International Journal of Engineering, Mathematical and Physical Sciences
ISSN: 2517-9934
Vol:1, No:12, 2007

A Comparison of the Nonparametric Regression Models using Smoothing Spline and Kernel Regression

Dursun Aydin

Abstract**—**This paper study about using of nonparametric models for Gross National Product data in Turkey and Stanford heart transplant data. It is discussed two nonparametric techniques called smoothing spline and kernel regression. The main goal is to compare the techniques used for prediction of the nonparametric regression models. According to the results of numerical studies, it is concluded that smoothing spline regression estimators are better than those of the kernel regression.

Keywords**—**Kernel regression, Nonparametric models, Prediction, Smoothing spline.

I. INTRODUCTION

IT is considered the following nonparametric regression model

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

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.

The basic aim of the nonparametric regression is to estimate unknown function $f \in C^2[a,b]$ (of all functions f with continuous first and second derivatives) in model (1). In parametric regression of the form $y = f(x) + \varepsilon$, where f is some known, smooth function, the modeler must determine the appropriate form of f. In nonparametric regression, f is some unknown, smooth function and is unspecified by the modeler. A data-driven technique determines the shape of the curve. This chapter describes two different estimation techniques of nonparametric regression model: Smoothing spline regression and Kernel regression.

Estimations of the model (1) using smoothing spline regression and kernel regression are discussed in sections 2 and 3. Section 4 shows the some of the performance criteria associated with the models. Numerical studies are conducted in section 5 for two real data, whereas conclusion is offered in section 6.

II. SMOOTHING SPLINE REGRESSION

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

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

for pre-specified value $\lambda > 0$. The first term in equation (2) denotes the residual sum of the squares 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 known as the smoothing parameter. As λ varies from 0 to $+\infty$, the solution varies from interpolation to a linear model. As $\lambda \to +\infty$, the roughness penalty dominates in (2) and the spline estimate is forced to be a constant. As $\lambda \to 0$, 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* represented by $\sum_{i=1}^{n}\{y_i - f(x_i)\}^2$ and *smoothnees of the estimate* measured by $\int_a^b \{f''(x)\}^2 dx$.

In this paper, it is used $\mathbf{R}$ and $\mathbf{S}-Plus$ programs for choice of the smoothing parameter. They choose the smoothing parameter, using either ordinary or generalized cross validation (see, [1], or supply an alternative argument, df, which specifies *the degrees of freedom* for smooth.

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 becomes a suitable approach of function f in model 1. Let $\mathbf{f} = (f(x_1), \ldots, f(x_n))$ be the vector of values of function f at the knot points $x_1, \ldots, x_n$. The smoothing spline estimate $\hat{\mathbf{f}}_\lambda$ of this vector or the fitted values for data $\mathbf{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$ are given by

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

where $\hat{f}_\lambda$ is a natural cubic spline with knots at $x_1,...,x_n$ 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 [2], [3] and [4].

III. Kernel Regression

The main philosophy of nonparametric regression is to estimate the regression function f using a weighted average of the raw data where the weights are a function of distance in the x-space. In particular, the weights are a decreasing function of distance. A weighting scheme of this type is proposed by Nadaraya -Watson (1964) in which the weight associated with observations y_j, for prediction at x_i is given by:

$$w_{ij} = \frac{K\left(\frac{x_i - x_j}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{x_i - x_j}{h}\right)} = \frac{K(u)}{\sum_{j=1}^{n} K(u)} \tag{4}$$

where $K(u)$ is a decreasing function of u, and $h > 0$ is called the bandwidth or smoothing parameter. $K(u)$, the kernel function, may be taken to be a probability density function such as a Gaussian. The kernel function should be symmetric (see [7] and [8]). The kernel estimate of the function f in model (1) at the any point x_i is expressed as

$$\hat{y}_i = \hat{f}(x_i) = \sum_{j=1}^{n} w_{ij} y_j = w_j{}' \mathbf{y}, \quad i = 1, 2, ..., n \tag{5}$$

Each of the n data points is assigned a distinct weight w_{ij}, $j = 1, 2, ..., n$ for any point of fit x_i In matrix notation, equation (5) is given by

$$\hat{\mathbf{f}} = \mathbf{W}\mathbf{y} \tag{6}$$

where

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

and $\mathbf{w}_i' = (w_{i1},...,w_{in})$ The matrix $\mathbf{W}$ is denoted as the *kernel hat matrix* or kernel smoother matrix. Similarly to ordinary least square (OLS) where the hat matrix is used for transform the y_j's to the $\hat{y}_i$'s, the kernel "hat" matrix is used for transform the y_j's to the $\hat{y}_i$'s. Kernel predictions at an any point, x_i, may be obtained by using equation (5), replacing the "i" by "1". Then kernel prediction at any point x_1 can be written as following:

$$\hat{f}(x_1) = \mathbf{w}_1'\mathbf{y} = (w_{11},...,w_{1n}) \begin{bmatrix} y_1 \\ . \\ . \\ . \\ y_n \end{bmatrix} \tag{7}$$

As mentioned above, similarly to smoothing spline estimation in equation (3), the kernel estimation of the nonparametric regression that expressed in (1) is given by

$$\hat{\mathbf{f}} = \begin{bmatrix} f(x_1) \\ . \\ . \\ . \\ f(x_n) \end{bmatrix}_{n\times 1} = \begin{bmatrix} w_{11} & . & . & . & w_{1n} \\ . & & & & . \\ . & & & & . \\ . & & & & . \\ w_{n1} & . & . & . & w_{nn} \end{bmatrix}_{n\times n} \begin{bmatrix} y_1 \\ . \\ . \\ . \\ y_n \end{bmatrix}_{n\times 1} \tag{8}$$

A. Selection of Kernel Function

The predictions of the kernel regession comes from the fact that the estimated regression function at x_i is obtained by taking a weighted average of the y_j values where the weights w_{ij} are produced by the kernel function, $K(u)$. It is concluded that the selection of smoothing parameter (bandwidth) is much more important than the selection of kernel function for the performance of the kernel regression estimator (see [9] and [1]). The kernel function, $K(u)$ is typically chosen to be nonnegative, symmetric about zero, continuous and twice differentiable. Some alternative popular kernel functions are given in the Table I.

TABLE I
Alternative Kernel Functions

Kernel	Explict Form
Gaussian Kernel	$K(u) = \frac{1}{\sqrt{2\pi}} \exp(-u^2), u \in [-\infty, \infty]$
Uniform Kernel	$K(u) = \frac{1}{2}, \ u \in [-1, 1]$
Triangular	$(1 - \lvert u \rvert), u \in [-1, 1]$
Epanechnikv	$\frac{3}{4}(1 - u^2), \ u \in [-1, 1]$

Quartic	$\frac{15}{16}(1-u^2)^2,\ u \in [-1,1]$
Triweight	$\frac{35}{32}(1-u^2)^3,\ u \in [-1,1]$

Since the selection of kernel function is not critical for the performance of the kernel regression estimator, it will be used the simplified Gaussian kernel. The kernel mentioned here is given by

$$K(u)=\frac{1}{\sqrt{2\pi}}\exp\left(-u^2\right)=\frac{1}{\sqrt{2\pi}}\exp\left(-\left(\frac{x_i-x_j}{h}\right)^2\right) \tag{9}$$

In this case, Nadarya –Watson (1964) kernel estimation at any point x_i may be obtained by

$$\hat{y}_i=\hat{f}(x_i)=\frac{\sum_{j=1}^{n}\frac{1}{\sqrt{2\pi}}\exp\left[-\left(\frac{x_i-x_j}{h}\right)^2\right]y_j}{\sum_{j=1}^{n}\frac{1}{\sqrt{2\pi}}\exp\left[-\left(\frac{x_i-x_j}{h}\right)^2\right]}=w_i'\mathbf{y} \tag{10}$$

As mentioned section 2, it is used $\mathbf{R}$ and $\mathbf{S}-Plus$ programs to perform kernel regression. They use the $"ksmooth"$ function. The some kernels available in $"ksmooth"$ are shown in Table I. It is recommended that “normal kernel” because it is simple to calculate. In practice, selection of h (bandwidth) is usually done by trial and error, or this procedure can be done by selection criteria such as cross validation and generalized cross validation [10].

IV. Performance Criteria of the Models

The performance of the model is related with how close are the prediction values for test data and the observed values. Three different prediction consistency criteria are used in order to compare the performances of obtained smoothing spline and kernel regression. These are mean square error (MSE) (or root mean square error (RMSE)), mean absolute error (MAE) and mean absolute percentage error (MAPE) respectively. These criteria are defined as follows:

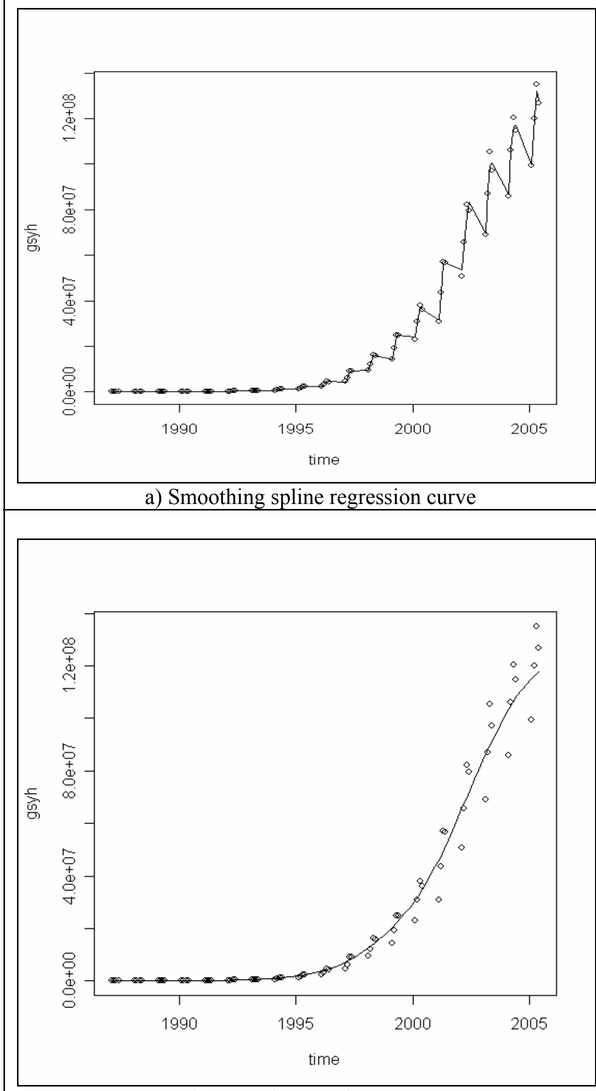
- $MSE=\frac{1}{T}\sum_{t=1}^{T}(y_t-\hat{y}_t)^2$ or $RMSE=\sqrt{MSE}$.
- $MAE=\frac{1}{T}\sum_{t=1}^{T}\left|y_t-\hat{y}_t\right|$.
- $MAPE=\frac{1}{T}\sum_{t=1}^{T}\frac{\left|y_t-\hat{y}_t\right|}{\left|y_t\right|}$ (%100)

V. Numerical Studies

In this section it is presented two examples, the one is gross national product at Fixed (1987) Prices in Turkey (Quarterly, YTL Thousand) (see, www.tcmb.gov.tr). Data related to variables used in this study consists of Quarterly time series which starts January, 1987 and ends December 2005, comprising $n=76$ observations. The other one is heart transplant data. These data have been taken from [11], and it contains survival times of patients on the waiting list for the Stanford heart transplant program. The variables mentioned here are defined as follows:

gsyh : Gross National Product (YTL Thousand)
time: Data quarterly from January 1987Q1 up to December 2005Q4
Age : Age of the patients
SurvivalTime: Survival Times in days of patients

A. Empirical Results



a) Smoothing spline regression curve

b) Kernel regression curve

Fig. 1 Plots of the covariates for gross national product data.

First of all, it is discussed that prediction of a non-parametric regression model using both smoothing spline and

Kernel regression for gross national product data. Results obtained with these models are given in Fig. 1 and their some of the performance criteria are also given the following Table II.

TABLE II
PERFORMANCE VALUES OF THE MODELS FOR GROSS NATIONAL PRODUCT DATA

Performance Criteria	Techniques	
	Spline Regression	Kernel Regression
MSE	1,76391E+12	2,39958E+13
RMSE	8,81954E+11	1,19979E+13
MAE	608277,872	2447521
MAPE	8,989789	22,11507

According to Table II, it is shown that performance criteria values obtained by smoothing spline regression are smaller than results of the Kernel regression. Hence, it can be say that smoothing spline regression is better than kernel regression for prediction of these models.

Figs. 1 (a) and (b) show the estimates (solid) and observations for smoothing spline and kernel regression, respectively for gross national product data in Turkey. The time variable in nonparametric model can be only displayed graphically, because it can't be expressed as parametric. As shown Figs. 1 (a) and (b), shape of the effects of time on gsyh is appears as a curve. Each plotted curve (regression curve) in these figures (a-b) is a contribution of a term to nonparametric predictor. These curves are closely following the real observations. This situation indicates that estimated values are very good. However, estimated values in Fig. 1(a) are better than Fig. 1(b), because the curve in Fig. 1(a) is very closely following the real observations.

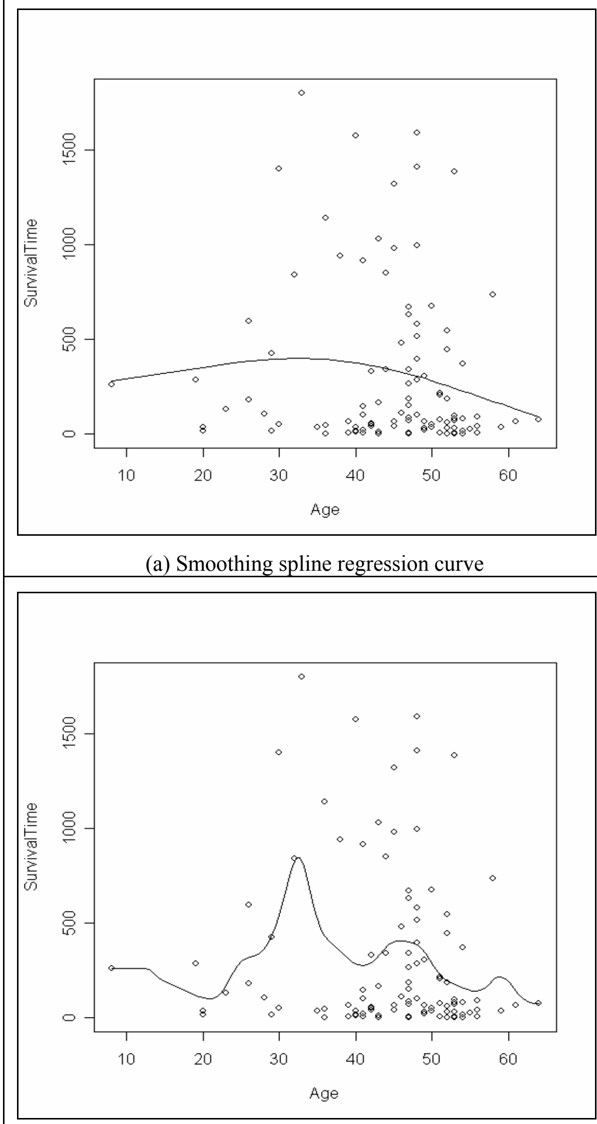
Secondly, It is discussed that prediction of nonparametric regression model using the same techniques for Stanford heart transplant data. Prediction results obtained by these models are given in Fig. 2 and their some of the performance criteria values are also given the Table III.

TABLE III
PERFORMANCE VALUES OF THE MODELS FOR STANFORD HEART TRANSPLANT DATA

Performance Criteria	Techniques	
	Spline Regression	Kernel Regression
MSE	174123,8	266454,9
RMSE	87061,9	133227,4
MAE	315,0	378,8
MAPE	1679,4	1494,8

According to Table III, it is shown that most of the performance criteria values obtained by smoothing spline regression are again smaller than results of the Kernel regression. Hence, it can be say that smoothing spline regression is better than kernel regression for prediction **Survival Times** in days of patients.

Figs. 2 (a) and (b) show the estimates (solid) and observations for smoothing spline and kernel regression, respectively for Stanford heart transplant data. Nonparametric model mentioned here relates **Survival Time** in days to **Age**. As shown the plots produced by spline and kernel regression models, Fig. 2 indicates a curvature in the relationship between **Age** and **survival time**. Effect on **Survival time** in days of **age** in spline fit is rather moderate, while the effect in kernel fit is rather bumpy. For this reason, it is advisable to use smoothing spline regression.



(a) Smoothing spline regression curve

(b) Kernel regression curve

Fig. 2 Plots of the covariates for Stanford heart transplant data

VI. CONCLUSION

In this paper, it has been discussed four nonparametric models based on smoothing spline and kernel regression. It is concluded that prediction of the two real data by these models. Results obtained with smoothing spline models have been compared to kernel regression models. In brief, from a closer inspection of the empirical results, the following observations were made:

- According to MSE, RMSE, MAE and MAPE values for gross national product data, the nonparametric model based on smoothing spline has indicated a good performance;
- The estimates obtained by smoothing spline have also indicated a good result in terms of graphically for gross national product data;
- The nonparametric model based on smoothing spline has also indicated a good performance according to MSE, RMSE and MAE values for Stanford heart transplant data;
- The nonparametric model based on kernel regression has indicated a good performance according to only MAPE value for Stanford heart transplant data ;

These results emphasize that estimates based on smoothing spline technique is better than the kernel regression.

REFERENCES

[1] Hastie, T.J. and Tibshirani, R.J., Generalized Additive Models, Chapman & Hall /CRC, 1999.
[2] Wahba, G., Spline Model For Observational Data, Siam, Philadelphia Pa., 1990.
[3] Green, P. J. and Silverman, B. W., Nonparametric Regression and Generalized Linear Models, Chapman & Hall, 1994.
[4] Eubank, R. L., Nonparametric Regression and Smoothing Spline, Marcel Dekker Inc., 1999.
[5] Nadarya, E.A., *On Estimating Regression,* Theory Pb. Appl., Vol.10, 1964, pp. 186-190.
[6] Watson, G. S., *Smooth Regression Analysis,* Sankhya, Series A, Vol.26, 1964, pp.359-372.
[7] Yatchew, A., Semiparametric Regression for the Applied Econometrician, Cambridge University Pres, 2003.
[8] Wand, M,. P. Ve M. C. Jones, *Kernel Smoothing,* New York: Chapman and Hall, 1995.
[9] Hardle, W., *Applied Nonparametric Regression*, Cambridge University Press, Cambridge ,1991.
[10] Schimek, M. G., Smoothing and Regression, Jhon. Willy. & Sons, 2000.
[11] John Crowley, Marie Hu, Covariance Analysis of Heart Transplant Survival Data**,** *Journal of the American Statistical Association*, Vol. 72, No. 357 Mar., 1977, pp. 27-36.

Dursun Aydin was born in 1969 in Ardahan, Turkey. He graduated from Anadolu University Science Faculty Statistics in 1992 in Eskişehir, Turkey. Then he graduated from Marmara University Graduate School of Social Sciences in 2001 İstanbul, Turkey. Finally he graduated Anadolu University Graduate School of the Natural and Applied Sciences Science in 2005.

He has been working for 12 years at Anadolu University as a teaching assistant, but now he is an instructor doctor at the same university. He has published a lot of papers in journals of various universities of Turkey and several international papers.

His research area includes: Multivariate Statistics, Basics Statistics, Parametric Regression, Nonparametric Regression, Semiparametric Regression, Generalized Additive Models, Time Series.