Choice of Bandwidth for Nonparametric Regression Models using Kernel Smoothing: A Simulation Study

Dursun Aydın[a*], Öznur İşçi Güneri[b], Akın Fit[c]

[a,b,c]*Department of Statistics Muğla Sıtkı Koçman University Muğla, 48000 Turkey*
[a]*Email: duaydin@mu.edu.tr*
[b]*Email: oznur.isci@mu.edu.tr*
[c]*Email: statistics48@hotmail.com*

Abstract

In this study, kernel smoothing method is considered in the estimation of nonparametric regression models. A crucial step in the implementation of this method is to select a proper bandwidth (smoothing parameter). In an attempt to address the specification of amount of smoothing, this article provides a comparative study of different methods (or criteria) for choosing the smoothing parameter. Given the need of automatic data-driven smoothing parameter selectors for applied statistics, this study is focused to explain and compare these methods. In this context, we generalized the selection methods used in the smoothing spline method for kernel smoothing. In order to explore and compare the performance of these methods, a simulation study is performed for data sets with different sample sizes. As a result of simulation, the appropriate selection criteria are provided for a suitable smoothing parameter selection.

Keywords: Kernel smoothing; Smoothing spline; Nonparametric regression; Bandwidth; Selection method.

----

* Corresponding author.

1. Introduction

Kernel smoothing is one of the most popular methods used for the estimation of the nonparametric regression models. The most important matter in the implementation of this method is to specify an appropriate "bandwidth (λ)". A central question in this article is how to choose the optimum bandwidth λ for nonparametric regression using Kernel smoothing. In this article, the amount of smoothing is determined by means of bandwidth selection methods, Cross-validation (CV), generalized cross-validation (GCV), improved version of Akaike information criterion (AIC$_c$), Mallows'Cp, risk estimation using classical pilots (RECP), restricted maximum likelihood (REML), Rice's T and a model selector of Shibata (SH). The essential purpose of this article is to provide a comparative study for the performance of eight selection methods. To accomplish this, simulation experiments were conducted to find out which selection methods have good performance about smoothing parameter selection. Turbo C program is used for the realization of this simulation study. Mean squares error (MSE) is taken as a criterion of the performance measures for assessing the quality of the kernel smoothing estimator.

This paper is mainly concerned with the selection of bandwidth (or penalty parameter) through Monte Carlo simulation study. Bandwidth parameters play a crucial role in this procedure. These parameters are said to control the tradeoff between fidelity to data and smoothness: too low values of bandwidth parameter overfit the data, whereas too high values oversmooth. In the literature, different selection methods are components of various studies for an appropriate smoothing parameter. Indeed, to a considerable extent, [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18] supplement on the selection of the smoothing parameter.

The rest of this article is organized as follows. The second section presents the kernel smoothing method for nonparametric regression. The section 3 discusses the CV, GCV, AIC$_c$), Mallows Cp, RECP, REML, Rice's T and SH methods for selecting the penalty parameter. Section 4 indicates the Monte Carlo simulation experiments and Section 5 gives the conclusions and recommendations.

2. Kernel Smoothing for Nonparametric Regression

Nonparametric regression model including a predictor variable and a response variable is defined as

$$y_i = f(x_i) + \varepsilon_i \tag{1}$$

where y_i's are observation values of the response variable, x_i's are observation values of the predictor variable, ε_i's are independent error terms with zero mean and common variance σ^2, and f is an unknown smooth function.

Generally speaking a kernel smoothing defines a set of weights $\{w_i(x), i = 1, 2, ..., n\}$ for each x and defines

$$\hat{f}(x) = \sum_{i=1}^{n} w_i(x) y_i \tag{2}$$

The function $\hat{f}(x)$ in equation (2) is also called as 'linear smoother' for appropriately defined weights $w_i(x)$. This linear representation leads to many nice statistical and computational properties. Also, because we think that points that are close together are similar, a kernel smoothing usually defines weights that decrease in a smooth fashion as one moves away from the target point.

Nadaraya and Watson estimator is considered in kernel smoothing. In this method, a x point is found in description area of the average function $f(x)$ and determined a smoothing window around of this point. The most commonly used smoothing window is displayed as simply the $(x - \lambda, x + \lambda)$ range. Kernel estimate is a weighted mean of the observations in the smoothing window, and is indicated as follows:

$$\hat{f}_\lambda(x) = \sum_{i=1}^{n} K\left(\frac{x_i - x}{\lambda}\right) y_i \Big/ \sum_{j=1}^{n} K\left(\frac{x_j - x}{\lambda}\right) \tag{3}$$

where λ corresponds to the radius of smoothing window and known as a smoothing parameter, n is the number of observations and K is the selected kernel function. Kernel smoothing can be displayed in the following way [19]:

$$\hat{f}_\lambda(x) = \sum_{i=1}^{n} \left(K\left(\frac{x_i - x}{\lambda}\right) \Big/ \sum_{j=1}^{n} K\left(\frac{x_j - x}{\lambda}\right) \right) y_i = \sum_{i=1}^{n} w_i(x) y_i \tag{4}$$

where the weights $w_i(x)$ depend on the $x - x_i$ distance and assigned to y_i from $i.$ observation in equation (4), so the weight w of $i.$ observation; $x - x_i$ as defined a function of $x - x_i$ distance. Regression function is estimated in x point as a weight average of dependent variable y_i with the help of weights $w_i(x)$ depending on x and coefficients $w_i(x)$ as follows

$$w_i(x) = K\left(\frac{x_i - x}{\lambda}\right) \Big/ \sum_{j=1}^{n} K\left(\frac{x_j - x}{\lambda}\right) = K(u) \Big/ \sum_{j=1}^{n} K(u) \tag{5}$$

Notice that $\sum_{i=1}^{n} w_i(x) = 1$. The kernel is a continuous, bounded and symmetric real valued function K which satisfies the following characteristics:

- $K(u) \geq 0 \quad \text{for all values of } u$

- $\int_{-\infty}^{+\infty} K(u) du = 1$

- $K(-u) = K(u)$

These features are also characteristics of a symmetric probability density function. Some of the kernel functions used in the application are provided in Table 1.

Given the all functions provide the properties above of the kernel functions in Table 1. Selection of the kernel function is less important than selection of smoothing parameter [6].

Table 1: Alternative kernel functions

Kernel	Explicit Form
Gaussian	$K(u) = \dfrac{1}{\sqrt{2\pi}} \exp(-\dfrac{1}{2}u^2), u \in [-\infty, \infty]$
Uniform	$K(u) = \dfrac{1}{2}, \ u \in [-1,1]$
Triangular	$K(u) = (1-\lvert u \rvert), \ u \in [-1,1]$
Epanechnikov	$K(u) = \dfrac{3}{4}(1-u^2), \ u \in [-1,1]$
Quartic	$K(u) = \dfrac{15}{16}(1-u^2)^2, \ u \in [-1,1]$
Triweight	$K(u) = \dfrac{35}{32}(1-u^2)^3, \ u \in [-1,1]$

In addition, the matrix and vector form of the model (6), $\hat{\mathbf{f}} = \mathbf{W}_\lambda \mathbf{y}$ is given by

$$\mathbf{W} = \begin{bmatrix} \boldsymbol{w}_1' \\ \vdots \\ \boldsymbol{w}_i' \\ \vdots \\ \boldsymbol{w}_n' \end{bmatrix} = \begin{bmatrix} w_{11} & \cdots & w_{1n} \\ \vdots & \cdots & \vdots \\ w_{i1} & \cdots & w_{in} \\ \vdots & \cdots & \vdots \\ w_{n1} & \cdots & w_{nn} \end{bmatrix} \tag{6}$$

The Kernel estimate of the function f in model (1) for any point x_i is expressed as

$$\hat{y}_i = \hat{f}(x_i) = \sum_{j=1}^{n} w_{ij} y_j = \boldsymbol{w}_j' y , \ i = 1, 2, ..., n \tag{7}$$

The kernel estimation of the non-parametric regression expressed in model (1) is also given in the following

way [16]:

$$\hat{\mathbf{f}} = \begin{bmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix}_{n \times 1} = \begin{bmatrix} w_{11} & \cdots & w_{1n} \\ \vdots & \ddots & \vdots \\ w_{n1} & \cdots & w_{nn} \end{bmatrix}_{n \times n} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}_{n \times 1} = \mathbf{W}_\lambda \, \mathbf{y} \qquad (8)$$

3. The Bandwidth Selection Methods

Selecting a proper bandwidth λ is a crucial step in estimating $f(x)$. As λ varies from 0 to $+\infty$, the solution varies from interpolation to a linear model [20]. Thus, the parameter λ controls the smoothness of the function estimate and greatly affects its appearance. Also, the parameter λ provides a compromise between variance and bias of $\hat{f}_\lambda(x)$. Each data set with the same λ value is not expected to work equally well [8]. Therefore, smoothing parameter $\lambda > 0$ has a major impact on the quality of $\hat{f}_\lambda(x)$. In this article, the different selection methods have been introduced for the selection of the a proper λ.

In this paper, it is discussed various bandwidth selection methods in the literature. Most of these methods are implemented in our simulation study. The positive value λ that minimizes any selection methods is selected as an appropriate smoothing parameter. The selection criteria used in our simulation study are classified as follows:

Cross Validation (CV): The basic idea of CV is to leave the points $\{x_i, y_i\}_{i=1}^{n}$ out one at a time and to select the smoothing parameter λ that minimizes the residual sum of squares and to estimate squared residual for a smooth function at x_i based on the remaining (n-1) points. The CV score function to be minimized is given by

$$CV(\lambda) = n^{-1} \sum_{i=1}^{n} \left\{ y_i - \hat{f}_i^{-i}(x_i) \right\}^2 = n^{-1} \sum_{i=1}^{n} \left\{ \frac{y_i - \hat{f}_\lambda(x_i)}{1 - \mathbf{W}_{ii}} \right\}^2 \qquad (9)$$

where $\hat{f}_\lambda$ is the fit (kernel smoother) for n pairs of measurements $\{x_i, y_i\}_{i=1}^{n}$ with smoothing parameter λ, $\hat{f}_\lambda^{(-i)}$ is the fit calculated by leaving out the ith data point and $\mathbf{W}_{ii}$ is the ith diagonal element of smoother matrix $\mathbf{W}$ in equation (6).

Generalized Cross Validation (GCV): GCV is a modified form of the CV which is a popular criterion for choosing the smoothing parameter. The GCV score is constructed by analogy to CV score obtained from dividing to the factors $1 - \mathbf{W}_{ii}$ of the ordinary residuals. The main idea of GCV is to replace the factors $1 - \mathbf{W}_{ii}$ in equation (9) with the average score $1 - n^{-1} tr(\mathbf{W}_\lambda)$. Thus, by summing of the squared residual corrected and factor $\left\{ 1 - n^{-1} tr(\mathbf{W}_\lambda) \right\}^2$, by the anology ordinary cross validation, the GCV score function is

obtained as fallow [14].

$$Cp(\lambda) = \frac{1}{n}\left\{\left\|(\mathbf{W}_\lambda - I)y\right\|^2 + 2\sigma^2 tr(\mathbf{W}_\lambda) + \sigma^2\right\} = \frac{1}{n}\left\{\left\|y - \hat{f}_\lambda\right\|^2 + 2\sigma^2 tr(\mathbf{W}_\lambda) + \sigma^2\right\} \qquad (10)$$

Mallows'Cp Criterion: This criterion is known as unbiased risk estimate (UBR) in smoothing spline literature. This type of estimate was suggested by [21] in regression case, and applied to smoothing spline by [1]. When σ^2 is known, an unbiased estimate of the residual sum of squares is given by Cp criterion:

$$Cp(\lambda) = \frac{1}{n}\left\{\left\|(\mathbf{W}_\lambda - I)y\right\|^2 + 2\sigma^2 tr(\mathbf{W}_\lambda) + \sigma^2\right\} = \frac{1}{n}\left\{\left\|y - \hat{f}_\lambda\right\|^2 + 2\sigma^2 tr(\mathbf{W}_\lambda) + \sigma^2\right\} \qquad (11)$$

Unless σ^2 is known, in practise an estimate for σ^2 is estimated by

$$\hat{\sigma}_\lambda^2 = \sum_{i=1}^{n}\left\{y_i - \hat{f}(x_i)\right\}^2 \bigg/ tr(I - \mathbf{W}_\lambda) = \left\|(\mathbf{W}_\lambda - I)y\right\|^2 \bigg/ tr(I - \mathbf{W}_\lambda) \qquad (12)$$

where $\hat{\lambda}$ is pre-chosen with any of the CV, GCV or AIC_c criteria ($\hat{\lambda}$ is an estimate of λ) [13].

Improved Version of Akaike Information (AIC_c): An improved version of a criterion based on the classical Akaike information criterion (AIC), AIC_c criterion, is used for choosing the smoothing parameter for nonparametric smoothers [7]. This improved criterion is defined as

$$AIC_c(\lambda) = \log\frac{\left\|(\mathbf{W}_\lambda - I)y\right\|^2}{n} + 1 + \frac{2\{tr(\mathbf{W}_\lambda) + 1\}}{n - tr(\mathbf{W}_\lambda) - 2} \qquad (13)$$

As can be seen from the equation (13), this criterion is easy to apply for choosing of smoothing parameter.

Risk Estimation Using Classical Pilots (RECP): Risk function measures the distance between the actual regression function (f) and its estimation ($\hat{f}_\lambda$). Actually, a good estimate must contain minumum risk. A direct computation leads to the the bias-variance decomposition for $R(f,\hat{f}_\lambda)$:

$$R(f,\hat{f}_\lambda) = \frac{1}{n}E\left\|f - \hat{f}_\lambda\right\|^2 = \frac{1}{n}\left\{\left\|(\mathbf{W}_\lambda - I)f\right\|^2 + \sigma^2 tr(\mathbf{W}_\lambda \mathbf{W}'_\lambda)\right\} \qquad (14)$$

It is straight forward to show that $R(f,\hat{f}_\lambda) = E\left\{C_p(\lambda)\right\}$. Because the risk $R(f,\hat{f}_\lambda)$ is unknown quantity, so-called risk is now estimated by computable quantity $R(\hat{f}_{\lambda_p},\hat{f}_\lambda)$. The obtained expression for $R(\hat{f}_{\lambda_p},\hat{f}_\lambda)$ is

$$R(\hat{f}_{\lambda_p}, \hat{f}_{\lambda}) = \frac{1}{n} E \left\| \hat{f}_{\lambda_p} - \hat{f}_{\lambda} \right\|^2 = \frac{1}{n} \left\{ \left\| (\mathbf{W}_\lambda - I)\hat{f}_{\lambda_p} \right\|^2 + \sigma_{\lambda_p}^2 tr(\mathbf{W}_\lambda \mathbf{W}'_\lambda) \right\} \tag{15}$$

where $\hat{\sigma}_{\lambda_p}^2$ and $\hat{f}_{\lambda_p}$ are the appropriate pilot estimates for σ^2 and f, respectively. The pilot λ_p selected by classical methods is used for computation of the pilot estimates [9, 12].

Restricted Maximum Likelihood (REML): The REML method proceeds by optimizing a function of λ. When λ is estimated by REML, the REML error variance estimate agrees with the smoothing theoretic variance estimate. The REML score can be expressed as

$$REML(\lambda) = \left\| (\mathbf{W}_\lambda - I)y \right\|^2 \Big/ tr(I - \mathbf{W}_\lambda) \tag{16}$$

The rigth term of Equation (16) equals $\left\| \mathbf{y} - \hat{\mathbf{y}} \right\| / n - tr(\mathbf{W}_\lambda)$, estimate of σ^2 that is based on viewing $tr(\mathbf{W}_\lambda)$ as the degrees of freedom of the smoother freedom of the smoother [22].

The REML and GCV, viewed as functions of λ, share a similar form and yield identical values. Also, they can be presented within a common framework that reveals some interesting connections between them. More specifically, the derivatives of GCV and REML criteria with respect to λ can be determined quite naturally in a common form [18].

Rice's T Criterion: On the basis of other criteria that are not very popular as a selection criterion in the literature. Rice's T criterion is defined as follows

$$T(\lambda) = \frac{1}{n} \left\| (I - \mathbf{W}_\lambda)y \right\|^2 / (1 - n^{-1} 2tr(\mathbf{W}_\lambda)) \tag{17}$$

The positive value λ that minimizes the equation (17) is selected as smoothing parameter.

A Model Selector of Shibata (SH) Criterion: This model is very popular in the literature, such as the above criterion T. Addition to other selection criteria gives poor results. Shibata criterion is expressed by

$$SH(\lambda) = \frac{1}{n} \left\| (I - \mathbf{W}_\lambda)y \right\|^2 (1 - n^{-1} 2tr(\mathbf{W}_\lambda)) \tag{18}$$

The positive value λ that minimizes expression $SH(\lambda)$ is selected as smoothing parameter.

4. Monte Carlo Simulation Study

This section reports the results of a Monte Carlo simulation study. This study is conducted to evaluate the

performances of the eight selection methods. By using TURBO C program, we generated the samples sized n = 50, 75 and 100. The number of replications was 500 for each of the samples. According to selection criteria, $\hat{f}$ functions are estimated by using the 'kernel smoothing' based on the Gaussian kernel function. In order to evaluate $\hat{\mathbf{f}}_\lambda$ computed according to each of the selection criterion we used the MSE given by

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left\{f(x_i) - \hat{f}_\lambda(x_i)\right\}^2 \text{ and } \hat{f}_\lambda(x_i) = (\hat{\mathbf{f}}_\lambda)_i) \tag{19}$$

where $f(x_i)$ is value at knots x_i of the appropriate function f defined in Table 2.

To find out if the difference between the MSE median values of any two selection methods is significant or not, the paired Wilcoxon tests were assessed. In this way, methods which complement the best smoothing parameter were determined by evaluating so-called selection methods.

4.1. Creating data and the experimental setup

The data sets used for sampling in the simulation experiment are obtained from models in Table 2. The simulation study is performed by TURBO C program and through SPSS (2010) program [23]. The experimental setup is designed as follows:

The experimental setup applied at this stage was designed to study the effects of four regression functions under the noise level factor. Thus four different cases are considered in the simulation experiments. The Monte Carlo simulations also examine the performance of the selection criteria as they relate to the sample size, the pattern of predictor values, the true regression function and the true standard deviation of the errors.

According to the each of selection criteria, MSE values are calculated by considering kernel estimators. Paired Wilcoxon test is applied to test whether MSE values considered as the performance measure of any two methods are significant or not. To see the performance of the selection methods for each set of experiments, factor level is changed six times (r = 1, 2, 3, 4, 5, 6).

Table 2: Simulation setup

Case	General Form	Regression Function
1	$y_{ir} = f_1(x_i) + \sigma_r \varepsilon_i$	$f_1(x) = \sin(15\pi x)$
2	$y_{ir} = f_2(x_i) + \sigma_r \varepsilon_i$	$f_2(x) = 0.3\exp\left\{-64(x-0.25)^2\right\} + 0.7\exp\left\{-256(x-0.75)^2\right\}$
3	$y_{ir} = f_3(x_i) + \sigma_r \varepsilon_i$	$f_3(x) = 10\exp(-10x)$
4	$y_{ir} = f_4(x_i) + \sigma_r \varepsilon_i$	$f_4(x) = \sqrt{x(1-x)}\sin\left[\dfrac{2\pi(1+2^{(9-4r)/5})}{x+2^{(9-4r)/5}}\right]$
	$\sigma_r = 0.02 + 0.04(r-1)^2, \ r = 1,2,3,4,5,6$ and $x \sim U(0,1), \ \varepsilon_i \sim N(0,1), \ i = 1,...,n$	

4.2. Experimental evaluations

For each simulated data set paired Wilcoxon tests were applied to test whether the difference between the median MSE values of any two methods is significant or not. The significance level used was 5 %. The selection methods were also ranked as follows: If median MSE value of a method is significantly less than the remaining five, it will be assigned a rank 1. If median MSE value of a method is significantly larger than one but less than the remaining four, it will be assigned to a rank 2, and similarly to ranks 3-8. Methods having non-significantly different median values will share the same averaged rank, on the other hand, method or methods having the smallest rank will be superior.

Table 3: Wilcoxon signed ranks test results for n=100, case 3 and r = 1

	GCV-CV	C_P-CV	AIC-CV	SH-CV	T-CV	ML-CV	RECP-CV
Z	0.00 (a)	0.00 (a)	0.00 (a)	-1.34(b)	0.00 (a)	0.00 (a)	0.00 (a)
P –value	1.00	1.00	1.00	0.18	1.00	1.00	1.00
	C_P-GCV	AIC-GCV	SH-GCV	T-GCV	ML-GCV	RECP-GCV	AIC-C_P
Z	0.00 (a)	0.00 (a)	-1.34 (b)	0.00 (a)	0.00 (a)	0.00 (a)	0.00 (a)
P value	1.00	1.00	0.18	1.00	1.00	1.00	1.00
	SH- C_P	T- C_P	ML- C_P	RECP- C_P	SH-AIC	T-AIC	ML-AIC
Z	-1.34 (a)	0.00 (b)	0.00 (b)	0.00 (b)	-1.34 (a)	0.00 (b)	0.00 (b)
P -value	0.18	1.00	1.00	1.00	1.00	1.00	1.00
	RECP-AIC	T-SH	ML-SH	RECP-SH	ML-T	RECP-T	RECP-ML
Z	0.00 (a)	-1.34 (b)	-1.34 (b)	-1.34 (b)	0.00 (a)	0.00 (a)	0.00 (a)
P -value	1.00	0.18	0.18	0.18	1.00	1.00	1.00

a.   The sum of negative ranks equals the sum of positive ranks.

b.   Based on positive ranks

Table 3 shows the paired Wilcoxon test results obtained from 500 repeated simulation experiments for n = 100, case 3 and r = 1. In addition, for n = 100, case 3 and r = 5 Wilcoxon test results obtained from the same simulation experiments are indicated in Table 4.

These binary comparisons are performed by using MSE median values for each case and each factor level. There are twenty-eight binary comparisons for each case and factor level (see Table 3). Thus, since a case consists of the six factor levels, binary comparisons carried out are 28x6 = 168. Total binary comparisons performed are 4x168 = 672 for four cases. This number is also equivalence to the number of experiments which are carried out in simulation. These binary comparisons are tested by the hypotheses,

$H_0 : M = 0$ and $H_1 : M \neq 0$ (20)

The hypotheses expressed in (20) are finalized by p-values.

Table 4: Wilcoxon signed ranks test results for n=100, case 3 and r = 5

	GCV-CV	C$_P$-CV	AIC-CV	SH-CV	T-CV	ML-CV	RECP-CV
Z	-1.68 (a)	-5.64 (a)	-8.33 (a)	-8.68 (a)	-8.15 (a)	-8.68 (b)	-8.68 (b)
P -value	0.09	0.00	0.00	0.00	0.00	0.00	1.00
	C$_P$-GCV	AIC-GCV	SH-GCV	T-GCV	ML-GCV	RECP-GCV	AIC-C$_P$
Z	-5.37 (a)	-8.374 (a)	-8.68 (a)	-8.10 (a)	-8.68 (b)	-8.68 (b)	-6.51 (a)
P value	0.000	0.000	0.00	0.00	0.00	0.00	0.00
	SH-C$_P$	T-C$_P$	ML-C$_P$	RECP-C$_P$	SH-AIC	T-AIC	ML-AIC
Z	-8.68 (a)	-6.09 (a)	-8.68 (b)	-8.68 (b)	-8.68 (a)	-2.36 (b)	-8.68 (b)
P -value	0.00	0.00	0.00	0.00	0.00	0.02	0.00
	RECP-AIC	T-SH	ML-SH	RECP-SH	ML-T	RECP-T	RECP-ML
Z	-8.68 (a)	-8.68 (a)	-8.68 (a)	-8.68(a)	-8.68 (a)	-8.68 (a)	-8.68 (a)
P -value	0.00	0.00	0.00	0.00	0.00	0.00	0.00

a.   The sum of negative ranks equals the sum of positive ranks.

b.   Based on positive ranks

As can be seen from Table 2,   totally, 72 numerical experiments (or configurations) are conducted by considering 4 cases, 6 factor levels and 3 samples, For this reason, it is not possible to display here all these configurations. Therefore, some configurations are given in Figures 1-3 for different samples sized n.
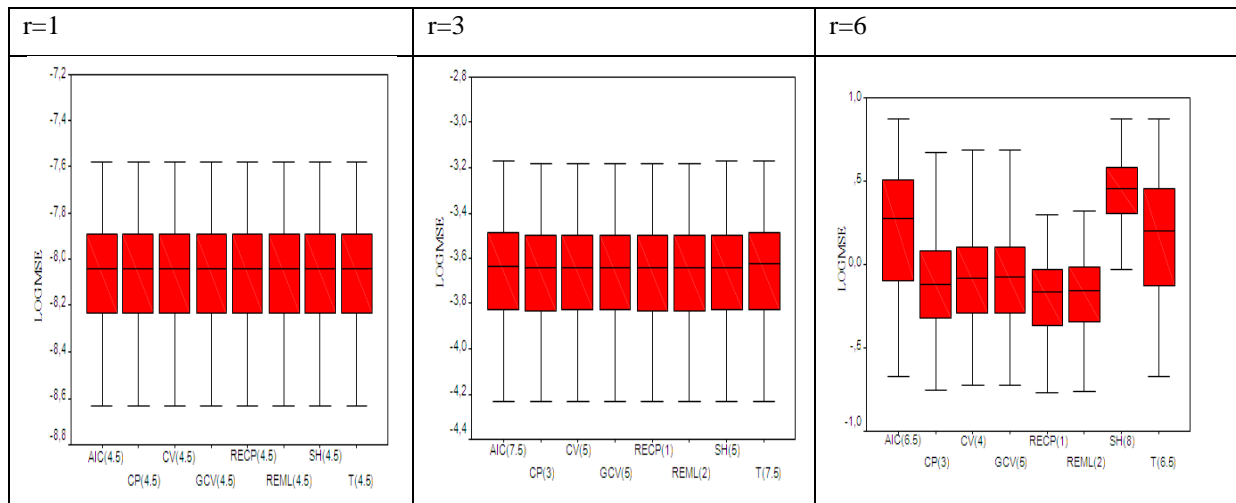

Figure 1: Simulation results correspond to case 1 for n = 50

These Figures display the boxplots of the   values for, from left to right, AIC, CP, CV, GCV, RECP, REML, SH and T. The numbers below the boxplots are the paired Wilcoxon test rankings. Comparison results presented in

Figures 1-3 indicate that RECP maintains its superiority over the other selection methods. Also, for 72 different simulation experiments, the averaged ranking values of the selection methods according to Wilcoxon tests are tabulated in Tables 3 and 4.
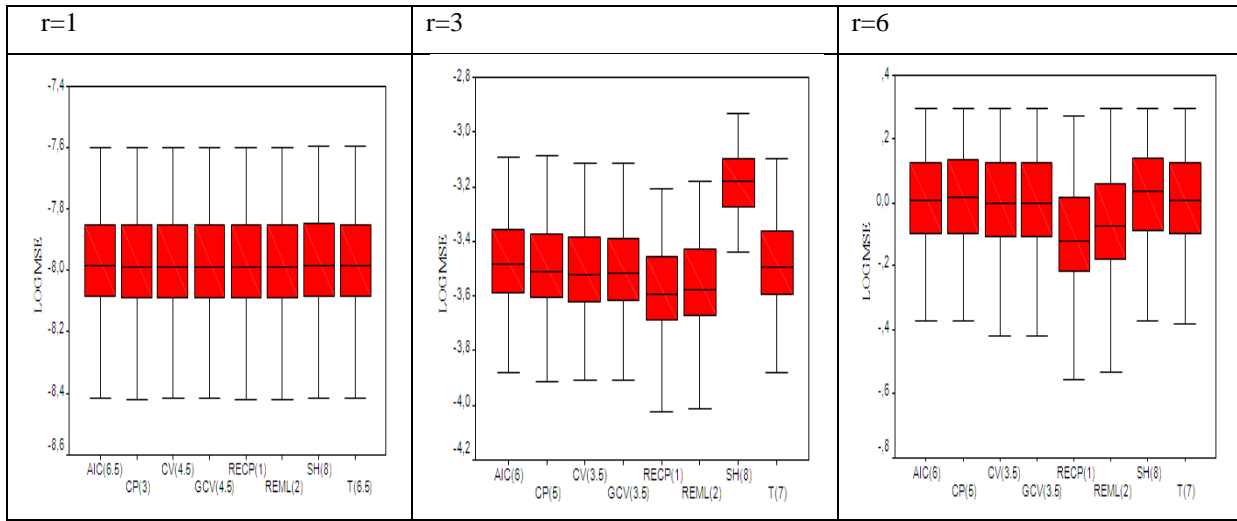

Figure 2: Simulation results correspond to case 2 for n =75

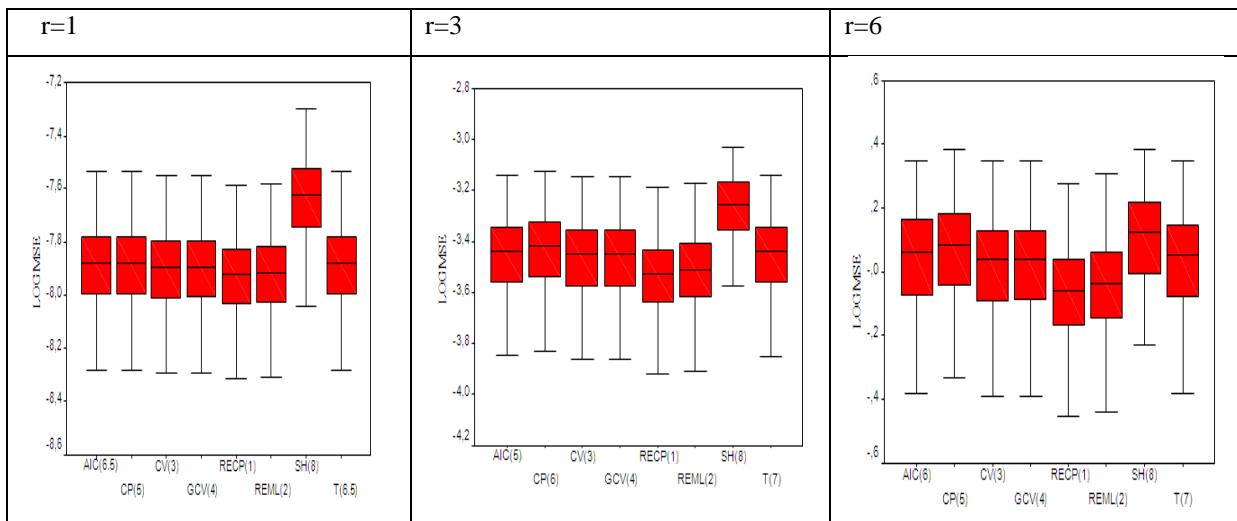**Figure 3:** Simulation results correspond to case 4 for n =100

According to the results in Table 5, for small sized samples (for n=50), RECP has had the best empirical performance for all cases. Furthermore, REML and Cp have shared a better performance after RECP criterion. In accordance with the overall Wilcoxon test rankings in Table 5, RECP and REML have also displayed a good performance. As shown in Table 5, AIC_c, SH and T methods produced the similar results under all experimental cases. For small samples, it is observed that T has produced the worst performance.

According to the results in Table 5, for medium sized samples (for n=75), RECP has also demonstrated the best empirical performance for all cases. Furthermore, REML and CV have shared a better performance after RECP criterion. In accordance with the overall Wilcoxon test rankings in Table 5, RECP has also displayed a highly

good performance. As shown in Table 5, AIC_c, and SH methods produced the similar results under all experimental cases. For small samples, it is observed that SH has produced the worst performance.

Table 5: Averaged Wilcoxon test ranking values for the eight selection methods

Criteria	Case 1	Case 2	Case 3	Case 4	Overall
For n=50					
CV	4.333	3.417	4.250	3.333	3.830
GCV	5.250	4.250	4.583	4.417	4.620
Cp	3.083	4.333	3.000	4.250	3.670
AIC_c	6.000	6.750	6.167	6.833	6.440
SH	6.833	6.750	6.667	8.000	7.060
T	6.333	6.500	7.333	6.167	6.580
REML	2.417	2.250	2.333	2.000	2.250
RECP	1.750*	1.417*	1.667*	1.000*	1.460*
For n=75					
CV	3.750	4.500	4.167	3.333	3.940
GCV	3.750	4.500	4.167	3.833	4.060
Cp	4.750	3.000	3.500	5.500	4.190
AIC_c	5.750	6.500	6.000	5.333	5.890
SH	7.417	8.000	7.417	8.000	7.710
T	5.750	6.500	6.333	7.000	6.390
REML	2.750	2.000	2.833	2.000	2.390
RECP	2.083*	1.000*	1.583*	1.000*	1.420*
For n=100					
CV	3.750	3.167	3.750	3.000	3.420
GCV	3.976	4.000	3.917	4.000	3.960
Cp	4.917	5.500	4.667	5.250	5.080
AIC_c	5.583	5.500	5.917	5.830	5.710
SH	7.417	8.000	7.417	8.000	7.710
T	5.583	6.833	6.333	6.917	6.420
REML	2.750	2.000	2.417	2.000	2.290
RECP	2.083*	1.000*	1.583*	1.000*	1.420*

(*): Indicates the selection methods having the best rankings

According to Table 5, for large sized samples (for n=100), RECP criterion has had the best empirical performance. Generally it is shown that REML, GCV and CV criteria have shared a good performance after RECP. According to the overall Wilcoxon test rankings in Table 5, RECP, REML, GCV and CV criteria can be ranked in terms of the performance. As shown in Table 5, generally, CV and GCV gave the similar results. SH

has also produced the worst performance for large samples.

5. Conclusions and Recommendations

The Monte Carlo simulation results confirmed that RECP has a good performance and can give slightly better estimates than the other selection methods for nonparametric models. The RECP method also outperforms the others in terms of the observed MSE. Generally, SH and T methods produced not good performance in all factors. The scores in Table 6 are obtained by taking the means of the averaged Wilcoxon test ranking values tie with each of the selection methods in Table 5.

Table 6: Means of the averaged Wilcoxon test ranking values for the selection methods

Criteria	Case 1	Case 2	Case 3	Case 4	Average
CV	3.940	3.694	4.055	3.222	3.727
GCV	4.305	4.250	4.222	4.083	4.215
Cp	4.249	4.277	3.722	5.000	4.312
AIC_c	5.777	6.250	6.027	5.998	6.013
SH	7.222	7.583	7.166	8.000	7.492
T	5.888	6.611	6.666	6.694	6.464
REML	2.638	2.083	2.527	2.000	2.312
RECP	1.972*	1.138*	1.611*	1.000*	1.430*

(*): Indicates the selection methods having the best rankings

Finally, by considering the simulation results and evaluations given above, the following suggestions have to be taken into account:

- For small, medium, large samples, RECP is recommended as being the best selection criterion;
- REML is also recommended as second selection criterion after RECP;
- The two selection methods, CV and GCV criteria have produced very similar results according to all regression functions. Cp criterion has given also similar results with these two criteria for small samples. However, as sample size is increased, Cp is getting a form different from these two criteria;
- SH and T criteria have produced the worst empirical performance for all the simulation experiments.

In this case, according to the above results, we can make the following suggestions: In all regression models and the general means, use the RECP criterion because of its superior empirical performance; otherwise use one of the REML criteria whose empirical performance is very close to RECP. The other four selection criteria, CV, GCV, Cp and AIC_c, give a good performance after the RECP and REML.

References

[1] P. Craven, and G. Wahba. "Smoothing Noisy Data With Spline Functions". Numerische Mathematik,

vol.31, pp.377–403, 1979.

[2] E. Heckman. "Spline Smoothing in a Partly Linear Model", Journal of the Royal Statistical Society. Series B (Methodological), vol.48, pp. 244-248, 1986.

[3] W. Hardle. "Applied Nonparametric Regression". Cambridge University Press, Cambridge, 1991.

[4] W. Hardle, P. Hall, and J.S. Marron. "How Far are Automatically Chosen Regression Smoothing Parameters From Their Optimum? (With Discussion)". Journal of the American Statistical Association, vol. 83, pp.86–89, 1988.

[5] G. Wahba. "Spline models for observation data". SIAM, Pennslylvania, 1990.

[6] T. Hastie, and R. Tibshirani. "Genaralized Additive Models". Chapman & Hall, London, 1990.

[7] C.M. Hurvich, J.S. Simonoff, and C.L., Tsai. "Smoothing parameter selection in nonparametric regression using an improved akaike information criterion". Journal of the Royal Statistical Society. Series B (Methodological), vol. 60, pp.271-293, 1998.

[8] R.L. Eubank. "Nonparametric Regression and Spline Smoothing". New York, 1999.

[9] T.C.M. Lee, and V.Solo. "Bandwidth Selection for Local Linear Regression: A Simulation Study". Computational Statistics, vol. 14, pp. 515–532, 1999.

[10] M.G. Schimek. "Estimation and Inference in Partially Linear Models with Smoothing Splines". Journal of Statistical Planning and Inference, vol. 91, pp. 525-540, 2000.

[11] E. Cantoni, and E. Ronchetti. "Resistant Selection of Smoothing Parameter for Smoothing Splines". Statisics and Computing, vol.11, pp.141–146, 2001.

[12] T.C.M. Lee. "A Stabilized Bandwidth Selection Method for Kernel Smoothing of Periodiagram". Signal Process, vol. 81, pp.419-430, 2001.

[13] S.C. Kou, and B. Efron. "Smoothers and the Cp, generalized maximum likelihood, and extended exponential criteria". Jasa, vol. 97, pp. 766-782, 2002.

[14] D. Ruppert, M. P. Wand, and R. J. Carroll. "Semiparametric Regression". Cambridge University Press, Cambridge, 2003.

[15] T.C.M. Lee. "Smoothing parameter selection for smoothing splines: a simulation study". Computational Statistics & Data Analysis,vol. 42, pp.139-148, 2003.

[16] D. Aydın. "A Comparison of the Nonparametric Regression Models Using Smoothing Spline and

Kernel Regression". International Journal of Mathematical, Computational, Physical, Electrical and Computer Engineering, vol .1, pp. 588-592, 2007.

[17] T. Krivobokova, and G. Kauermann. "A note on penalized spline smoothing with correlated errors". Journal of the American Statistical Association, vol, 102, pp. 1328-1337, 2007.

[18] D. Aydın, and M. Memmedli. "Optimum Smoothing Parameter Selection for Penalized Least Squares in Form of Linear Mixed Effect Models". Optimization: A Journal of Mathematical Programming and Operations Research, vol. 61(4), pp. 459-476, 2012.

[19] C. Loader. "Smoothing: Local Regression Techniques". Papers / Humboldt-Universität Berlin, Center for Applied Statistics and Economics (CASE), No.12, 2004.

[20] W. Hardle, M. Muller, S. Sperlich, and A. Werwatz. "Nonparametric and Semiparametric Models", Springer, New York, 305, 2004.

[21] C.L. Mallows. "Some comments on CP". Technometrics, vol.15, pp. 661-675, 1973.

[22] P.T. Reis, and R.T. Ogden. "Smoothing Parameter Selection for A Class of Semiparametric Linear Models". Journal of the Royal Statistical Society. Series B (Methodological), vol.71, pp. 505–523, 2009.

[23] SPSS. "SPSS for Windows". Version 16.0, Chicago, SPSS Inc.,USA, 2010.