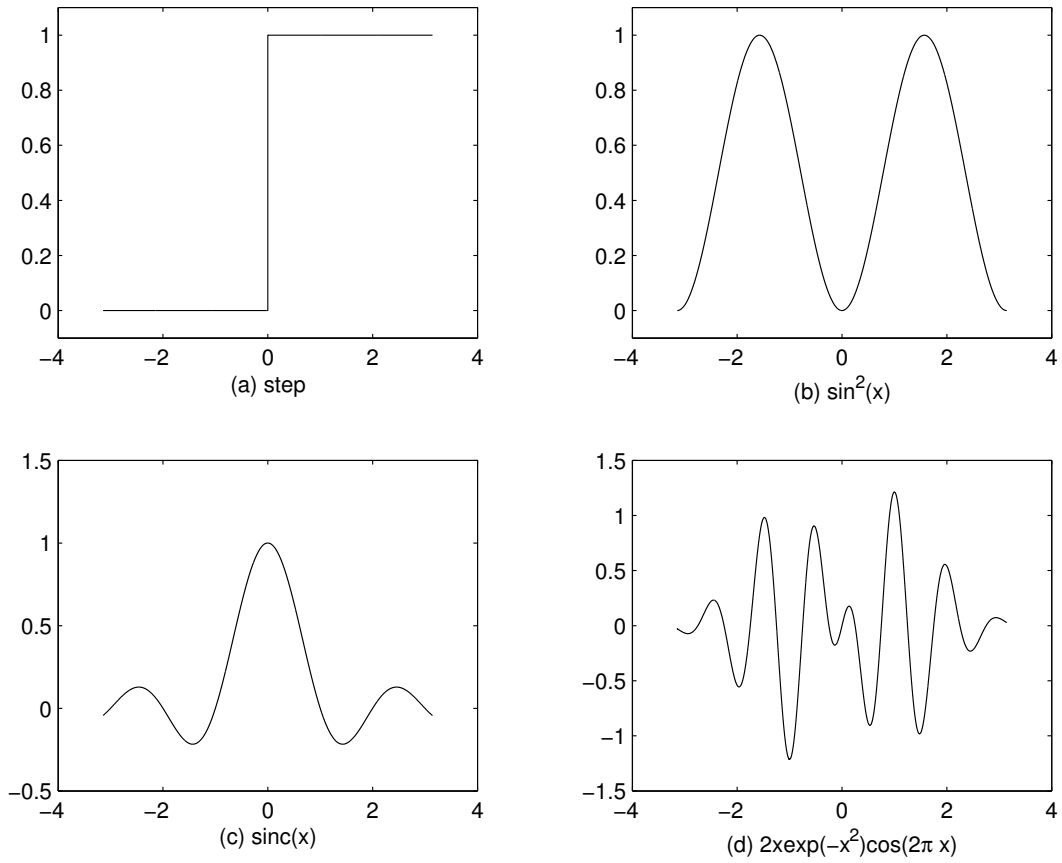



**Fig. 1.** The target functions for the simulation of model selection:

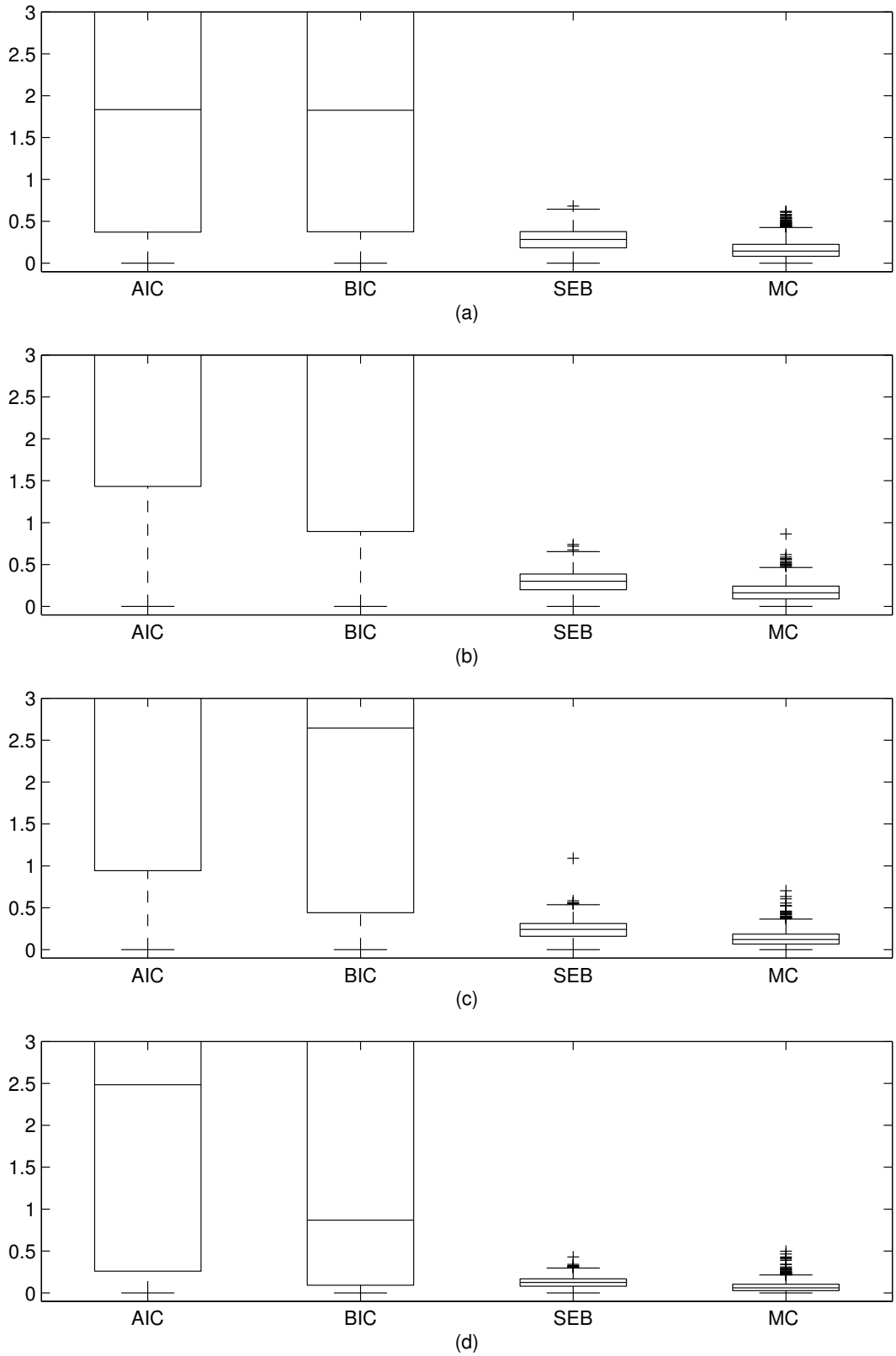



**Fig. 2.** The box-plots of risk ratios $r_R$ for the regression of step function: (a), (b), (c), and (d) represent the box-plots of $r_R$ with $\sigma = 0$, 0.025, 0.05, and 0.1 respectively.

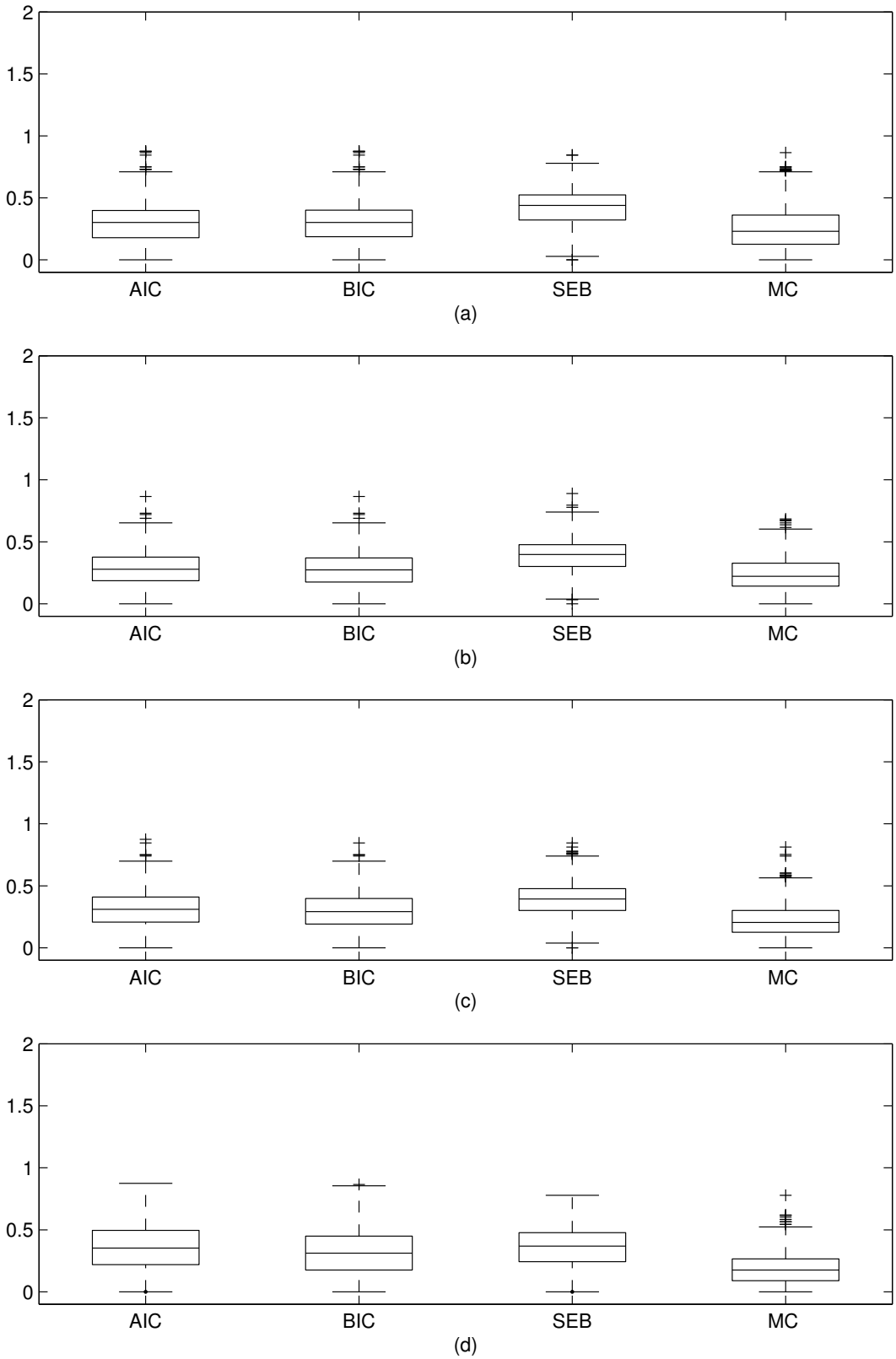



**Fig. 3.** The box-plots of node ratios $r_n$ for the regression of step function: (a), (b), (c), and (d) represent the box-plots of $r_n$ with $\sigma = 0$, 0.025, 0.05, and 0.1 respectively.

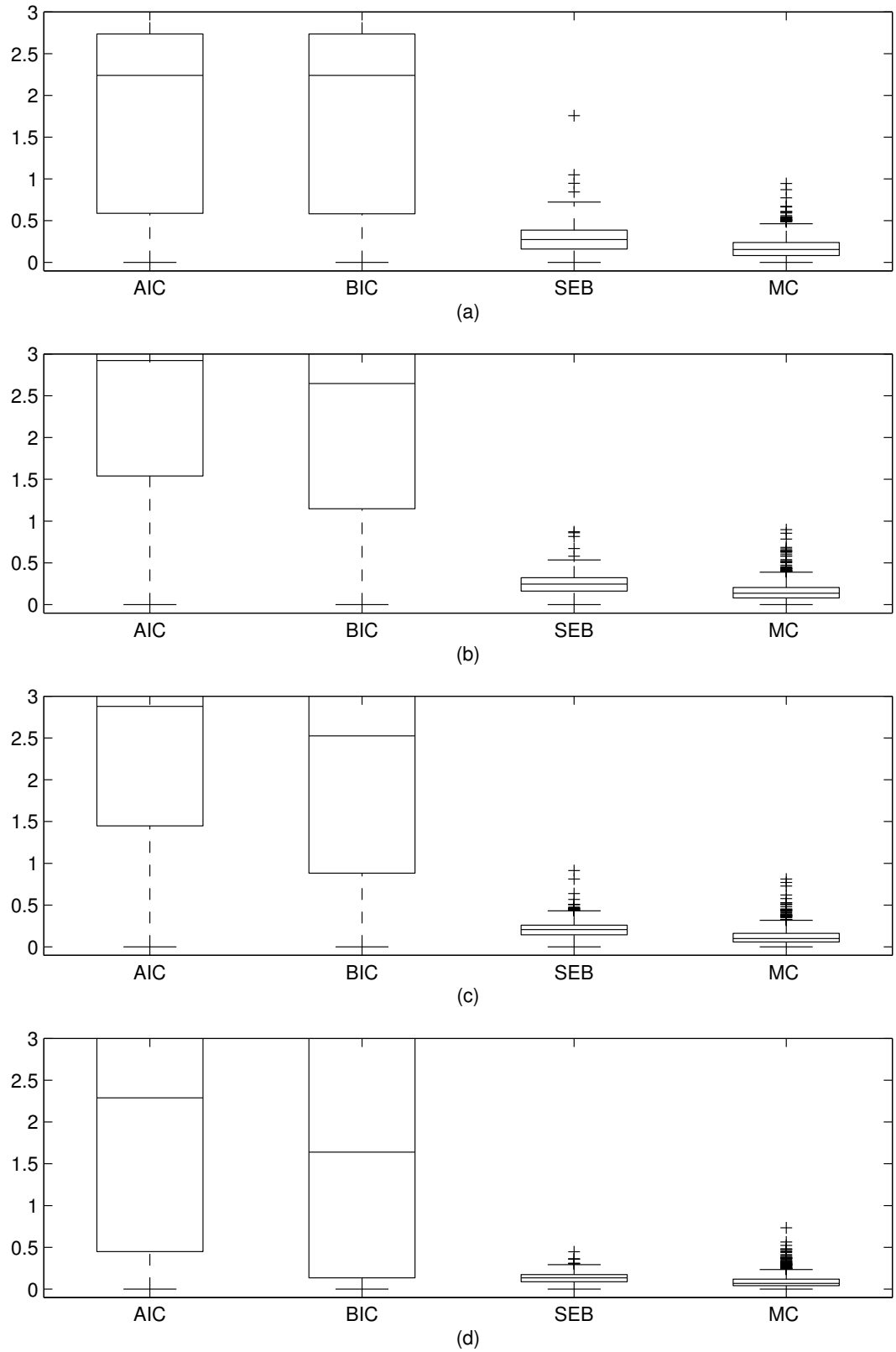



**Fig. 4.** The box-plots of risk ratios $r_R$ for the regression of sine-square function: (a), (b), (c), and (d) represent the box-plots of $r_R$ with $\sigma = 0$, 0.025, 0.05, and 0.1 respectively.

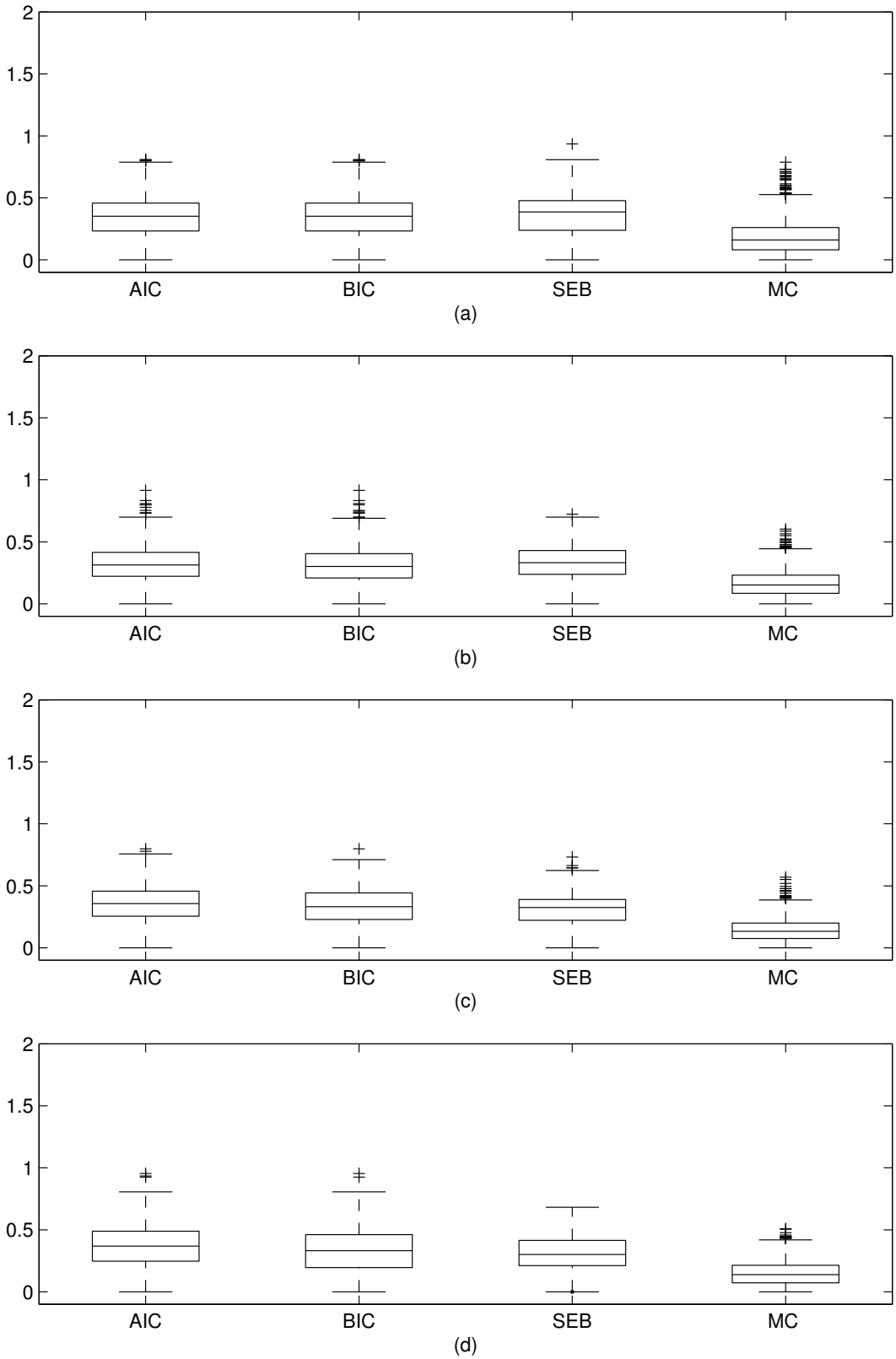



**Fig. 5.** The box-plots of node ratios $r_n$ for the regression of sine-square function: (a), (b), (c), and (d) represent the box-plots of $r_n$ with $\sigma = 0$, 0.025, 0.05, and 0.1 respectively.

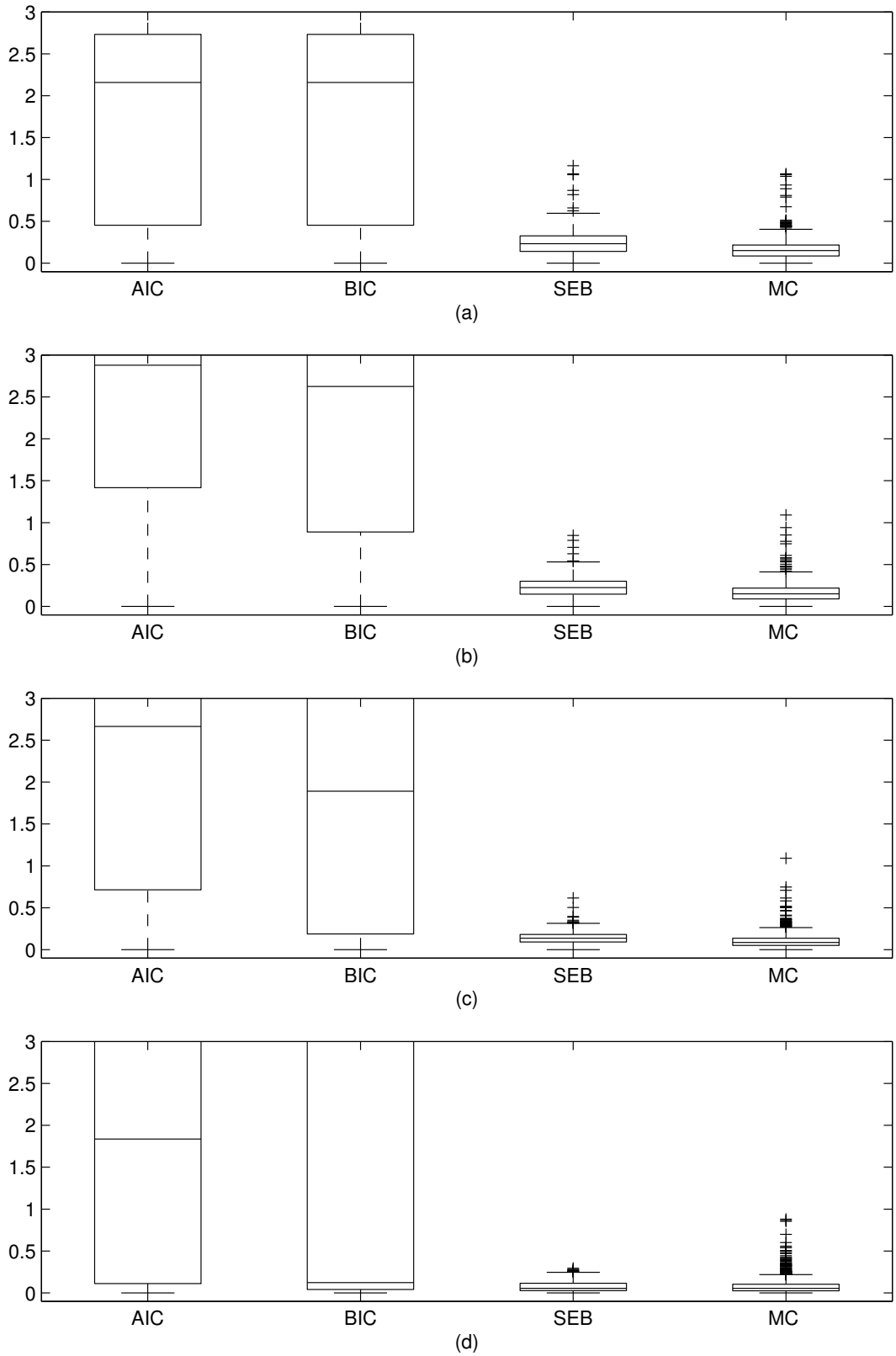



**Fig. 6.** The box-plots of risk ratios $r_R$ for the regression of sinc function: (a), (b), (c), and (d) represent the box-plots of $r_R$ with $\sigma = 0$, 0.025, 0.05, and 0.1 respectively.

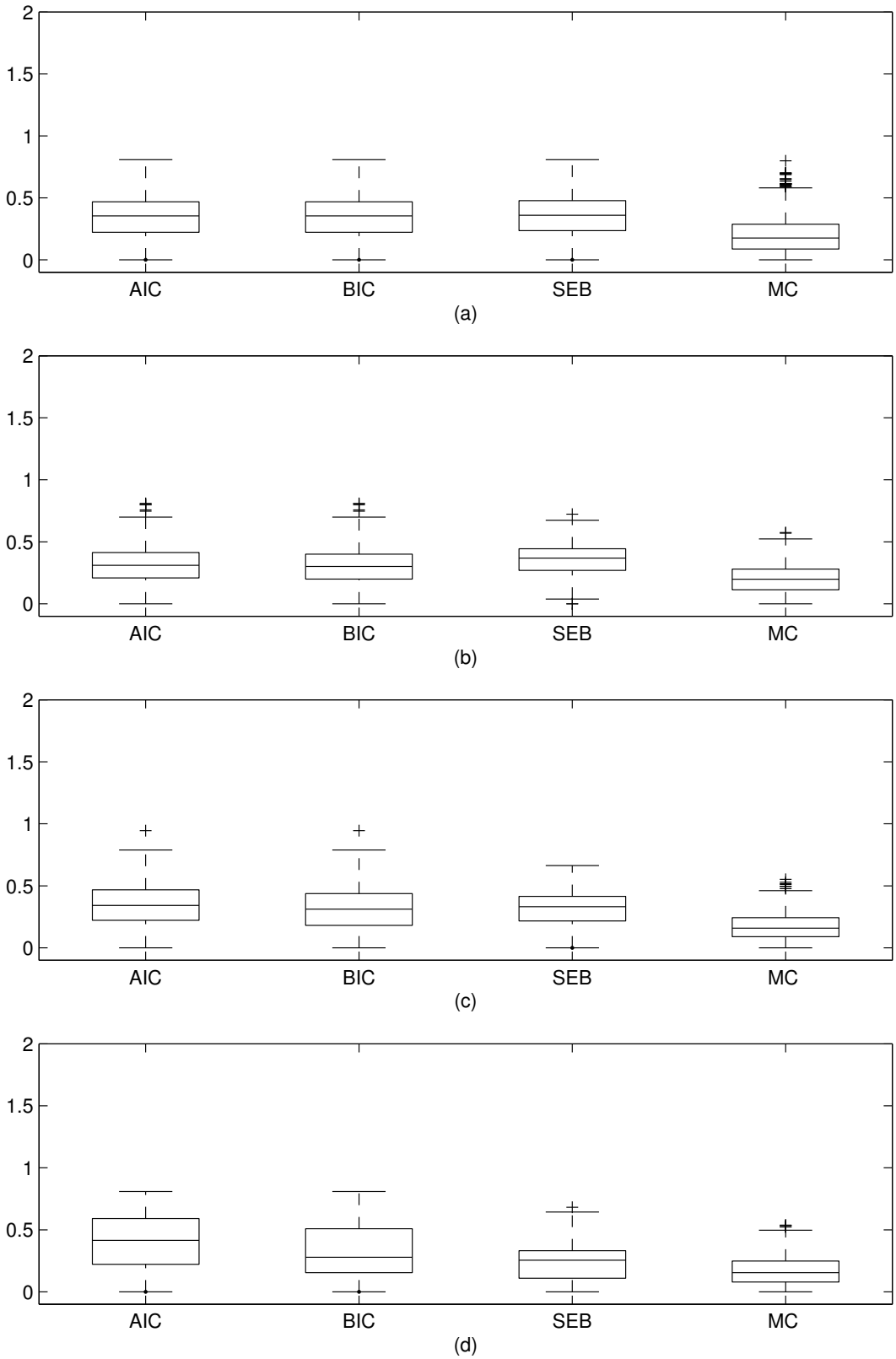



**Fig. 7.** The box-plots of node ratios $r_n$ for the regression of sinc function: (a), (b), (c), and (d) represent the box-plots of $r_n$ with $\sigma = 0$, 0.025, 0.05, and 0.1 respectively.

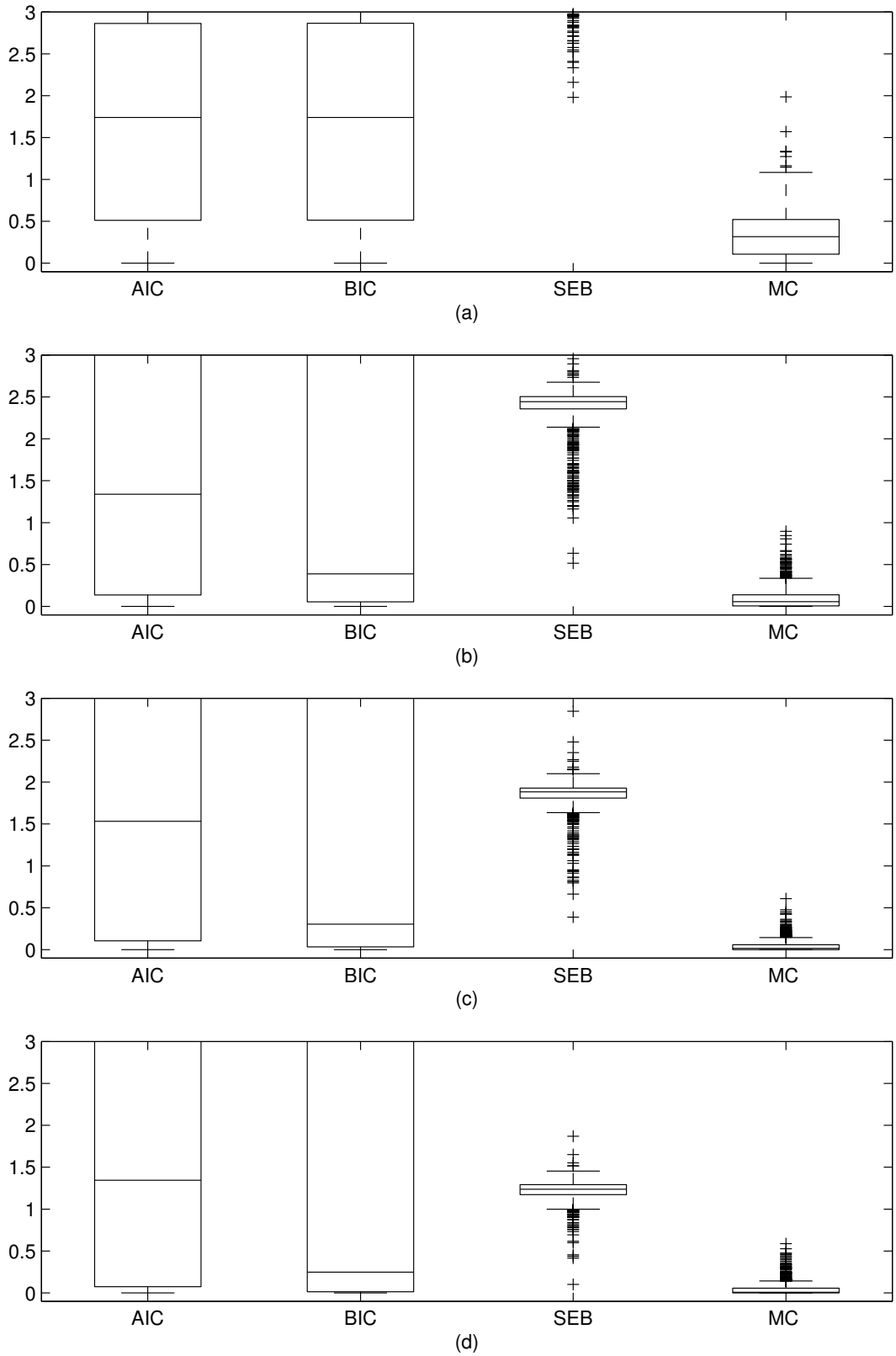



**Fig. 8.** The box-plots of risk ratios $r_R$ for the regression of combined function: (a), (b), (c), and (d) represent the box-plots of $r_R$ with $\sigma = 0$, 0.025, 0.05, and 0.1 respectively.

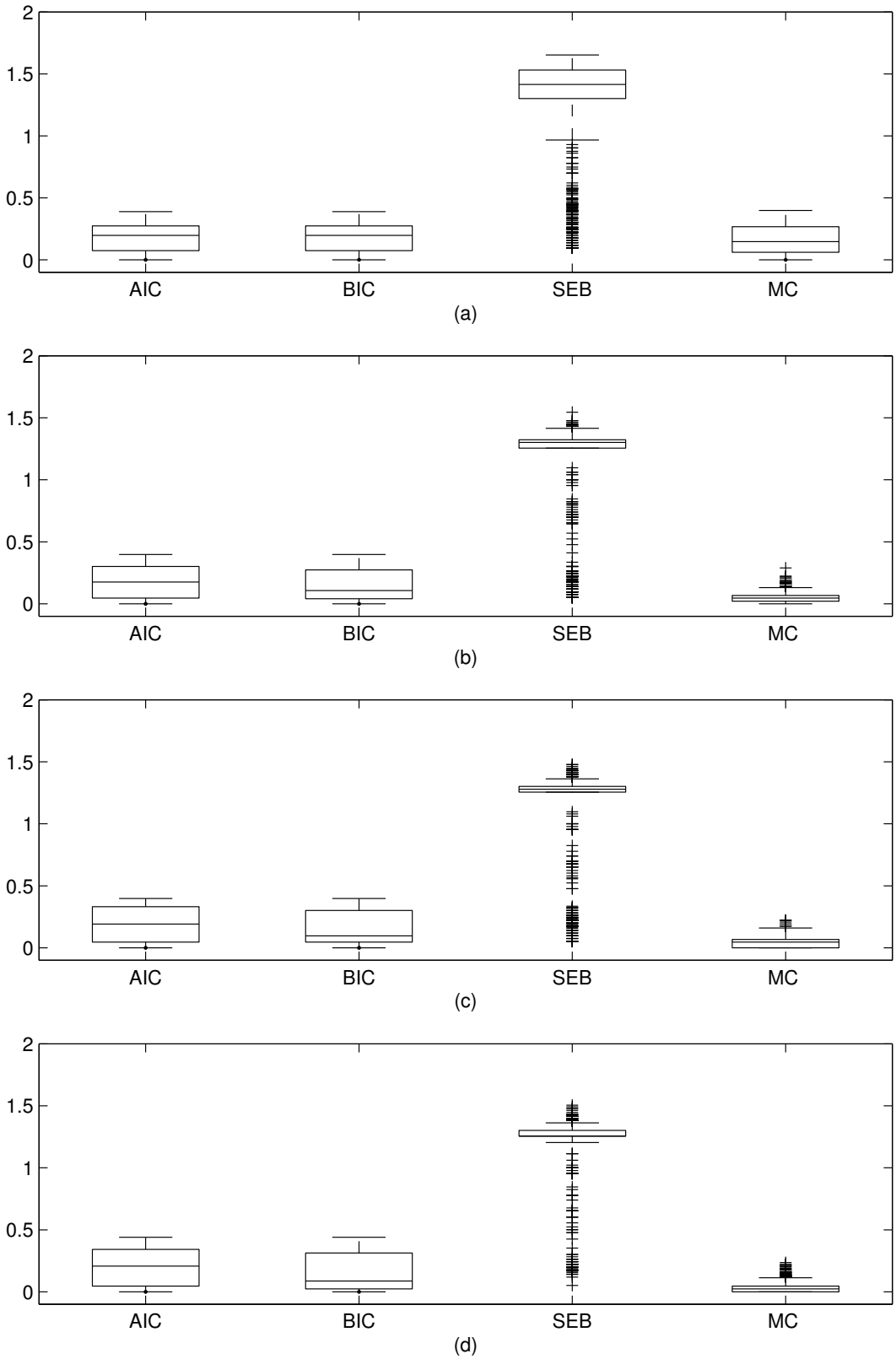



**Fig. 9.** The box-plots of node ratios $r_n$ for the regression of combined function: (a), (b), (c), and (d) represent the box-plots of $r_n$ with $\sigma = 0$, 0.025, 0.05, and 0.1 respectively.

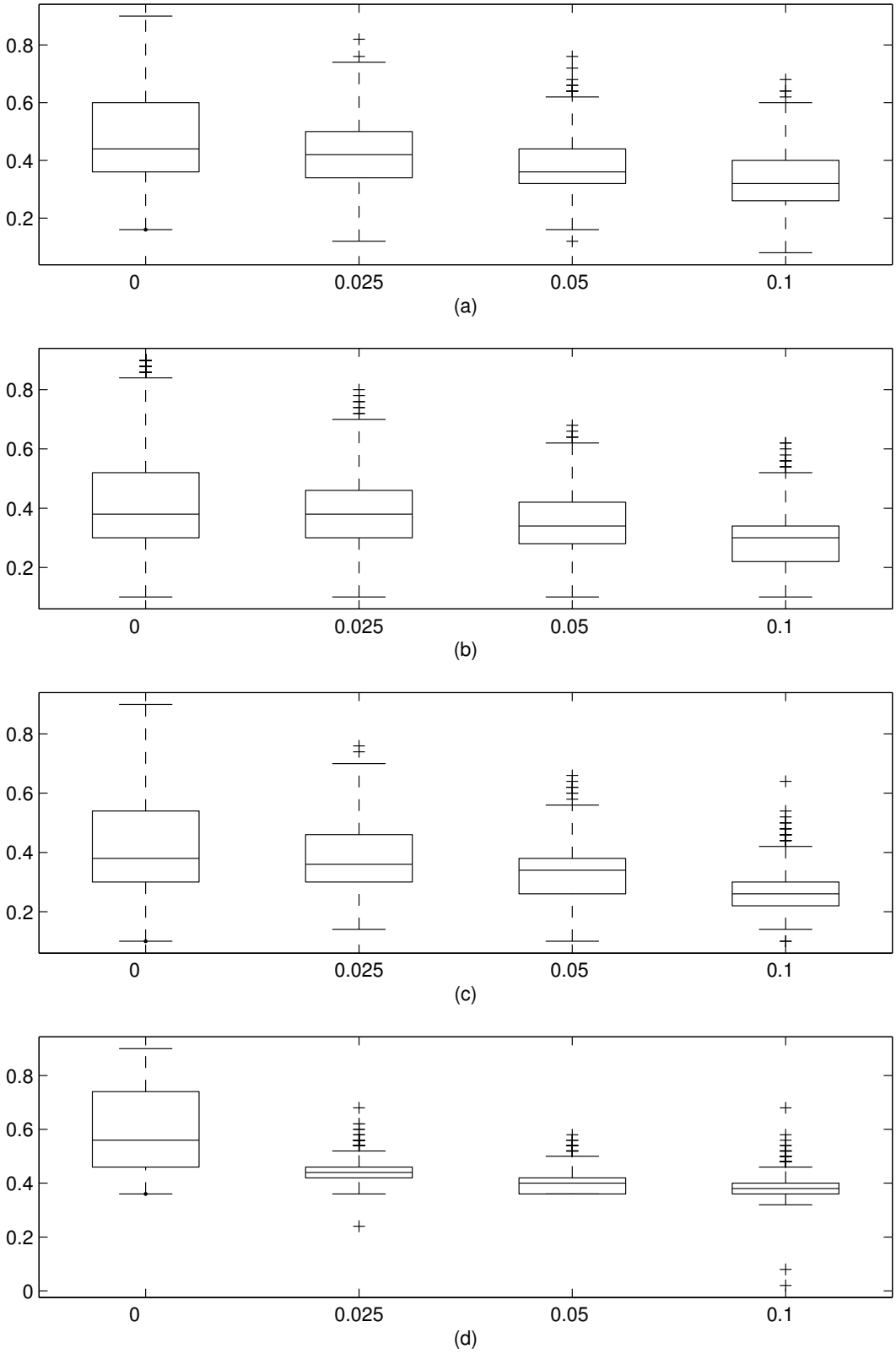
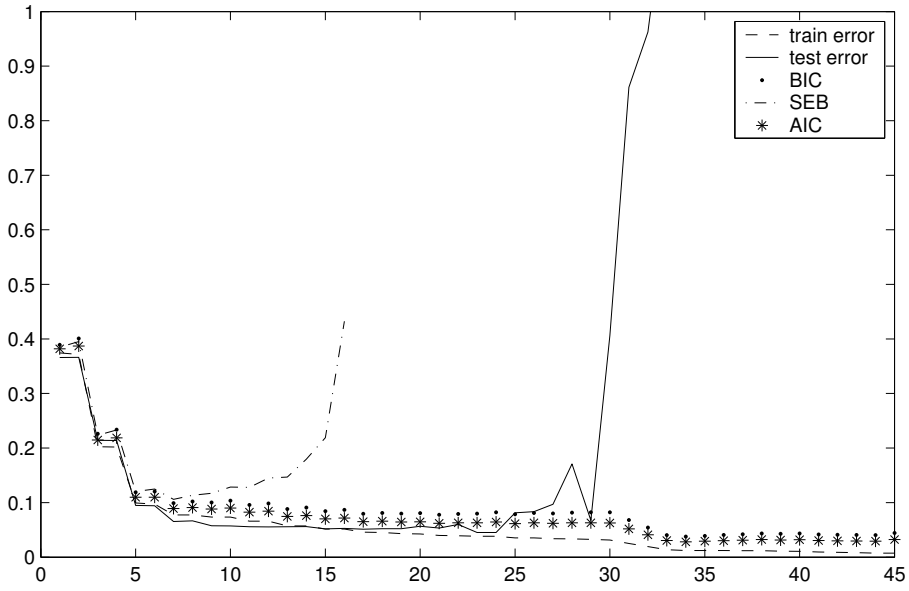
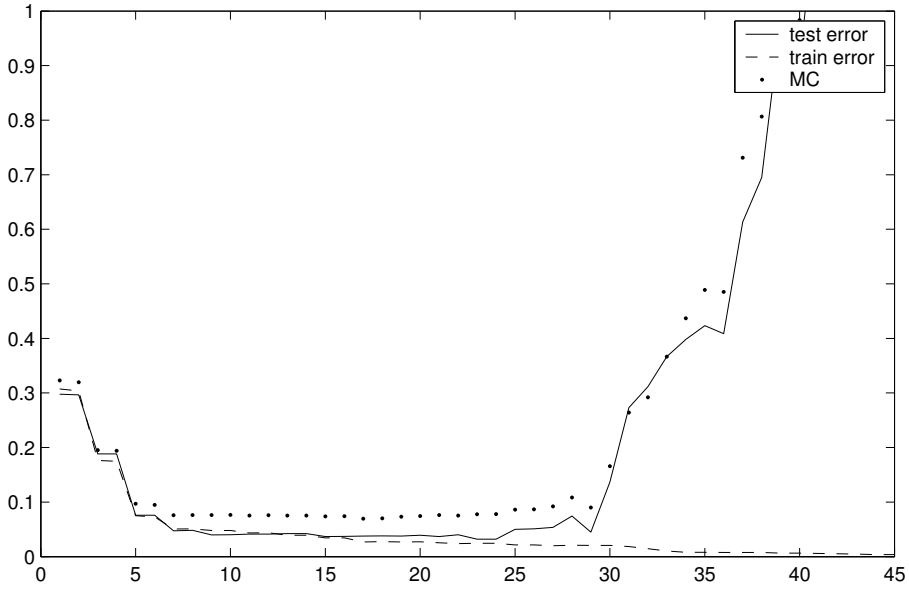



**Fig. 10.** The box-plots for the optimal number of nodes to the number of samples $n^*/l$: (a), (b), (c), and (d) represent the box-plots of $n^*/l$ in the regression of step, sine-square, sinc, and combined functions respectively.

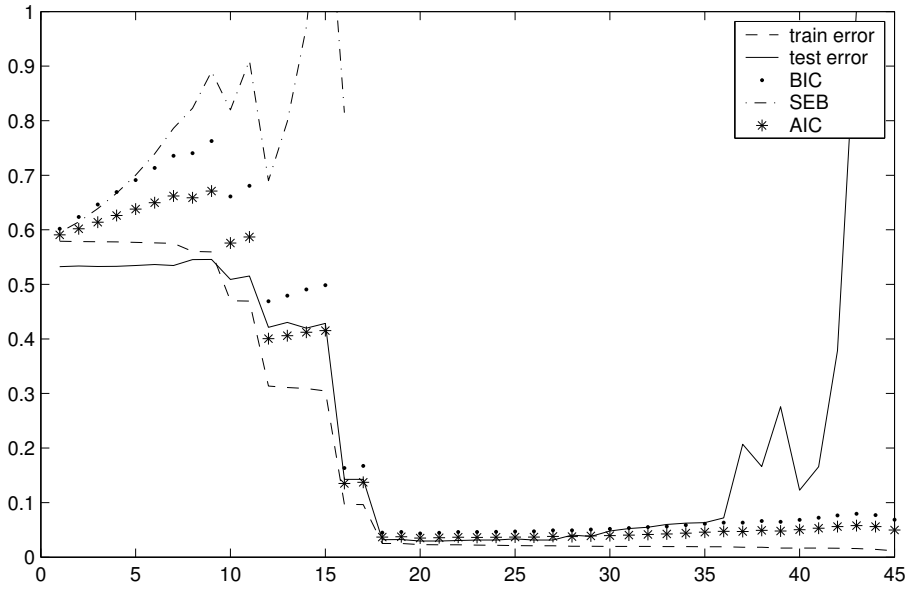



(a)

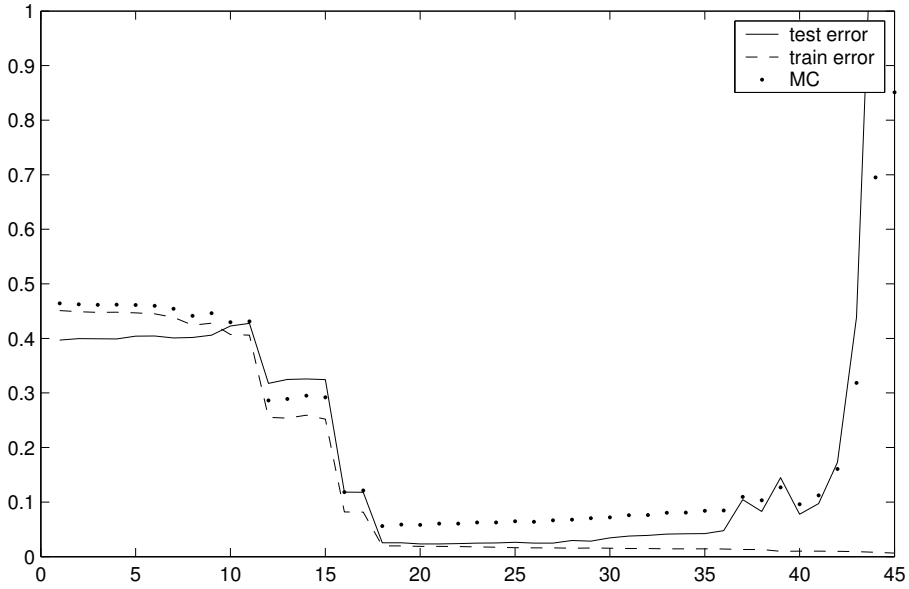



(b)

**Fig. 11.** Performance Predictions of model selection methods in the case of 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.




(a)




(b)

**Fig. 12.** Performance Predictions of model selection methods in the case of 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.

## 5 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 sum of absolute weights as well as the structural information such as the number of nodes in the regression model. To verify the validity of the suggested bound, the model selection of regression models using the trigonometric polynomials in function approximation problem was performed. As a result, the suggested method showed the better performance for various types of target functions from the view points of risk and node ratios compared to other model selection methods. We also demonstrated that the risk prediction using the MC method was able to catch the trend of test errors. This was mainly due to the fact that the MC based method was performed using the risk function bounds incorporating both the information of learned results as well as the structural information of the estimation network, the main criteria of other model selection methods. Furthermore, the suggested MC method can be easily extended to various types of regression models with nonlinear kernel functions which have some smoothness constraints.

# References


1. Akaike, H.: Information theory and an extansion of the maximum likelihood principle. In Petrov, B. and Csaki, F., editors, 2nd Interanational Symposium on Information Theory, VOL. **22** (1973), 267–281, Budapest.
2. Schwartz, G.: Estimating the dimension of a model. Ann. Stat., **6** (1978),461–464.
3. Wahba, G., Golub, G., and Heath, M.: Generalized cross-validation as a method for choosing a good ridge parameter. Technometrics, **21** (1979), 215–223.
4. Rissanen, J.: Stochastic complexity and modeling. Annals of Statistics, **14** (1986), 1080–1100.
5. Barron, A., Rissanen, J., and Yu, B.: The minimum description length principle in coding and modeling. IEEE Transactions on Infomation Theory, **44** (1998), 2743–2760.
6. Dean P. Foster and Edward I. George: The risk inflation criterion for multiple regression. Annals of Statistics, **22** (1994), 1947–1975.
7. Vladimir N. Vapnik: Statistical Learning Theory. J. Wiley, 1998
8. Olivier Chapelle, Vladimir Vapnik, Yoshua Bengio: Model selection for small sample regression. Machine Learning, **48** (2002), 315-333, Kluwer Academic Publishers.
9. Vladimir Cherkassky, Xuhui Shao, Filip M. Mulier, Vladimir N. Vapnik: Model complexity control for regression using VC generalization bounds. IEEE transactions on Neural Networks, VOL. **10**, NO. 5 (1999),1075–1089.
10. G.G. Lorentz: Approximation of functions. Chelsea Publishing Company, New York, 1986.

# Appendix

### A-1. Proof of Theorem 1

The main idea of proving this theorem is to decompose the risk functional $R(f_n)_{L_1}$ into four components: the modulus of continuity for target function $w(f,h)$, the modulus of continuity for approximation function $w(f_n,h)$, additive noise and the empirical risk functional $|y_i - f_n(x_i)|$. First, we will prove the theorem for noiseless target function, that is, $y = f(x)$.

For each $x \in X$, there is $x_i \in X_i,\ i = 1, \cdots, N$, with $|x - x_i| < h$. This implies that

$$\begin{aligned} |f(x) - f_n(x)| &\leqslant |f(x) - f(x_i)| + |f(x_i) - f_n(x_i)| + |f_n(x) - f_n(x_i)| \\ &\leqslant \omega(f,h) + |f(x_i) - f_n(x_i)| + |f_n(x) - f_n(x_i)|. \end{aligned} \tag{37}$$

Let $f_n(x) = \sum_{i=1}^{n} w_i \phi_i(x)$. For each $x, y \in X$ with $|x - y| \leqslant h$, we have

$$\begin{aligned} |f_n(x) - f_n(y)| &= |\sum_{i=1}^{n} w_i(\phi_i(x) - \phi_i(y))| \leqslant \max_{1 \leqslant i \leqslant n} |\phi_i(x) - \phi_i(y)| \sum_{i=1}^{n} |w_i| \\ &\leqslant \max_{1 \leqslant i \leqslant n} \{\omega(\phi_i, h)\} \sum_{i=1}^{n} |w_i|. \end{aligned} \tag{38}$$

That is,

$$|f(x) - f_n(x)| \leqslant \omega(f,h) + |f(x_i) + f_n(x_i)| + \max_{1 \leqslant i \leqslant n} \{\omega(f,h)\} \sum_{i=1}^{n} |w_i| \tag{39}$$

from (37) and (38). By taking the integral of (39) in $X_i$ satisfying $P(X_i) = 1/N$, for $i = 1, \cdots, N$, we get

$$\int_{X_i} |f(x) - f_n(x)| dP(x) \leqslant \frac{1}{N} \left( \omega(f,h) + |f(x_i) - f_n(x_i)| + \max\{\omega(\phi_i, h)\} \sum_{i=1}^{n} |w_i| \right). \tag{40}$$

By taking the integral of (39) in $X$, we get

$$\begin{aligned} R(f_n)_{L_1} &= \int_X |f(x) - f_n(x)| dP(x) = \sum_{i=1}^{N} \int_{X_i} |f(x) - f_n(x)| dP(x) \\ &\leqslant R_{emp}(f_n)_{L_1} + \omega(f,h) + \max\{\omega(\phi_i, h)\} \sum_{i=1}^{n} |w_i|. \end{aligned} \tag{41}$$

Now, let us consider the target function with noise, that is, $y = f(x)+\epsilon$, where $\epsilon$ represents a random variable with mean zero and variance $\sigma^2$. Then, it follows that

$$\begin{aligned} |y - f_n(x)| &\leqslant |y - y_i| + |y_i - f_n(x_i)| + |f_n(x_i) - f_n(x)| \\ &\leqslant w(f,h) + |\epsilon - \epsilon_i| + |y_i - f_n(x_i)| + |f_n(x_i) - f_n(x)|. \end{aligned} \tag{42}$$

Here, $\epsilon_i,\ i = 1, \cdots, N$ are random variables with mean zero and variance $\sigma^2$, We also assume that $\epsilon$ and $\epsilon_i$ are independent. In this case,

$$\begin{aligned} P\{|\epsilon - \epsilon_i| > u\} &\leqslant \frac{\mathbb{E}|\epsilon - \epsilon_i|^2}{u^2} \ \text{ (by Markov's inequality)} \\ &\leqslant \frac{2\sigma^2}{u^2} \end{aligned} \tag{43}$$

Therefore, with the probability at least $1 - \delta$ the following inequality holds:

$$\begin{aligned} R(f_n)_{L_1} = \int_{X\times\mathbb{R}} |y - f_n(x)| dP(x,\epsilon) \leqslant \\ R_{emp}(f_n)_{L_1} + \omega(f,h) + \max\{\omega(\phi_i,h)\} \sum_{i=1}^{n} |w_i| + \sigma\sqrt{\frac{2}{\delta}}. \end{aligned} \tag{44}$$

Q. E. D.