The Empirical Performances of the Selection Criteria for Nonparametric Regression Using Smoothing Spline

DURSUN AYDIN, RABIA ECE OMAY
Department of Statistics
Anadolu University
Eskisehir 26470
TURKEY

Abstract: - In this paper, the smoothing parameter selection problem has been examined in nonparametric regression using smoothing spline method for different data sets. For that reason a Monte Carlo simulation study has been performed with a program that coded in MATLAB. This simulation study provides a comparison of the five popular smoothing parameter selection criteria. Thus, the emprical performances of the five selection criteria have been investigated and suitable criteria, which used for smoothing parameter selection, have been obtained.

Key-Words: - Nonparametric regression; Smoothing spline; Smoothing parameter; Selection criteria

1 Introduction

Smoothing spline method is one of the most popular methods used for the prediction of the nonparametric regression models. The aim of this method is to estimate the nonparametric function that minimizes penalized least squares criterion. A roughness penalty term multiplied by a positive λ smoothing parameter is added to the residual sum of squares in smoothing spline regression. For this reason, the estimation of the unknown function depends on smoothing parameter λ whose values is generally obtained from data. Therefore, the determination of an optimum smoothing parameter in the interval $(0, \infty)$ has been arisen as an important problem. As related with subject in theory, many studies based on the different selection methods have been discussed for choosing an appropriate smoothing parameter. For references on choosing of the smoothing parameter, see, for example, Craven and Wahba (1979); Hurvich, et al. (1998); Eubank (1999); Lee and Solo (1999); Cantoni and Ronchetti (2001); Lee (2003 and 2004); Kou and Lee (2003).

In this study, the empirical performances of five smoothing parameter selection methods which are used for selection of the smoothing parameter have been compared. The selection criteria mentioned here are given as fallowing: Cross-validation (CV), generalized cross-validation (GCV), improved version of Akaike information criterion (AIC$_c$), Mallows' C$_p$ and risk estimation using classical pilots (RCP). A simulation study is conducted to find out which selection methods are the best in smoothing parameter selection. Thus, the small and large samples are obtained via the mentioned simulation study and the six selection methods are evaluated.

The rest of this paper is organized as fallows. Nonparametric regression and its prediction is presented in Section 1. Five different smoothing parameter selection methods are reviewed in Section 3. Section 4 compares these methods via a simulation study, while conclusion and recommendations are offered in Section 5.

2 Nonparametric Regression Model and Its Prediction

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

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

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 ($\varepsilon_i \sim N(0,\sigma^2)$).

The essential 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 arises as solution to the following minimization problem: Find $\hat{f} \in C^2[a,b]$ that minimizes the penalized residual sum of squares

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

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 (RS) and it imposes a penalty on roughness. In other words, it penalizes the curvature of function. The λ in (2) is known as the smoothing parameter. As λ varies from 0 to +∞, the solution varies from interpolation to a linear model. When $\lambda \to +\infty$, the roughness penalty dominates in (2) and the spline estimate is forced to be a constant. When $\lambda \to 0$, the roughness penalty disappears in (2) and the spline estimate interpolates the data. Thus, the smoothing parameter λ plays a key role in controlling the trade-off between *the goodness of fit* (the closeness to data) represented by $\sum_{i=1}^{n} \{y_i - f(x_i)\}^2$ and *smoothnees 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 $x_1,...,x_n$. From this point of view, a special structured spline interpolation which depends on a chosen value λ becomes a suitable approach of function f in model 1. Let $\mathbf{f} = (f(x_1),...,f(x_n))$ be the vector of values of function f at the knot points $x_1,...,x_n$. The smoothing spline estimate $\hat{\mathbf{f}}_\lambda$ of this vector or the fitted values for data $\mathbf{y} = (y_1,...,y_n)^{\mathrm{T}}$ are given by

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

where $\hat{f}_\lambda$ is a natural cubic spline with knots at $x_1,...,x_n$ for a fixed smoothing parameter $\lambda > 0$, and S_λ is a well-known positive-definite (symmetrical) smoother matrix which depends on λ and the knot points $x_1,...,x_n$, but not on $\mathbf{y}$. For general references about smoothing spline, see Eubank (1988), Green and Silverman (1994) and Wahba (1990).

3 Smoothing Parameter Selection Criteria

Smoothing spline estimator solves the problem of allowing fits with variable slope, but it creates a new problem. In other words, it creates 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. For this purpose, the selection methods have been introduced for the selection of smoothing parameter λ in equation (2). *The positive value λ that minimizes any selection criter is selected as an appropriate smoothing parameter.*

Cross-Validation: 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

$$\mathrm{CV}(\lambda) = \frac{1}{\mathrm{n}} \sum_{i=1}^{n} \{y_i - \hat{f}_\lambda^{(-i)}(x_i)\}^2 \equiv \mathrm{CV}(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{y_i - \hat{f}_\lambda(x_i)}{1-(S_\lambda)_{ii}} \right\}^2 \qquad (4)$$

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 Wahba, 1990; Green and Silverman,1994).

Generalized cross-validation: GCV is a modified form of the CV which is a popular criter for choosing the smoothing parameter. The GCV score is constructed by analogy to CV score obtained from dividing to the factors $1-(S_\lambda)_{ii}$ of the ordinary residuals. The main idea of GCV is to replace the factors $1-(S_\lambda)_{ii}$ in equation (4) with the average score $1-n^{-1}tr(S_\lambda)$ Thus, by summing of the squared residual corrected and factor $\{1-n^{-1}tr(S_\lambda)\}^2$, by the analogy ordinary cross-validation, the GCV score function is obtained as fallow (Wahba, 1990):

$$\mathrm{GCV}(\lambda)=\frac{1}{n}\frac{\sum_{i=1}^{n}\{y_i-\hat{f}_\lambda(x_1)\}^2}{\{1-n^{-1}tr(S_\lambda)\}^2}=\frac{n^{-1}\|(I-S_\lambda)\mathbf{y}\|^2}{\left[n^{-1}tr(I-S_\lambda)\right]^2} \qquad (5)$$

İmproved Akaike information criterion: 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 (Hurvich et al., 1988). This improved criterion is defined as

$$\begin{aligned}\mathrm{AIC_c}(\lambda)&=\log\frac{\sum\{y_i-\hat{f}_\lambda(x_i)\}^2}{n}+1+\frac{2\{tr(S_\lambda)+1\}}{n-tr(S_\lambda)-2}\\ &=\log\frac{\|(S_\lambda-I)\mathbf{y}\|^2}{n}+1+\frac{2\{tr(S_\lambda)+1\}}{n-tr(S_\lambda)-2}\end{aligned} \qquad (6)$$

This criterion is easy to apply for choosing of smoothing parameter, as can be seen from the equation (6).

Mallows' C$_P$ criterion: C_p criterion is known as *unbiased risk esimate* (UBR) in smoothing spline literature. This type of estimate was suggested by Mallows (1973) in regression case, and applied to smoothing spline by Craven and Wahba (1979). When σ^2 is known, an unbiased estimate of the residual sum of squares is given by C_p criterion:

$$\begin{aligned}C_\mathrm{p}(\lambda)&=\frac{1}{n}\left\{\|(S_\lambda-I)\mathbf{y}\|^2+2\sigma^2tr(S_\lambda)+\sigma^2\right\}\\ &=\frac{1}{n}\left\{\|\mathbf{y}-\hat{\mathbf{f}}_\lambda\|^2+2\sigma^2tr(S_\lambda)+\sigma^2\right\}\end{aligned} \qquad (7)$$

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

$$\hat{\sigma}^2=\hat{\sigma}^2_{\hat{\lambda}}=\frac{\sum_{\mathrm{i}=1}^{\mathrm{n}}\left(y_\mathrm{i}-\hat{f}_{\hat{\lambda}}(x_\mathrm{i})\right)^2}{tr\left(I-S_{\hat{\lambda}}\right)}=\frac{\|(S_{\hat{\lambda}}-I)\boldsymbol{y}\|^2}{tr\left(I-S_{\hat{\lambda}}\right)} \qquad (8)$$

where $\hat{\lambda}$ is pre-chosen with any of the CV, GCV or AIC$_C$ criteria ($\hat{\lambda}$ is an estimate of λ) For reference, see Lee (2002), Lee (2003), and Wahba (1990).

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

$$\begin{aligned}R(\mathbf{f},\hat{\mathbf{f}}_\lambda)&=\frac{1}{\mathrm{n}}E\left\|\mathbf{f}-\hat{\mathbf{f}}_\lambda\right\|^2\\ &=\frac{1}{n}\left\{\|(S_\lambda-I)\mathbf{f}\|^2+\sigma^2tr(S_\lambda S_\lambda^T)\right\}\end{aligned} \qquad (9)$$

It is straightforward to show that $R\left(\mathbf{f},\hat{\mathbf{f}}_\lambda\right)=E\left\{C_\mathrm{p}(\lambda)\right\}$ Because the risk $R(\mathbf{f},\hat{\mathbf{f}}_\lambda)$ is an unknown quantity, so-called risk is now estimated by computable quantity $R(\hat{\mathbf{f}}_{\lambda_\mathrm{p}},\hat{\mathbf{f}}_\lambda)$. The obtained expression for $R(\hat{\mathbf{f}}_{\lambda_\mathrm{p}},\hat{\mathbf{f}}_\lambda)$ is

$$\begin{aligned}R(\hat{\mathbf{f}}_{\lambda_\mathrm{p}},\hat{\mathbf{f}}_\lambda)&=\frac{1}{\mathrm{n}}E\left\|\hat{\mathbf{f}}_{\lambda_\mathrm{p}}-\hat{\mathbf{f}}_\lambda\right\|^2\\ &=\frac{1}{n}\left\{\left\|(S_\lambda-I)\hat{\mathbf{f}}_{\lambda_\mathrm{p}}\right\|^2+\hat{\sigma}^2_{\lambda_\mathrm{p}}tr(S_\lambda S_\lambda^T)\right\}\end{aligned} \qquad (10)$$

where $\hat{\sigma}^2_{\lambda_p}$ and $\hat{\mathbf{f}}_{\lambda_\mathrm{p}}$ are the appropriate *pilot estimates* for σ^2 and $\mathbf{f}$, respectively. The pilot λ_p selected by classical methods is used for computation of the pilot estimates.

4 Simulation Study

In this section we use Monte Carlo simulations to examine the properties of the various selection criteria. The simulations investigates the performance of the selectors as they relate to the four regression functions and standart deviations of the errors. The four regression functions for the experimental setup adopted here were used in earlier Monte Carlo studies (Ruppert et al., 1995; Hart and Yi, 1996; Herrmann, 1997), but we used different error standart deviations; besides generated the samples sized $n=200$ from each of regression functions. The number of replications were 300. For each simulated data sets, the mean squared-errors (MSE) was used for evaluate the quality of any curve estimate $\hat{f}$. Whether the difference between the MSE median values of any two selection methods is significant or not were tested with paired Wilcoxon tests. In this way, methods which select the best smoothing parameter were determined by evaluating so-called selection methods.

4.1 Experimental setup

The experimental setup applied here was designed to study the effects of the factors over the following four regression functions:

1. $y_{ir} = f(x_i) + \sigma_r \varepsilon_i = \sin(15\pi x_i) + \sigma_r \varepsilon_i$,

2. $y_{ir} = f(x_i) + \sigma_r \varepsilon_i = 1 - 48x_i + 218x_i^2 - 315x_i^3 + 145x_i^4 + \sigma_r \varepsilon_i$

3. $$y_{ir} = f(x_i) + \sigma_r \varepsilon_i = 0.3\exp\left\{-64(x_i - 0.25)^2\right\} + 0.7\exp\left\{-256(x_i - 0.75)^2\right\} + \sigma_r \varepsilon_i$$

4. $y_{ir} = f(x_i) + \sigma_r \varepsilon_i = 10\exp(-10x_i) + \sigma_r \varepsilon_i$

Where $\sigma_r = 0.02 + 0.04\,(r-1)^2,\ i = 1,\ldots,n$ and $r = 1,\ldots,6;\ x_i = \dfrac{i-0.5}{n};\ \varepsilon_i \sim iid\, N(0,1)$

The simulation study was performed with MATLAB program and the experimental setup was designed as the following way:

- To see the performance of the selection methods for each set of experiments, factor level r is changed six times ($r = 1,\ 2,\ 3,\ 4,\ 5,\ 6$).
- We generated sample sizes n = 200 for each level of the regression functions. The number of replications were m = 300 for each data set of the generated samples
- We computed the appropriate smoothing spline estimators $\hat{\mathbf{f}}_\lambda$ in equality (3) by selecting the smoothing parameter λ which minimizes the selection methods.
- We used the MSE values to evaluate $\hat{\mathbf{f}}_\lambda$ computed according to each of the selection criterion:

$$MSE = \frac{1}{n}\sum_{\mathrm{i}=1}^{\mathrm{n}}\left\{f(x_{\mathrm{i}}) - \hat{f}_\lambda(x_{\mathrm{i}})\right\}^2 ,$$

$$(\hat{f}_\lambda(x_i) = (\hat{\mathbf{f}}_\lambda)_i). \tag{11}$$

- Paired Wilcoxon test was applied to test whether MSE values considered as the performance measure of any two methods are significant or not.
- By considering 4 functions and 6 factor levels, we performed totally 24 numerical experiments.

4.2 Empirical performance of the selection criteria

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 in the following manner: 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 3-5. 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.

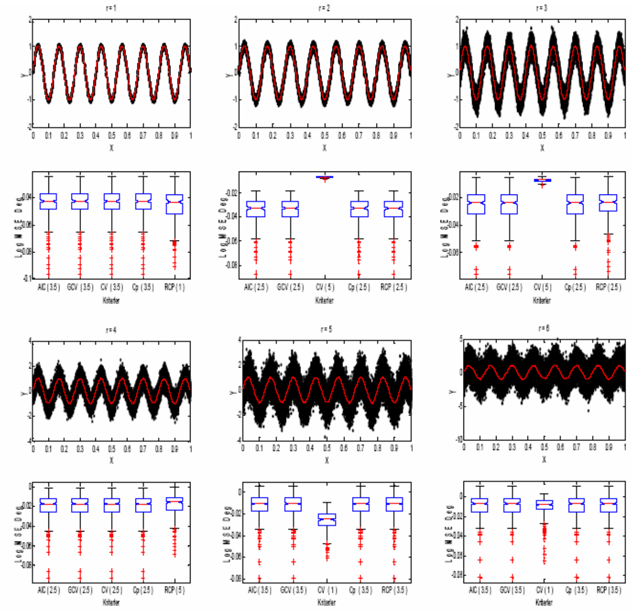

Figure 1: In each of fanels the top row plots display the true regression function together with one typical simulated data set. The bottom row plots display the boxplots of the $\log_e$ MSE values for, from left to right, AIC$_c$, GCV, CV, Cp and RCP criteria. The numbers below the boxplots are the paired Wilcoxon test rankings.

According to the simulation, the resulting rankings are given in Figures 1-4 to five selection methods, which is express in section 3. For 24 different simulation experiments, the averaged ranking values of the selection methods according to Wilcoxon tests are tabulated in Table 1.

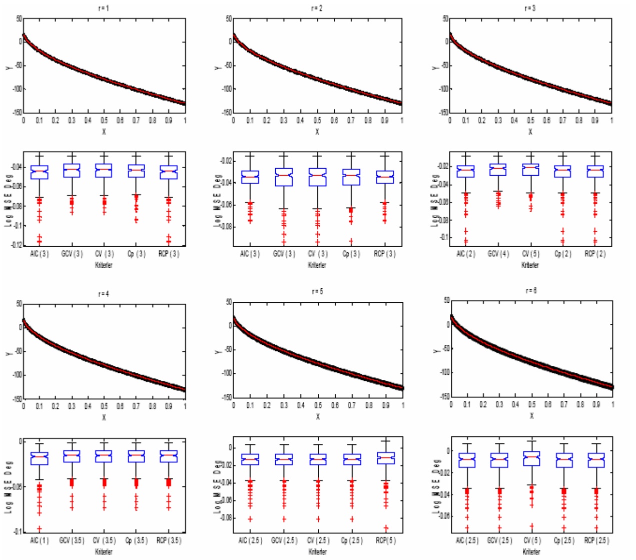

Figure 2: Similar to Figure 1, but the simulation results of the regression function 2

5 Conclusions and Recommendations

According to all regression functions and general means in Table 1, it may seem that AIC$_c$ is the best criter. However, AIC$_c$ was not better than the other criteria in the 24 different smimulation configurations. For example, when Figure 1 is examine, RCP has taken the best ranking for r =1, then CV has taken the best ranking for r = 4 and 5. As similar, the same postions may seem into the other figures. In brief, from a closer inspection of the simulation results, the following observations were made:

- According to all regresion and functions general means, AIC$_c$ has had the superior emprical performance.

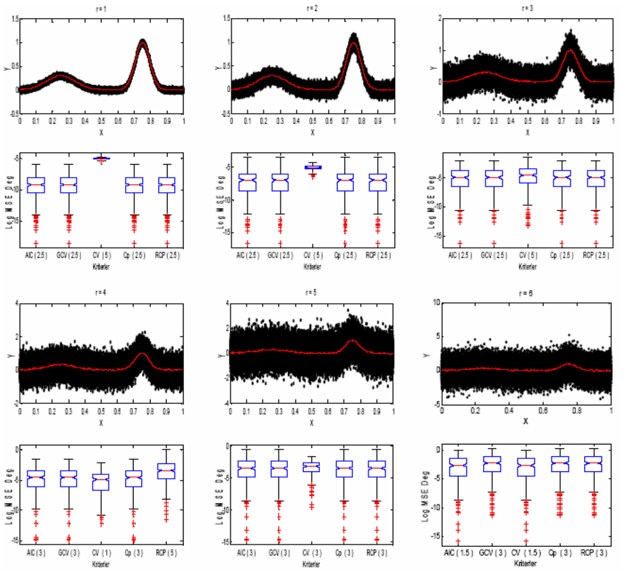

Figure 3: Similar to Figure 1, but the simulation results of the regression function 3

- CV criterion indicated the worst performance according to all regression functions and general means;
- The two selection methods, GCV and Cp criteria, gave very similar results according to all regression functions;
- The two selection methods, GCV and Cp criteria, gave a very good performance after the AIC$_c$.

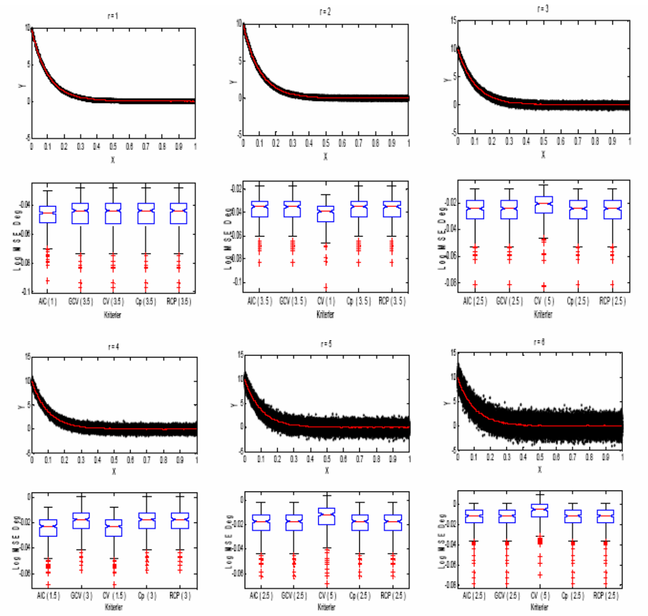

Figure 4: Similar to Figure 1, but the simulation results of the regression function 4

Table 1: Averaged Wilcoxon test ranking values for the six selection criteria

	Func.1	Func.2	Func.3	Func.4	Average
AIC$_c$	**2.667**	**2.333**	**2.500**	**2.083**	**2.583**
GCV	**2.667**	3.083	2.750	2.917	2.854
CV	3.000	3.667	3.417	3.500	3.396
Cp	**2.667**	2.750	2.750	2.917	2.761
RCP	3.000	3.167	3.083	2.917	3.045

According this simulation results, our recommendation is as fallows: In all regressions moodels and general means, use the AIC$_c$ criterion because of its superior emprical performance; otherwise use one of the GCV and Cp criteria that emprical performance is very close to AIC$_c$.

References:

[1] Cantoni, E., Ronchetti, Resistant Selection of Smoothing Parameter for Smoothing Splines, *Statis. Comput*, Vol.11, 2001, pp. 141-146.

[2] Craven.,P., and Wahba, G., Smoothing Noisy Data with Spline Functions, *Num. Math.*, Vol.31, 1997, pp. 377-403.

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

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

[5] Hart, J. D., and Yi, S., One-sided cross-validation, Unpublished, 1996.

[6] Herrmann, E., Local bandwidht choice in kernel regression estimation, *J. Comput. Graph. Statist*, Vol.24, pp., 1997, 1619-1647.

[7] Hurvich C. M., and Simonoff J. S., and Tasi C.-L., Smoothing parameter selection in nonparametric regreession using an improved Akaike İnformation criterion, *J.R. Statist. Soc*. B, Vol.60, 1998, 271-293.

[8] Kou, S. C., On the effiency of selection criteria in spline regression, *Probab. theory relat. fields*, Vol.127, 2003, pp. 153-176.

[9] Lee, T.C.M., Smoothing prameter selection for soomthing splines: a simulation study, *Computational statistics & Data analysis*, Vol.42, 2003, pp. 139-148.

[10] Lee, T.C.M., İmproved Smoothing spline Regression by Combining Estimates of different Smoothness, *Satatistics & Probability Letters*, 2004.

[11] Lee, T.M.C. Ve Solo, V., Bandwidth selection for local linear regression: a simulation study, *Comput. Statist*, Vol.14, 1999, pp. 515-532.

[12] Ruppert, D., Sheather, S. J. and Wand, M. P., An effective bandwidht selector for local least square regression, *J. Am. Statist.*, Vol.90, 1995, pp. 1257-1270.

[13] Wahba, G., Spline Model For Observational Data. Siam, Philadelphia Pa., 1990.