EUROPEAN JOURNAL OF PURE AND APPLIED MATHEMATICS
Vol. 6, No. 2, 2013, 222-238
ISSN 1307-5543 – www.ejpam.com

Smoothing Parameter Selection for Nonparametric Regression Using Smoothing Spline

Dursun Aydin [1,*], Memmedaga Memmedli[2], Rabia Ece Omay[2]

[1] *Department of Statistics, Mugla Sitki Kocman University 48000 Mugla, Turkey*
[2] *Department of Statistics, Anadolu University 26470 Eskisehir, Turkey*

Abstract. In this paper, the smoothing parameter selection problem has been examined in respect to a smoothing spline implementation in predicting nonparametric regression models. For this purpose, a simulation study has been performed by using a program written in MATLAB. The simulation study provides a comparison of the nine smoothing parameter selection methods. In this connection, 500 replications have been performed in simulation for sample sets with different sizes. Thus, the appropriate selection criteria are provided for a suitable smoothing parameter selection.

2010 Mathematics Subject Classifications: 62E17, 65C05, 65C20, 65C60
Key Words and Phrases: Nonparametric regression, Smoothing spline, Smoothing parameter, Selection criteria, Generalized cross validation.

1. Introduction

Smoothing spline method is one of the most popular methods used for the prediction of the nonparametric regression models. The role of this method is to estimate the nonparametric function that minimizes penalized least squares criterion. A roughness penalty term multiplied by a positive smoothing parameter is added to the residual sum of squares in smoothing spline regression. In the light of this approach, the estimation of the unknown function depends on smoothing parameter λ. Therefore, the determination of an optimum smoothing parameter in the interval $(0, \infty)$ was found to be an underlying complication. In the literature, different selection methods are components of various studies for an appropriate smoothing parameter. Indeed, to a considerable extent, Craven and Wahba [5], Hardle [8], Hardle, Hall and Marron [9], Wahba [26], Hurvich, et al. [11], Eubank [6], Lee and Solo [17], Hastie and Tibshirani [10], Schimek [22], Cantoni and Ronchetti [4], Ruppert, Wand and Carroll [21], Lee [15, 16], and Kou [12] supplement on the selection of the smoothing parameter.

In this study, the empirical performances of the selection methods used in selection of the smoothing parameter are compared. Selection methods used in our simulation study are an

[*] Corresponding author.

Email addresses: duaydin@mu.edu.tr (D. Aydin)

© 2013 EJPAM All rights reserved.

improved version of Akaike information criterion (AICc), robustified cross-validation (RCV), average predictive square error (PSE), parallel of Akaike's information criterion (GFAIC), generalized cross-validation (GCV), cross-validation (CV), Mallows' Cp criterion,risk estimation using classical pilots (RCP) and local risk estimation (LRS). A simulation study was conducted to find out which selection methods are the best in smoothing parameter selection. To throw light on this issue, the samples differing in small and large sizes are secured by means of the above mentioned simulation, and moreover, nine selection methods are evaluated.

This paper is mainly concerned with the selection of smoothing parameter (or penalty parameter) through Monte Carlo simulation study. Smoothing parameters play a crucial role in this procedure. These parameters are said to control the trade off between fidelity to the data and smoothness: too low values of smoothing parameter overfit the data, whereas too high values oversmooth. Krivobokova and Kauermann [14] showed that using the REML to estimate smoothing parameter outperforms other methods such as (generalized) CV or Akaike criterion especially when the error correlation structure is misspecified. Krivobokova et al. [13] formulated a hierarchical mixed model to estimate local smoothing parameter to achieve adaptive penalized spline smoothing. Yanrong Cao et al. [27] discussed different methods of choosing the important smoothing parameter and recommend GCV as the choice for penalized spline smoothing parameter selection for both computational efficiency and accuracy of the functional coefficient regression models. Aydin and Memmedli [3] recommended GCV and REML as being good smoothing parameter selection criteria for small and medium sized samples.

Nonparametric regression and its prediction are discussed in section 2. Section 3 reviews nine different smoothing parameter selection methods. Section 4 compares these methods via a simulation study, and finally, the conclusion and recommendations are presented in section 5.

2. Nonparametric Regression Model and its Prediction

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

$$y_i = f(x_i) + \varepsilon_i, \quad a < x_1 < \ldots < x_n < b \tag{1}$$

where $f \in C^2[a,b]$ is an unknown smooth function, $(y_i)_{i=1}^n$ are observation values of the response variable y, $(x_i)_{i=1}^n$ are observation values of the predictor variable x and $(\varepsilon_i)_{i=1}^n$ are normal distributed random errors with zero mean and common variance σ^2 ($\varepsilon_i \tilde{} N(0,\sigma^2)$).

The basic aim of the nonparametric regression is to estimate unknown function $f \in C^2[a,b]$ (the class of all functions f with continuous first and second derivatives) in model (1). Smoothing spline estimate of the f function appears as a solution to the following minimization problem: Find $\hat{f} \in C^2[a,b]$ that minimizes the penalized residual sum of squares

$$S(f) = \sum_{i=1}^{n} \{y_i - f(x_i)\}^2 + \lambda \int_a^b \{f''(x)\}^2 dx \tag{2}$$

for pre-specified value $\lambda > 0$. The first term in equation (2) denotes the residual sum of the squares (RSS) and it penalizes the lack of fit. The second term which is weighted by λ denotes the roughness penalty and it imposes a penalty on roughness. In other words, it penalizes the curvature of the function f. The λ in (2) is recognized to be the smoothing parameter. As λ varies from 0 to + , the solution varies from interpolation to a linear model. As $\lambda \to +\infty$, the roughness penalty dominates in (2) and the spline estimate is compelled to be a constant. As $\lambda \to\to 0$, the roughness penalty disappears in (2) and the spline estimation interpolates the data. Thus, the smoothing parameter λ plays a key role in controlling the trade-off between the goodness of fit represented by $\sum_{i=1}^{n} \{y_i - f(x_i)\}^2$ and smoothness of the estimate measured by $\int_a^b \{f''(x)\}^2 dx$.

The solution based on smoothing spline for minimum problem in the equation (2) is known as a "natural cubic spline" with knots at $x_1, \ldots, x_n$. From this point of view, a special structured spline interpolation which depends on a chosen value λ develops into a suitable approach of function f in model 1. Let $f = (f(x_1), \ldots, f(x_n))$ be the vector of values of function f at the knot points $x_1, \ldots, x_n$. The smoothing spline estimate $\hat{f}_\lambda$ of this vector or the fitted values for data $y = (y_1, \ldots, y_n)^T$ are projected by

$$\hat{\mathbf{f}}_\lambda = \begin{bmatrix} \hat{f}_\lambda(x_1) \\ \hat{f}_\lambda(x_2) \\ \vdots \\ \hat{f}_\lambda(x_n) \end{bmatrix}_{(n\times 1)} = S_\lambda \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{(n\times 1)} \quad \text{or } \hat{\mathbf{f}}_\lambda = S_\lambda \mathbf{y} \tag{3}$$

where $\hat{f}_\lambda$ is a natural cubic spline with knots at $x_1, \ldots, x_n$ for a fixed $\lambda > 0$, and S_λ is a well-known positive-definite (symmetrical) smoother matrix which depends on λ and the knot points $x_1, \ldots, x_n$, but not on y. Function $\hat{f}_\lambda$, the estimation of function f, is obtained by cubic spline interpolation that rests on condition $\hat{f}(x_i) = (\hat{f})_i$, $i = 1, 2, \ldots, n$. To gain better perspective on smoothing spline, Eubank [6], Green and Silverman [7] and Wahba [26] state studied opinions.

3. Smoothing Parameter Selection Methods

Although smoothing spline estimator solves the problem of allowing fits with variable slope, a new dilemma emerges. In fact, it generates the determination of the appropriate value for the smoothing parameter λ for a given data set. The same value of λ is unlikely to work equally well with every data set. As such, the estimation methods have been introduced for the selection of smoothing parameter λ in equation (2). The positive value λ that minimizes any smoothing parameter selection methods is selected as an appropriate smoothing parameter.

3.1. Selection Methods used in Simulation Study

Various smoothing parameter selection methods are featured in the literature. Most of these suggested methods were implemented in our simulation study. Moreover, a selection criterion from previous studies in the literature to provide an effective performance was also introduced in this particular study. The selection criteria used in our simulation study are classified as:

Average predictive squared error: In selection the smoothing parameter, it is essential not to try to minimize the mean squared error at each x_i, but instead, the focus should centre on a global measures such as average predictive squared error

$$PSE(\lambda) = \left\{ 1 + \frac{tr(S_\lambda S_\lambda^T)}{n} \right\} \sigma^2 + \frac{\left\| (I - S_\lambda) f \right\|^2}{n} \tag{4}$$

Where $tr(S_\lambda S_\lambda^T)$ is trace of matrix $S_\lambda S_\lambda^T$ and $\left\| (I - S_\lambda) f \right\|$ is norm of matrix $(I - S_\lambda) f$. If σ^2 is not known, in practice an estimation for σ^2 can be given by

$$\hat{\sigma}^2 = \frac{RSS(\lambda^*)}{\left\{ n - tr\left(2S_{\lambda^*} - S_{\lambda^*} S^T{}_{\lambda^*}\right) \right\}} = \frac{\left\| (S_{\lambda^*} - I) y \right\|^2}{\left\{ n - tr\left(2S_{\lambda^*} - S_{\lambda^*} S^T{}_{\lambda^*}\right) \right\}},$$

where $RSS(\lambda^*)$ is the residual sum of square from a smooth $S_{\lambda^*} y$ and λ^* is a pilot λ selected by any selection methods [see 10].

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

$$\mathrm{CV}(\lambda) = \frac{1}{n} \sum_{i=1}^n \left\{ y_i - \hat{f}_\lambda^{(-i)}(x_i) \right\}^2 \equiv \mathrm{CV}(\lambda) = \frac{1}{n} \sum_{i=1}^n \left\{ \frac{y_i - \hat{f}_\lambda(x_i)}{1 - (S_\lambda)_{ii}} \right\}^2, \tag{5}$$

where $\hat{f}_\lambda$ is the fit (spline smoother) for n pairs of measurements $\{x_i, y_i\}_{i=1}^n$ with smoothing parameter λ, and $\hat{f}_\lambda^{(-i)}$ is the fit calculated by leaving out the ith data point and $(S_\lambda)_{ii}$ is the ith diagonal element of smoother matrix S_λ [see 26, 7].

Using the approximations $(S_\lambda)_{ii} \approx \left\{ S_\lambda S_\lambda^T \right\}_{ii}$ and $1/\left(1 - (S_\lambda)_{ii}\right)^2 \approx 1 + 2(S_\lambda)_{ii}$, signifies that $E\{CV(\lambda)\} \approx PSE(\lambda) + 2/n + \sum_{i=1}^n (S_\lambda)_{ii} b_i^2(\lambda)$ [see 10].

Generalized cross-validation: GCV is a modified form of the CV which is a conventional method for choosing the smoothing parameter. The GCV score which is constructed by analogy to CV score can be obtained from the ordinary residuals by dividing by the factors $1 - (S_\lambda)_{ii}$. The underlying design of GCV is to replace the factors $1 - (S_\lambda)_{ii}$ in equation (5) with the average score $1 - n^{-1} tr(S_\lambda)$ Thus, by summing the squared corrected residual and factor $\left\{ 1 - n^{-1} tr(S_\lambda) \right\}^2$, by the analogy ordinary cross-validation, the GCV score function can be procured as follow [see 5, 26]:

$$GCV(\lambda) = \frac{1}{n}\frac{\sum_{i=1}^{n}\left\{y_i - \hat{f}_\lambda(x_1)\right\}^2}{\left\{1 - n^{-1}tr(S_\lambda)\right\}^2} = \frac{n^{-1}\left\|\left(I - S_\lambda\right)y\right\|^2}{\left[n^{-1}tr\left(I - S_\lambda\right)\right]^2} \tag{6}$$

Mallows' CP criterion: In the literature, C_p criterion is referred to as an unbiased risk estimate (UBR). This type of estimate was suggested by Mallows [18] in the regression case, and applied to smoothing spline by Craven and Wahba [5]. If σ^2 is recognized, an unbiased estimate of the residual sum of squares is provided by C_p criterion:

$$C_p(\lambda) = \frac{1}{n}\left\{\left\|(S_\lambda - I)y\right\|^2 + 2\sigma^2 tr(S_\lambda) - \sigma^2\right\} = \frac{1}{n}\left\{\left\|y - \hat{f}_\lambda\right\|^2 + 2\sigma^2 tr(S_\lambda) - \sigma^2\right\} \tag{7}$$

Unless σ^2 is known, in practice an estimation for σ^2 can be given by

$$\hat{\sigma}^2 = \hat{\sigma}^2_{\hat{\lambda}} = \frac{\sum_{i=1}^{n}\left(y_i - \hat{f}_{\hat{\lambda}}(x_i)\right)^2}{tr\left(I - S_{\hat{\lambda}}\right)} = \frac{\left\|(S_{\hat{\lambda}} - I)y\right\|^2}{tr\left(I - S_{\hat{\lambda}}\right)} \tag{8}$$

where $\hat{\lambda}$ is pre-chosen with any of the CV, GCV or AICC criteria ($\hat{\lambda}$ is an estimate of λ) For reference, see [15, 16, 26]. According to Hastie and Tibshirani [10] and Ruppert, et. al. [21], GCV is approximately equal to C_p.

Improved Akaike information criterion: An improved version of a criterion based on the classical Akaike information criterion (AIC), AICc criterion, is used for choosing the smoothing parameter for nonparametric smoothers [11]. This improved criterion is defined as

$$\begin{aligned} AIC_C(\lambda) &= \log\frac{\sum\left\{y_i - \hat{f}_\lambda(x_i)\right\}^2}{n} + 1 + \frac{2\left\{tr(S_\lambda) + 1\right\}}{n - tr(S_\lambda) - 2} \\ &= \log\frac{\left\|\left(S_\lambda - I\right)y\right\|^2}{n} + 1 + \frac{2\left\{tr(S_\lambda) + 1\right\}}{n - tr(S_\lambda) - 2}. \end{aligned} \tag{9}$$

This criterion is easy to apply for the selection of smoothing parameter, as can be seen from the equation (9).

Robustified Cross-Validation: The smoothing parameter λ can be chosen by method of robustified cross-validation (RCV). This serves to reduce risk of smoothing parameter misspecification in small sample sizes. Most of the selections methods have been proposed to optimize λ are related to a distributional. RCV criterion that does not refer to a distributional assumption is proposed by Robinson and Moyeed [20]. Robustified cross-validation score is given by

$$RCV(\lambda) = n^{-1}\frac{1 + n^{-1} + tr\left(S_\lambda\right)^2}{\left(1 + n^{-1} + tr\left(S_\lambda\right)\right)^2}\left\|\left(I - S_\lambda\right)y\right\|^2. \tag{10}$$

The minimum of (10) was empirically obtained by successively searching interlocked grid intervals of flexible sizes.

Parallel of Akaike's information criterion: A result of Stone [24] states that in the case of independent observation, the sum over i of the values of log-likelihood function in $y = (y_1, \ldots, y_n)$ with parameters estimated by maximum likelihood based on $\{x_i\}$ is asymptotically equivalent to Akaike's information criterion (AIC). This result is not directly applicable to splines, instead of a simple, maximum-likelihood estimate and σ^2 as additional noise parameter has to be taken into account. However, there is parallel of AIC [22]. This parallel criterion is given by

$$GF_{AIC} = n^{-1} \left\| (I - S_\lambda)\, y \right\|^2 + \exp(2n^{-1} tr(S_\lambda)) \tag{11}$$

Minimization of GF_{AIC} looks like maximization of the log-likelihood function in y.

Risk estimation using classical pilots: Risk function measures the distance between the actual regression function (f) and its estimation $(\hat{f}_\lambda)$. Needless to say that, a good estimate must contain minimum risk. A direct computation leads to the bias-variance decomposition for $R(f, \hat{f}_\lambda)$:

$$R(f, \hat{f}_\lambda) = \frac{1}{n} E \left\| f - \hat{f}_\lambda \right\|^2 = \frac{1}{n} \left\{ \left\| (S_\lambda - I)\, f \right\|^2 + \sigma^2 tr(S_\lambda S_\lambda^T) \right\} \tag{12}$$

A clear-cut explanation shows that $R(f, \hat{f}_\lambda) = E\left\{C_p(\lambda)\right\}$. Because the risk $R(f, \hat{f}_\lambda)$ is an 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

$$R(\hat{f}_{\lambda_p}, \hat{f}_\lambda) = \frac{1}{n} E \left\| \hat{f}_{\lambda_p} - \hat{f}_\lambda \right\|^2 = \frac{1}{n} \left\{ \left\| (S_\lambda - I)\, \hat{f}_{\lambda_p} \right\|^2 + \hat{\sigma}^2_{\lambda_p} tr(S_\lambda S_\lambda^T) \right\}, \tag{13}$$

where $\hat{\sigma}^2_{\lambda_p}$ 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.

Local risk estimation: The LRS method proposed by Lee [16], aims to select the $\hat{f}_\lambda(x_i)$ that minimizes the local risk $R_\lambda(x_i) = E\left\{f(x_i) - \hat{f}_\lambda(x_i)\right\}^2$ for the each knot points x_i A direct computation leads to the bias-variance decomposition for $R_\lambda(x_i)$:

$$R_\lambda(x_i) = \left\{ (S_\lambda f)\,(x_i) - f(x_i) \right\}^2 + \sigma^2 s_\lambda(x_i) \tag{14}$$

In the above equation, $(S_\lambda f)\,(x_i)$ is the ith element of vector $S_\lambda f$ and $s_\lambda(x_i)$ is the ith diagonal element of the square matrix $S_\lambda S_\lambda^T$. An estimator for $R_\lambda(x_i)$ is firstly computed and the $\hat{f}_\lambda(x_i)$ is selected in order to minimizes it. This process is repeated for all x_i's and at the end of the process a final mixed estimate for f is derived. The LRS method can be practically performed with the following five steps [15]:

(i) For a set of pre-selected smoothing parameters $\lambda_1 < \ldots < \lambda_m$, calculate the corresponding set of smoothing spline estimates: $F = \left\{\hat{f}_{\lambda_1}, \ldots, \hat{f}_{\lambda_m}\right\}$;

(ii) Select the pilot value λ_p from AIC_c criterion in (9) by using the elements in F;

(iii) For λ_p, calculate the estimates $\hat{f}_{\lambda_p}$, and $\hat{\sigma}^2_{\lambda_p}$ by using the (8);

(iv) Substitute the pilots $\hat{f}_{\lambda_p}$ and $\hat{\sigma}^2_{\lambda_p}$ into the expression

$$\hat{R}_\lambda(x_i) = \left\{ \left(S_\lambda \hat{f}_{\lambda_p} \right)(x_i) - \hat{f}_{\lambda_p}(x_i) \right\}^2 + \hat{\sigma}^2_{\lambda_p} s_\lambda(x_i)$$

and obtain the estimates $\hat{R}_\lambda(x_i)$

(v) For each x_i, find the $\hat{f}_\lambda(x_i)$ from F which minimizes $\hat{R}_\lambda(x_i)$ and the final estimate accept the appropriate values $\hat{f}_\lambda(x_i)$ for $f(x_i)$

3.2. Other Selection Methods

We also explored most of the selection methods that are not used in our simulation. Hardle, Hall and Marron [9] recommended that $AIC(\lambda)$, $FPE(\lambda)$, and $SH(\lambda)$ should not be used because of their trivial minimum at the no-smoothing point. They are also preference $T(\lambda)$ for bandwidth selector, but the statement of Rice [19] that "these result are suggestive", but $T(\lambda)$ hasn't had a good performance in our simulation study. It is accepted that $GLM(\lambda)$ estimates undersmooths relative to the $GCV(\lambda)$ estimate (see, Wahba [25]). Moreover, Hastie and Tibshirani [10], also point out that minimizing $ASR(\lambda)$ over the smoothing parameter leads to an interpolation estimate.

Furthermore, we have also performed a simulation study for all the selection methods. In such cases, the selection methods produced invalidating results when compared with the selection methods such as AICc, GFAIC, GCV, Cp and RCV used in our simulation study. For this reason, in this paper, the simulation results of the AIC, FPE, SH, T, GLM and ASR criteria are not displayed. For smoothing spline, these selection criteria can be given as:

(a) Akaike's information criterion [2],

$$AIC(\lambda) = \frac{1}{n} \left\| (I - S_\lambda)\, y \right\|^2 \exp(1 - n^{-1} tr(S_\lambda));$$

(b) finite prediction error [1],

$$FPE(\lambda) = \frac{1}{n} \left\| (I - S_\lambda)\, y \right\|^2 \frac{1 + n^{-1} tr(S_\lambda)}{1 - n^{-1} tr(S_\lambda)};$$

(c) a model selector of [23];

$$SH(\lambda) = \frac{1}{n} \left\| (I - S_\lambda)\, y \right\|^2 \left(1 - n^{-1} 2 tr(S_\lambda) \right)$$

(d) Rice's T criterion [19];

$$T(\lambda) = \frac{1}{n} \left\| (I - S_\lambda)\, y \right\|^2 / \left(1 - n^{-1} 2 tr(S_\lambda) \right)$$

(e) A generalization of the maximum likelihood (GLM),

$$GLM(\lambda) = y' \left(I - S_\lambda\right) y / \left[\det^+ \left(I - S_\lambda\right)\right]^{1/n-m};$$

where $\det^+ \left(I - S_\lambda\right)$ is the product of $n - m$ nonzero eigenvalues of $\left(I - S_\lambda\right)$.

(f) Average squared residual,

$$ASR(\lambda) = \frac{1}{n}\sum_{i=1}^{n} \left\{y_i - \hat{f}_\lambda(x_i)\right\}^2 .$$

4. Simulation Study

This section reports the results of a Monte Carlo simulation study. This study was conducted to evaluate the performances of the nine selection methods mentioned above. The experimental setup in this paper is adopted from Professor Steve Marron. By using a program coded in MATLAB, we generated the samples sized $n = 25, 50, 100, 150, 200, 350, 400$. The number of replications was 500 for each of the samples. For each simulated data sets, the mean squared-errors (MSE) was used for evaluate the quality of any curve estimate $\hat{f}$. 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. Experimental Plan and Generation of the Data

The experimental setup applied at this stage was designed to study the effects of the following three factors which vary an independent and effective approach: Noise level; Spatial variation; Variance function.

The setup specification is listed Table 1. The simulation study was performed according to MATLAB program, and the experimental setup was designed in the following framework:

- Totally three sets of numerical experiments were performed. For each set of experiments, only one of the above three experimental factors (e.g., noise level, degree of spatial variation and noise variance function) has been altered while the remaining two have been kept unchanged.

- Within each set of experiments, the factor levels was modified four times (i.e., $r = 1, 2, 3, 4$) to detect the effects of any experimental factor in Table 1.

- To see the performance of the small and large samples of the selection methods. For each factor level r within each set of experiments, we generated 7 different samples with sample sizes $n = 25, 50, 100, 150, 200, 350, 400$.

- The number of replications was 500 for each of the 84 numerical experiments.

- We computed the appropriate smoothing spline estimators $\hat{f}_\lambda$ in equality (3) by selecting the smoothing parameter λ which minimizes the selection methods.

- We used the MSE values to evaluate $\hat{f}_\lambda$ computed according to each of the selection criterion: $MSE = \frac{1}{n}\sum_{i=1}^{n}\left\{f(x_i) - \hat{f}_\lambda(x_i)\right\}^2$, $(\hat{f}_\lambda(x_i) = (\hat{f}_\lambda)_i)$, (where $f(x_i)$ is value at knots x_i of the appropriate function f defined in Table 1)

- Paired Wilcoxon test was applied to test whether MSE values was considered as the performance measure of any two methods are significant or not.

- The factor levels indicated as r was changed four times (i.e., $r = 1, 2, 3, 4$) in order to detect the effects of any factor from three factors in Table 1.

- By considering 3 factors, 4 factor levels and 7 samples, totally, 84 numerical experiments were conducted.

Table 1: Averaged Wilcoxon test ranking values for the nine selection methods in small sample sizes

Factors	Generic Form	Particular Choices
Noise Level	$y_{ir} = f(x_i) + \sigma_r \varepsilon_i$	Factors
Spatial variation	$y_{ir} = f(x_i) + \sigma_r \varepsilon_i$	$\sigma = 0.2$, $f_r(x) = \sqrt{x(1-x)}\sin\left(\frac{2\pi\{1+2^{(9-4r)/5}\}}{x+2^{(9-4r)/5}}\right)$
Variance function	$y_{ir} = f(x_i) + \sqrt{v_r(x_i)\varepsilon_i}$	$v_r(x) = [0.15\{1 + 0.4(2r-7)(x-0.5)\}]^2$
$r = 1, \ldots, 4; x_i = \frac{i-0.5}{n}; \varepsilon_i \sim iid\, N(0,1); f(x) = 1.5\theta\left(\frac{x-0.35}{0.15}\right) - \theta\left(\frac{x-0.8}{0.04}\right); \theta(u) = \frac{1}{\sqrt{2\pi}}\exp\left(\frac{-u^2}{2}\right)$ $n = 25, 50, 100, 200, 350, 400$ (it was taken seven sample size).		

4.2. Experimental Evaluations

For each simulated data set used in the experiments, the MSE values were used in order to evaluate the quality of any curve estimate $\hat{f}$. 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 a rank 2, and similarly for ranks 3-9. 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.

In this simulation study, because totally 84 different configurations are made, it is not possible to display here all these configurations. Therefore, only 24 different configurations are given in Figures A1-A6 (in the appendix) for different samples sized n. The head row plots of the figures A1-A6 display the true regression function together with all typical simulated data set. The bottom row plots display the boxplots of the $\log_e MSE$ values for, from left to right, AICc, RCV, FSE, GFAIC, GCV, CV, Cp, RCP and LRS. The numbers below the boxplots are the paired Wilcoxon test rankings. For 84 different simulation experiments, the averaged ranking values of the selection methods according to Wilcoxon tests are tabulated in Tables A1 and A2. All results tables are shown in the appendix; $(*)$ indicates the selection methods with the best rankings.

According to the results in Table A1, for small sized samples (for $n = 25, 50, 100$), GCV has had the best empirical performance for all factors. Furthermore, GFAIC and RCV have shared the better performance after GCV criterion. In accordance with the overall Wilcoxon test rankings in Table A1, GCV, GFAIC and RCV have also displayed a good performance. As shown Table A1, because of $R\left(f, \hat{f}_\lambda\right) = E\left\{C_p(\lambda)\right\}$, C_p and RCP methods produced the same results under all experimental factors. In this situation, for small sample sizes, for which reason the effects of the replication of simulation, C_p is approximately equal to its $E\left\{C_p\right\}$. For small samples, it is observed that PSE has produced the worst performance. According to Table A2, for large sized samples (for $n = 150, 200, 350400$), GFAIC criterion has had the best empirical performance. Generally it is shown that AICc, RCV and GCV (C_p) criteria have shared a better performance after than GFAIC. According to the overall Wilcoxon test rankings in Table A2, GFAIC, AICc and RCV criteria can be ranked in terms of the performance. As shown in Table A2, generally, C_p and GCV gave the same results. This can be interpreted to follow the accepted view that GCV is asymptotically equal to GCV [see 10]. That is to say that, for large sized samples, because of the effect of the replication of simulation, performance of the GCV and C_p is approximately equal. PSE has also produced the worst performance for large samples.

5. Conclusion and recommendations

The scores in Tables A3 and A4 in the appendix are obtained by taking the means of the averaged Wilcoxon test ranking values tie with each of the selection methods in Table A1-A2 respectively.

As shown in Table A3, according to the means of the small samples for all factors, GCV, GF_{AIC} and RCV criteria have had the best empirical performance respectively. When it is compared to the other criteria, PSE, C_p(RCP), AIC_c criteria have resulted in the worst performance.

According to the means of the large samples for all factors in Table A4, it is observed that GFAIC, AIC_c, GCV (C_p) have had the best empirical performance. However, PSE and CV have produced the worst result

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

- For both large and small samples, GF_{AIC} and GCV are recommended as being the best selection criteria;

- For especially large samples, the use GF_{AIC} would seem to be more appropriate. As for small samples, we propose the implementation of GCV criterion;

- For large samples, the implementation of AIC_c criterion, in addition to GFAIC and GCV (C_p) criteria would be more beneficial. For small samples, RCV criterion in addition to GFAIC and GCV criteria should prove fruitful.

Naturally, the above recommended suggestions have to be considered with a fair amount of caution as they are only an appraisal based on simulation results.

References

[1] H Akaike. Statistical predictor identification. *Annals of the Institute of Statistical Mathematics (AISM)*, 22:203–217, 1970.

[2] H Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

[3] D Aydin and M Memmedli. Optimum smoothing parameter selection for penalized least squares in form of linear mixed effect models. *Optimization*, iFirst:1–18, 2011.

[4] E Cantoni and E Ronchetti. Resistant selection of smoothing parameter for smoothing splines. *Statisics and Computing*, 11:141–146, 2001.

[5] P Craven and G Wahba. Smoothing noisy data with spline functions. *Numerische Mathematik*, 31:377–403, 1979.

[6] R L Eubank. *Nonparametric Regression and Smoothing Spline*. Marcel Dekker Inc., New York, 1999.

[7] P J Green and B W Silverman. *Nonparametric Regression and Generalized Linear Model*. Chapman & Hall, New York, USA, 1994.

[8] W Hardle. *Applied Nonparametric Regression*. Cambridge University Press, Cambridge, 1991.

[9] 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*, 83:86–89, 1988.

[10] T Hastie and R J Tibshirani. *Generalized Additive Models*. Chapman & Hall, London, 1999.

[11] C M Hurvich, J S Simonoff, and C L Tasi. Smoothing parameter selection in nonparametric regression using an improved akaike information criterion. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 60:271–293, 1998.

[12] S C Kou. On the efficiency of selection criteria in spline regression. *Probability Theory and Related Fields*, 127:153–176, 2003.

[13] T Krivobokova, C M Crainiceanu, and G Kauermann. Fast adaptive penalized splines. *Journal of Computational and Graphical Statistics*, 17:1–20, 2008.

[14] T Krivobokova and G Kauermann. A note on penalized spline smoothing with correlated errors. *Journal of the American Statistical Association*, 102:1328–337, 2007.

[15] T C M Lee. Smoothing parameter selection for smoothing splines: A simulation study. *Computational Statistics & Data Analysis*, 42:139–148, 2003.

[16] T C M Lee. Improved smoothing spline regression by combining estimates of different smoothness. *Statistics & Probability Letters*, 67:133–140, 2004.

[17] T C M Lee and V Solo. Bandwidth selection for local linear regression: A simulation study. *Computational Statistics & Data Analysis*, 14:515–532, 1999.

[18] C Mallows'. Some comments on cp. *Technometrics*, 15:661–375, 1973.

[19] J Rice. Bandwidth choice for nonparametric regression. *Annals of Statistics*, 12:1215–1230, 1984.

[20] T Robinson and R Moyeed. Making robust the cross-validation choice of smoothing parameter in spline regression. *Communications in Statistics - Theory and Methods*, 18:523–539, 1989.

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

[22] G M Schimek. *Smoothing and Regression: Approaches, Computation, and Application*. John Willey & Sons, Inc., USA, 2000.

[23] R Shibata. An optimal selection of regression variables. *Biometrika*, 68(1):45–45, 1981.

[24] M Stone. An asymptotic equivalence of choice of model by cross-validation and akaike's criterion. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 39:44–47, 1977.

[25] G Wahba. A comparison of gcv and gml for choosing the smoothing parameter in the generalized spline smoothing problem. *Annals of Statistics*, 13:1378–1402, 1985.

[26] G Wahba. *Spline Model For Observational Data*. Siam, Philadelphia, 1990.

[27] C Yanrong, Z W Tracy L Haiqun, and Y Yan. Penalized spline estimation for functional coefficient regression models. *Computational Statistics and Data Analysis*, 54:891–905, 2010.

Simulation Results

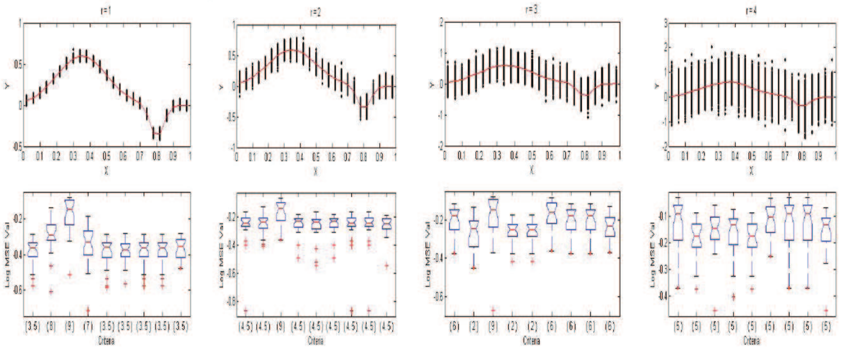

Figure A1: Simulation results correspond to the noise level factor for $n = 25$.

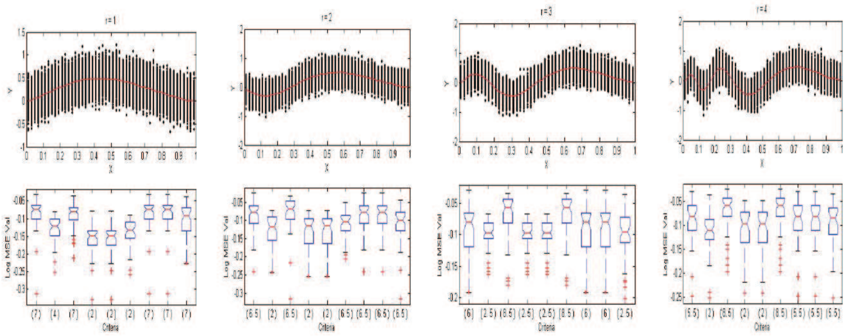

Figure A2: Simulation results correspond to the spatial variation factor for $n = 50$.

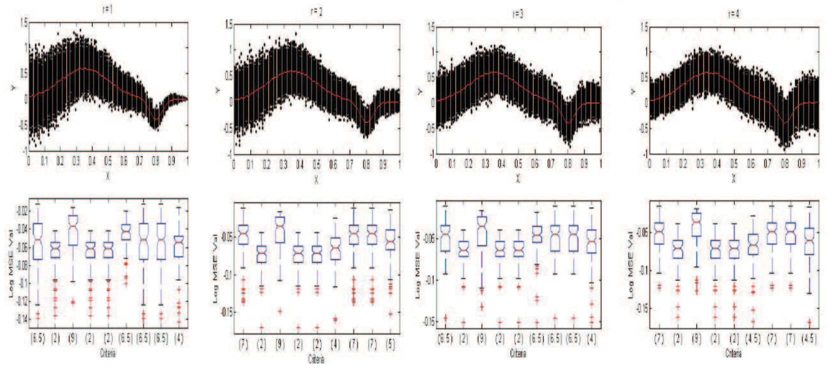

Figure A3: Simulation results correspond to the variance factor for $n = 100$.

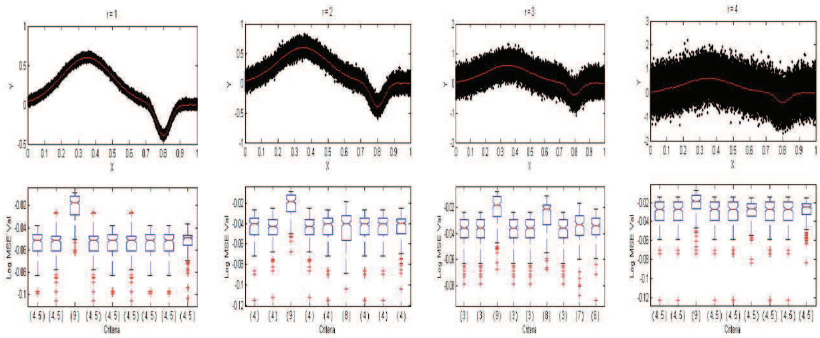

Figure A4: Simulation results correspond to the noise level factor for $n = 200$.

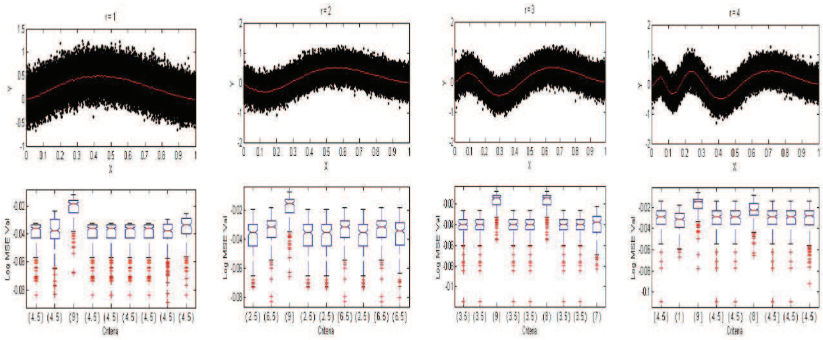

Figure A5: Simulation results correspond to the spatial variation factor for $n = 350$.

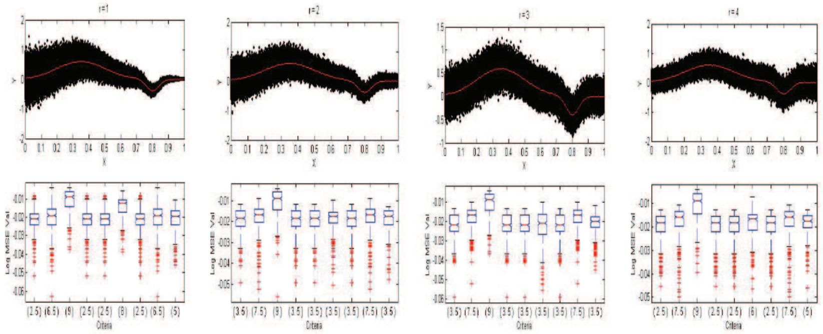

Figure A6: Simulation results correspond to the variance factor for $n = 400$.

Table A1: Averaged Wilcoxon test ranking values for the nine selection methods in small sample sizes

Criteria	Noise level	Spatial variation	Variance function	Overall
For $n = 25$				
AICc	4.75	5.375	5.25	5.125
RCV	4.875	3.250*	4.250*	4.125
PSE	8	8.5	7	7.83
GFAIC	4.625	3.250*	4.250*	4.042
GCV	3.750*	3.250*	4.250*	3.708*
CV	4.75	5.25	5.25	5.083
Cp	4.75	5.375	5.25	5.125
RCP	4.75	5.375	5.25	5.125
LRS	4.75	5.375	4.250*	4.792
For $n = 50$				
AICc	6.125	6.25	5.625	6
RCV	3.625	2.625*	2.750*	3
PSE	7	7.625	9	7.875
GFAIC	3.625	2.625*	2.750*	3
GCV	2.625*	2.625*	2.750*	2.667*
CV	4.375	7.625	6.125	6.042
Cp	6.125	6.25	5.625	6
RCP	6.125	6.25	5.625	6
LRS	5.375	5.375	4.75	5.167
For $n = 100$				
AICc	6.625	6.625	6.75	6.417
RCV	2.875	2.875	2.000*	2.583
PSE	8.125	8.5	9	8.542
GFAIC	2.875	2.000*	2.000*	2.292
GCV	2.000*	2.000*	2.000*	2.000*
CV	6.625	5.375	5.375	5.792
Cp	6.625	6.625	6.75	6.417
RCP	6.625	6.625	6.75	6.417
LRS	4.875	4.875	4.375	4.7

Table A2: Averaged Wilcoxon test ranking values for the nine selection methods in large sample sizes

Criteria	Noise level	Spatial variation	Variance function	Overall
For $n = 150$				
AICc	4.25	3.500*	3.250*	3.667
RCV	3.500*	4.75	3.250*	3.833
PSE	9	8.625	8.875	8.833
GFAIC	3.500*	3.500*	3.250*	3.417*
GCV	4.25	3.500*	4.875	4.208
CV	6.375	6.875	7	6.75
Cp	4.25	3.500*	3.250*	3.667
RCP	5.125	4.5	5.75	5.125
LRS	4.75	6.25	6	5.667
For $n = 200$				
AICc	4.000*	3.750*	2.500*	3.417*
RCV	4.000*	3.875	2.500*	3.458
PSE	9	9	9	9
GFAIC	4.000*	3.750*	2.500*	3.417*
GCV	4.000*	3.750*	4.625	4.125
CV	6.25	6.75	6.5	6.5
Cp	4.000*	3.750*	4.625	4.125
RCP	5	4.75	6.5	5.417
LRS	4.75	5.625	5.25	5.208
For $n = 350$				
AICc	3.875	3.375*	2.875*	3.375
RCV	6.5	6.625	7	6.708
PSE	9	8.75	9	8.917
GFAIC	2.625*	3.375*	2.875*	2.950*
GCV	3.875	3.375*	3.75	3.667
CV	7.75	7.25	5.875	6.958
Cp	3.875	3.375*	3.75	3.667
RCP	5	4.625	7	5.542
LRS	3	4.125	2.875*	3.333
For $n = 400$				
AICc	4.5	3.375*	3.000*	3.625
RCV	6.375	6.75	7.25	6.792
PSE	9	9	9	9
GFAIC	3.500*	3.375*	3.000*	3.292*
GCV	4.5	3.375*	3.000*	3.625
CV	5.75	6	5.25	5.667
Cp	4.5	3.375*	3.000*	3.625
RCP	5.5	5	7.25	5.917
LRS	4.5	4.75	4.25	4.5

Table A3: Means of the averaged Wilcoxon test ranking values for the nine selection methods

Criteria	Noise level	Spatial variation	Variance function	Overall
(Means for $n = 25, 50, 100$)				
AICc	5.667	5.958	5.875	5.833
RCV	3.792	2.917	3.000*	3.236
PSE	7.708	8.208	8.333	8.083
GFAIC	3.708	2.625*	3.000*	3.111
GCV	2.792*	2.625*	3.000*	2.806*
CV	5.125	6.083	5.583	5.597
Cp	5.667	6.083	5.875	5.875
RCP	5.667	6.083	5.875	5.875
LRS	5	5.208	4.458	4.889

Table A4: Means of the averaged Wilcoxon test ranking values for the nine selection methods

Criteria	Noise level	Spatial variation	Variance function	Overall
(Means for $n = 150, 200, 350, 400$)				
AICc	4.156	3.406*	2.906*	3.489
RCV	5.094	5.5	5	5.198
PSE	9	8.313	8.969	8.761
GFAIC	3.281*	3.406*	2.906*	3.198*
GCV	4.156	3.406*	4.063	3.875
CV	6.531	6.719	6.156	6.469
Cp	4.156	3.406*	3.656	3.739
RCP	5.156	4.719	6.625	5.5
LRS	4.25	5.186	4.594	4.677