Journal of Sciences, Islamic Republic of Iran 31(2): 175 - 183 (2020) http://jsciences.ut.ac.ir
University of Tehran, ISSN 1016-1104

Use of Two Smoothing Parameters in Penalized Spline Estimator for Bi-variate Predictor Non-parametric Regression Model

A. Islamiyati[1*], N. Sunusi[1], A. Kalondeng[1], F. Fatmawati[2], N. Chamidah[2]

[1] *Department of Statistics, Faculty of Mathematics and Natural Sciences, Hasanuddin University, Makassar, Indonesia*
[2] *Department of Mathematics, Faculty of Sciences and Technology, Airlangga University, Surabaya, Indonesia*

Received: 10 August 2019 / Revised: 10 March 2020 / Accepted: 18 April 2020

Abstract

Penalized spline criteria involve the function of goodness of fit and penalty, which in the penalty function contains smoothing parameters. It serves to control the smoothness of the curve that works simultaneously with point knots and spline degree. The regression function with two predictors in the non-parametric model will have two different non-parametric regression functions. Therefore, we propose the use of two smoothing parameters in the bi-variate predictor non-parametric regression model. We demonstrated its ability through longitudinal data simulation studies with a comparison of one smoothing parameter. It was done on several numbers of subjects with repeated measurements. The generalized cross validation value which is a measure of the model's ability is poured through the box plot. The results show that the use of two smoothing parameters is more optimal than one smoothing parameter. It was seen through a smaller generalized cross validation value on the use of two smoothing parameters. Application of blood sugar level data for patients with two smoothing parameters produced a penalized spline bi-variate predictor regression model with several segments of change patterns. There are five patterns at the time of treatment and blood pressure with the number of smoothing parameters is two, namely 0.39 and 0.73.

Keywords: Bi-variate; Longitudinal Data; Penalized Spline; Smoothing Parameter.

Introduction

The penalized spline estimator as one of the estimators in non-parametric regression model has advantages in visual and estimation. The ability of the estimator is caused by smoothing parameters and knot points simultaneously in controlling the smoothness of the curve [1]. The ability of the penalized spline to estimate the regression curve has been seen through the asymptotic properties of the estimator in the cross

* Corresponding author: Tel: +628124261584; Fax: +62 411 585188; Email: annaislamiyati701@gmail.com

section data [2, 3]. In addition, researchers have also developed robust penalized spline to overcome outliers in data [4, 5]. In further development, Islamiyati et al. [6, 7] have developed penalized spline estimators in longitudinal data. For the problem of smoothing parameters in the penalty function, it has become a separate topic of research in the non-parametric regression model. The penalty function in spline regression can refer to the ridge regression concept for time series data [8,9]. Heckman and Ramsay [10] have worked on the penalty function of a linear differential operator with possibly nonconstant coefficients. Yao and Lee [11] have used smoothing parameters derived from the variance loss function. Aydin and Yilmaz [12] used a modified spline estimator for non-parametric regression models with right-censored data. Chamidah and Lestari [13], Chamidah and Rifada [14], Lestari et al. [15] and Chamidah et al. [16] have selected smoothing parameters through the minimum Generalized Cross Validation (GCV) method. However, the study only used one smoothing parameter with one predictor in the model.

There are several longitudinal data studies that have used multi-predictor. Lin and Zhang [17] used double penalized quasi likelihood and Ni et al. [18] used double penalized in semiparametric regression. Double penalized is intended as the use of two penalties to select significant variables in the model. However, the basis of its use in the study is not based on the predictor function that contains the knots. Lai and Wang [19] examined bi-variate penalized spline but they only used one smoothing parameter. Therefore, this article proposes the use of two smoothing parameters based on two predictors in the regression model. Each predictor function contains a truncated element that considers the optimal knot point. We will show the ability of two smoothing parameters through simulation studies by comparing GCV values from one smoothing parameter.

Next, we describe the material and methods of the bi-variate predictor non-parametric regression model of longitudinal data. After that, we describe the results and discussion about the estimation of bi-variate predictor regression model with the penalized spline estimator. In addition, we show the ability of two smoothing parameters in the penalized spline bi-variate predictor through simulation studies on polynomial function and sinus trigonometry. We simulate different predictor functions in longitudinal data and compare GCV values between one and two smoothing parameters. Furthermore, the model was applied to the blood sugar level data of diabetic patients based on the time of treatment and the blood pressure.

Material and Methods

Given a pair of observation data $\left(t_{ij1}, t_{ij2}, y_{ij}\right)$, with $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots m_i$, for bi-variate predictor non-parametric regression models in longitudinal data is as follows:

$$y_{ij} = f\left(t_{ij1}, t_{ij2}\right) + \varepsilon_{ij}, i = 1, 2, \ldots, n, j = 1, 2, \ldots, m_i \quad (1)$$

where y_{ij} is the response to i^{th} subject, j^{th} measurement, t_{ij1} is 1st predictor to i^{th} subject, j^{th} measurement, t_{ij2} is 2nd predictor to i^{th} subject, j^{th} measurement and ε_{ij} is error to i^{th} subject, j^{th} measurement.

The bi-variate predictor non-parametric regression function in longitudinal data in equation (1) is expressed as the sum of the functions of each predictor, namely:

$$f\left(t_{ij1}, t_{ij2}\right) = f\left(t_{ij1}\right) + f\left(t_{ij2}\right) \quad (2)$$

Based on equation (2), the model in equation (1) becomes:

$$y_{ij} = f\left(t_{ij1}\right) + f\left(t_{ij2}\right) + \varepsilon_{ij} \quad (3)$$

The function of $f\left(t_{ij1}\right)$ and $f\left(t_{ij2}\right)$ in equation (3) is a spline function where the shape is unknown. Its functions are as follows:

$$\left.\begin{aligned} f\left(t_{ij1}\right) &= \sum_{u_1=0}^{q_1} \beta_{u_1}\left(t_{ij1}\right)^{u_1} + \sum_{v_1=1}^{d_1} \beta_{(q_1+v_1)1}\left(t_{ij1} - K_{v_1}\right)_+^{q_1} \\ f\left(t_{ij2}\right) &= \sum_{u_2=0}^{q_2} \beta_{u_2}\left(t_{ij2}\right)^{u_2} + \sum_{v_2=1}^{d_2} \beta_{(q_2+v_2)2}\left(t_{ij2} - K_{v_2}\right)_+^{q_2} \end{aligned}\right\} \quad (4)$$

The spline function in equation (4) can be formed into $f\left(t_{ij1}\right) = \mathbf{X}_1\boldsymbol{\beta}_1$ and $f\left(t_{ij2}\right) = \mathbf{X}_2\boldsymbol{\beta}_2$. Matrix $\mathbf{X}_1$ is the matrix **X** in the first predictor containing t_{ij1} and knots point, $\mathbf{X}_1 = \left(\mathbf{1} \quad \mathbf{t}_1 \quad \ldots \quad \mathbf{t}_1^{q_1} \quad \left(\mathbf{t}_1 - K_{11}\right)_+^{q_1} \quad \ldots \quad \left(\mathbf{t}_1 - K_{d_1 1}\right)_+^{q_1}\right)$. Vector $\boldsymbol{\beta}_1$ is spline regression coefficient for 1st predictor. Matrix $\mathbf{X}_2$ is the matrix **X** in the second predictor, $\mathbf{X}_2 = \left(\mathbf{1} \quad \mathbf{t}_2 \quad \ldots \quad \mathbf{t}_2^{q_2} \quad \left(\mathbf{t}_2 - K_{12}\right)_+^{q_2} \quad \ldots \quad \left(\mathbf{t}_2 - K_{d_2 2}\right)_+^{q_2}\right)$. Vector $\boldsymbol{\beta}_2$ is spline regression coefficient for 2nd predictor. As a result the spline function $f\left(t_{ij1}, t_{ij2}\right)$ in equation (2) can be expressed as a vector as follows:

$$\mathbf{f}\left(\mathbf{t}_1, \mathbf{t}_2\right) = \mathbf{X}\boldsymbol{\beta} \quad (5)$$

where $\mathbf{X} = \left(\mathbf{X}_1 \quad \mathbf{X}_2\right)$ and $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1 \quad \boldsymbol{\beta}_2\right)^T$.

Based on equation (5), we can state the bi-variate predictor non-parametric regression model in longitudinal data in equation (1) in the form of a matrix, namely:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad (6)$$

Results and Discussion

Penalized Spline Estimator with Two Smoothing Parameters

We estimate equation (6) using the penalized least squares (PLS) estimator. The PLS estimator combines the function of goodness of fit and the penalty function which we give a symbol with P, namely:

$$\mathrm{P} = \boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} + \lambda_1^* \int_{a_1}^{b_1} \left[g^{c_2}\left(t_{ij1}\right)\right]^2 dt_{ij1} + \lambda_2^* \int_{a_2}^{b_2} \left[g^{c_2}\left(t_{ij2}\right)\right]^2 dt_{ij2} \qquad (7)$$

Equation (7) shows the estimation criteria of the penalized spline which contains two refining parameters. The component $\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon}$ is a goodness of fit function of a bi-variate predictor non-parametric regression model in longitudinal data. That is the sum of quadratic errors obtained from equation (6) and it can be stated as follows.

$$\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} = \mathbf{y}^T\mathbf{y} - 2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}\,.$$

Furthermore, the component $\lambda_1^* \int_{a_1}^{b_1} \left[g^{c_2}\left(t_{ij1}\right)\right]^2 dt_{ij1}$ is a penalty function for the 1st predictor and $\lambda_2^* \int_{a_2}^{b_2} \left[g^{c_2}\left(t_{ij2}\right)\right]^2 dt_{ij2}$ for the 2nd predictor. Both are expressed as a penalty function in the PLS criteria in the bi-variate predictor non-parametric regression model, as shown in Theorem 1.

Theorem 1.

If the penalty function for the 1st and 2nd predictors are stated by $\mathrm{L} = \lambda_1^* \int_{a_1}^{b_1} \left[g^{(c_1)}\left(t_{ij1}\right)\right]^2 dt_{ij1}$ and $\mathrm{M} = \lambda_2^* \int_{a_2}^{b_2} \left[g^{(c_2)}\left(t_{ij2}\right)\right]^2 dt_{ij2}$, then the penalty function in the penalized spline estimator in the bivariate predictors non-parametric regression model for longitudinal data are as follows:

$$\lambda_1^* \int_{a_1}^{b_1} \left[g^{(c_1)}\left(t_{ij1}\right)\right]^2 dt_{ij1} + \lambda_2^* \int_{a_2}^{b_2} \left[g^{(c_2)}\left(t_{ij2}\right)\right]^2 dt_{ij2} = \boldsymbol{\beta}^T\mathbf{D}_{\boldsymbol{\lambda}}\boldsymbol{\beta} \qquad (8)$$

where $\boldsymbol{\lambda} = \left(\lambda_1, \lambda_2\right), \lambda_1 = \lambda_1^* C_1, \lambda_2 = \lambda_2^* C_2$, λ_1 and λ_2 are smoothing parameters for 1st and 2nd predictors, C_1 and C_2 are constant, a_1 and a_2 are lower boundary of the integrals of the functions $g\left(t_{ij1}\right)$ and $g\left(t_{ij1}\right)$, b_1 and b_1 are the upper limits of the integrals of the functions $g\left(t_{ij1}\right)$ and $g\left(t_{ij1}\right)$, functions $g\left(t_{ij1}\right)$ and $g\left(t_{ij1}\right)$ are the first and second predictor functions contained in the Sobolev Space.

Proof.

The penalty function in the criteria for the penalized spline bivariate predictor contains smoothing parameter for each predictor function. It is known that the penalty function for the 1st predictor is stated by $\mathrm{L} = \lambda_1^* \int_{a_1}^{b_1} \left[g^{(c_1)}\left(t_{ij1}\right)\right]^2 dt_{ij1}$ and for the 2nd predictor stated by $\mathrm{M} = \lambda_2^* \int_{a_2}^{b_2} \left[g^{(c_2)}\left(t_{ij2}\right)\right]^2 dt_{ij2}$. If a penalty function from bi-variate predictor spline regression is symbolized by JF, then it can be stated the additive penalty of predictor function as follows:

$$\mathrm{JF} = \lambda_1^* \int_{a_1}^{b_1} \left[g^{(c_1)}\left(t_{ij1}\right)\right]^2 dt_{ij1} + \lambda_2^* \int_{a_2}^{b_2} \left[g^{(c_2)}\left(t_{ij2}\right)\right]^2 dt_{ij2} \qquad (9)$$

where λ_1^* and λ_2^* are smoothing parameter, $g\left(t_{ij1}\right)$ and $g\left(t_{ij2}\right)$ is the element function of the Sobolev Space $W_2^{c_h}$ which is formed from the spline function.

The function of $g\left(t_{ij1}\right)$ in L is formed from the spline function as in equation (4) with the Dirac Delta function so we get:

$$g\left(t_{ij1}\right) = \beta_{01} + \ldots + \beta_{q_1 1}\left(t_{ij1}\right)^{q_1} + \beta_{(q_1+1)1}\left(t_{ij1} - K_{11}\right)_+^{q_1+1} + \ldots + \beta_{(q_1+d_1)1}\left(t_{ij1} - K_{d_1 1}\right)_+^{q_1+1} \qquad (10)$$

where truncated element of the function $g\left(t_{ij1}\right)$ is $\left(t_{ij1} - K_{11}\right)_+^{q_1+1}, \ldots, \left(t_{ij1} - K_{d_1 1}\right)_+^{q_1+1}$.

Next, we get the function $g\left(t_{ij2}\right)$ in M is:

$$g\left(t_{ij2}\right) = \beta_{02} + \ldots + \beta_{q_2 2}\left(t_{ij2}\right)^{q_2} + \beta_{(q_2+1)2}\left(t_{ij2} - K_{12}\right)_+^{q_2+1} + \ldots + \beta_{(q_2+d_2)2}\left(t_{ij2} - K_{d_2 2}\right)_+^{q_2+1} \qquad (11)$$

where truncated element of the function $g\left(t_{ij2}\right)$ is $\left(t_{ij2}-K_{12}\right)_{+}^{q_2+1},\ldots,\left(t_{ij2}-K_{d_2 2}\right)_{+}^{q_2+1}$.

Next, we describe the functions in equations (10) and (11) at L and M so we get:

$$C_1\sum_{v_1=1}^{d_1}\beta_{(q_1+v_1)1}^2 = C_1\left(\beta_{(q_1+1)1}^2+\beta_{(q_1+2)1}^2+\ldots+\beta_{(q_1+d_1)1}^2\right)=C_1\boldsymbol{\beta}_1^T\mathbf{D}_1\boldsymbol{\beta}_1$$

$$C_2\sum_{v_2=1}^{d_2}\beta_{(q_2+v_2)2}^2 = C_2\left(\beta_{(q_2+1)2}^2+\beta_{(q_2+2)2}^2+\ldots+\beta_{(q_2+d_2)2}^2\right)=C_2\boldsymbol{\beta}_2^T\mathbf{D}_2\boldsymbol{\beta}_2 \quad (12)$$

Based on equation (12), the penalty function on the 1st predictor was symbolized by L is expressed as $\mathrm{L}=\lambda_1^* C_1\boldsymbol{\beta}_1^T\mathbf{D}_1\boldsymbol{\beta}$ and the penalty function on the 2nd predictor was symbolized M is expressed as $\mathrm{M}=\lambda_2^* C_2\boldsymbol{\beta}_2^T\mathbf{D}_2\boldsymbol{\beta}_2$, consequently the JF in equation (9) becomes:

$$\mathrm{JF}=\lambda_1^* C_1\boldsymbol{\beta}_1^T\mathbf{D}_1\boldsymbol{\beta}_1+\lambda_2^* C_2\boldsymbol{\beta}_2^T\mathbf{D}_2\boldsymbol{\beta}_2 \quad (13)$$

If $\lambda_1^* C_1=\lambda_1$, $\lambda_2^* C_2=\lambda_2$, the lambda value as a scalar is multiplied by the matrix **D** as the diagonal matrix (0,1) and expressed as $\mathbf{D}_{\boldsymbol{\lambda}}$, then we can express equation (13) as follows:

$$\lambda_1^*\int_{a_1}^{b_1}\left[g^{(c_1)}\left(t_{ij1}\right)\right]^2 dt_{ij1}+\lambda_2^*\int_{a_2}^{b_2}\left[g^{(c_2)}\left(t_{ij2}\right)\right]^2 dt_{ij2}=\boldsymbol{\beta}^T\mathbf{D}_{\boldsymbol{\lambda}}\boldsymbol{\beta} \quad (14)$$

where $\boldsymbol{\beta}^T=\left(\boldsymbol{\beta}_1^T,\boldsymbol{\beta}_2^T\right)$, $\mathbf{D}_{\boldsymbol{\lambda}}=\begin{bmatrix}\mathbf{D}_{\lambda_1} & \mathbf{0}\\ \mathbf{0} & \mathbf{D}_{\lambda_2}\end{bmatrix}$, and $\boldsymbol{\beta}=\begin{bmatrix}\boldsymbol{\beta}_1\\ \boldsymbol{\beta}_2\end{bmatrix}$.

Parameter $\boldsymbol{\beta}$ is the spline regression coefficient of bivariate predictor longitudinal data, $\boldsymbol{\beta}_1$ is spline regression coefficient for 1st predictor and $\boldsymbol{\beta}_2$ is spline regression coefficient for 2nd predictor. Parameter $\boldsymbol{\lambda}=\left(\lambda_1,\lambda_2\right)$ is smoothing parameter for 1st and 2nd predictor. Matrix $\mathbf{D}_{\boldsymbol{\lambda}}$ is a diagonal matrix $(0,\lambda)$ that contains **D** matrix for 1st and 2nd predictor. Matrix $\mathbf{D}_{\lambda_1}=\mathrm{diag}\left(a_{01}\lambda_1,a_{11}\lambda_1,\ldots,a_{q_1 1}\lambda_1,a_{(q_1+1)1}\lambda_1,a_{(q_1+2)1}\lambda_1,\ldots,a_{(q_1+d_1)1}\lambda_1\right)$ is a diagonal matrix $(0,\lambda_1)$ for 1st predictor where $a_{01}=a_{11}=\ldots=a_{q_1 1}=0$ and $a_{(q_1+1)1}=a_{(q_1+2)1}=\ldots=a_{(q_1+d_1)1}=1$. Matrix $\mathbf{D}_{\lambda_2}=\mathrm{diag}\left(a_{02}\lambda_2,a_{12}\lambda_2,\ldots,a_{q_2 2}\lambda_2,a_{(q_2+1)2}\lambda_2,a_{(q_2+2)2}\lambda_2,\ldots,a_{(q_2+d_2)2}\lambda_2\right)$ is a diagonal matrix $(0,\lambda_2)$ for 2nd predictor where $a_{02}=a_{12}=\ldots=a_{q_2 2}=0$ and $a_{(q_2+1)2}=a_{(q_2+2)2}=\ldots=a_{(q_2+d_2)2}=1$.

The PLS criteria in equation (7) can be written in the form of vectors and matrices based on the penalty function in Theorem 1 as follows:

$$\mathrm{P}=\mathbf{y}^T\mathbf{y}-2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y}+\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\beta}^T\mathbf{D}_{\boldsymbol{\lambda}}\boldsymbol{\beta}$$

Next, we get the estimation results from $\boldsymbol{\beta}$ is :

$$\hat{\boldsymbol{\beta}}=\left(\mathbf{X}^T\mathbf{X}+\mathbf{D}_{\boldsymbol{\lambda}}\right)^{-1}\mathbf{X}^T\mathbf{y} \quad (15)$$

Based on equation (15), the estimation of bi-variate predictors non-parametric regression function in longitudinal data based on the penalized spline estimator is as follows:

$$\hat{\mathbf{f}}\left(\mathbf{t}_1,\mathbf{t}_2\right)=\mathbf{X}\hat{\boldsymbol{\beta}}=\mathbf{X}\left(\mathbf{X}^T\mathbf{X}+\mathbf{D}_{\boldsymbol{\lambda}}\right)^{-1}\mathbf{X}^T\mathbf{y} \quad (16)$$

where $\mathbf{A}\left(\boldsymbol{\lambda}\right)=\mathbf{X}\left(\mathbf{X}^T\hat{\boldsymbol{\Omega}}^{-1}\mathbf{X}+\mathbf{D}_{\boldsymbol{\lambda}}\right)^{-1}\mathbf{X}^T\hat{\boldsymbol{\Omega}}^{-1}$ is smoothing parameter matrix. Therefore, we can state the function estimation at (16) as an estimator of smoothing function is $\hat{\mathbf{f}}_{\boldsymbol{\lambda}}$, namely:

$$\hat{\mathbf{f}}_{\boldsymbol{\lambda}}=\mathbf{A}\left(\boldsymbol{\lambda}\right)\mathbf{y}. \quad (17)$$

Furthermore, the value of GCV of bi-variate predictor non-parametric regression model is based on the penalized spline estimator is as follows:

$$\mathrm{GCV}(\boldsymbol{\lambda})=\frac{\left(\mathbf{y}-\hat{\mathbf{f}}\left(\mathbf{t}_1,\mathbf{t}_2\right)\right)^T\left(\mathbf{y}-\hat{\mathbf{f}}\left(\mathbf{t}_1,\mathbf{t}_2\right)\right)}{\left(\sum_{i=1}^{n}m_i\right)^{-1}\left(tr\left(\mathbf{I}-\mathbf{A}\left(\boldsymbol{\lambda}\right)\right)\right)^2} \quad (18)$$

where

$$\mathbf{y}=\left(y_{11},y_{12},\ldots,y_{1m_1},y_{21},y_{22},\ldots,y_{2m_2},\ldots,y_{n1},y_{n2},\ldots,y_{nm_n}\right)^T.$$

Simulation Study

In this section, we simulate longitudinal data in the form of experimental functions in the 1st and 2nd predictors are quadratic and sinus trigonometry. The smoothing parameters used in this simulation are two parameters that correspond to different predictor functions at each predictor. The regression model is:

$$y_{ij}=f\left(t_{ij1}\right)+f\left(t_{ij2}\right)+\varepsilon_{ij},$$

where functions in each predictor are:

$$f\left(t_{ij1}\right)=-2.5+4t_{ij1}-0.2t_{ij1}^2$$

$$f\left(t_{ij2}\right)=3\sin\left(2\pi t_{ij2}\right)$$

We simulate data of observations from N = 80 to 1511 with the size of subjects, i.e. n = 5, 10, 20, 50 and 100, each subject measured from 10-20 times the measurement. We selected two smoothing parameters on the 1st and 2nd predictors based on minimum GCV values. Next, we show GCV values from the use of two and one smoothing parameter. We show the results in the box plots for various types of observations. Figure 1 shows the GCV value of using two smoothing parameters giving smaller values compared to one smoothing parameter. This means that the bi-variate predictor non-parametric regression model should involve two smoothing parameters. In a small subject, we see a significant difference in GCV values.

Next, we perform another simulation with the same predictor function form t_1 and t_2. The function that we created is a polynomial function like the following:

$$f\left(t_{ij1}\right) = -0.7 \ -1.5t_{ij1} + 2.5t_{ij1}^2$$
$$f\left(t_{ij2}\right) = 2.5 \ +3t_{ij2} - 0.2t_{ij2}^2$$

The number of subjects that were tried was the same as the first simulation, n = 5, 10, 20, 50, 75, and 100, where each subject was measured repeatedly starting 3-10 times. We do the iteration process up to 50 times for each number of subjects by recording the GCV value for each iteration. The GCV values shown from various simulations of the number of subjects are shown through the box plot in Figure 2. The purpose of this simulation is to show the ability to use two smoothing parameters better than one smoothing parameter in the bi-variate predictor case as seen in the previous simulation. Figure 2 shows the same conditions as the first simulation. We can see in the plot box, the GCV value for two smoothing parameters is smaller than the use of one smoothing parameter. This means that the

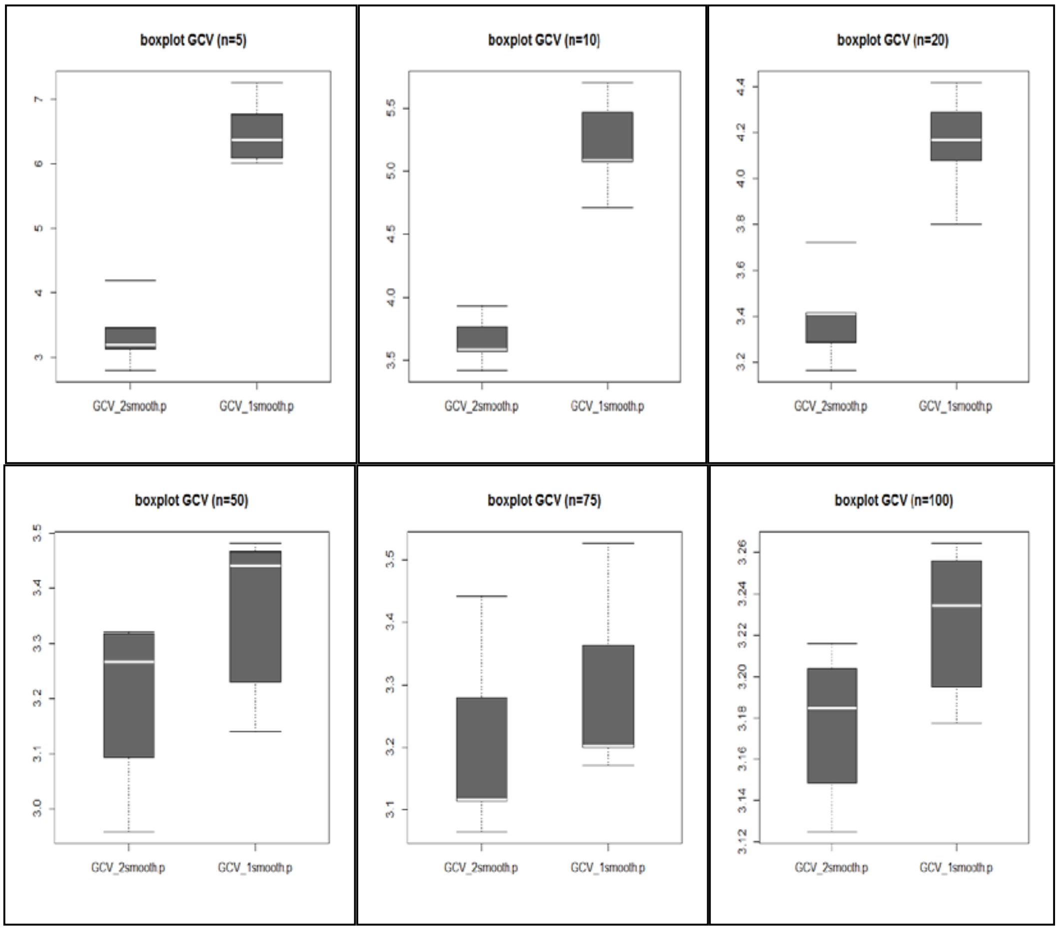

Figure 1. Box plot of GCV values of two and one smoothing parameter for first simulation, n = 5, N = 80; n = 10, N = 153; n = 20, N = 289; n = 50, N = 735; n = 75, N = 1133; n = 100, N = 1511.

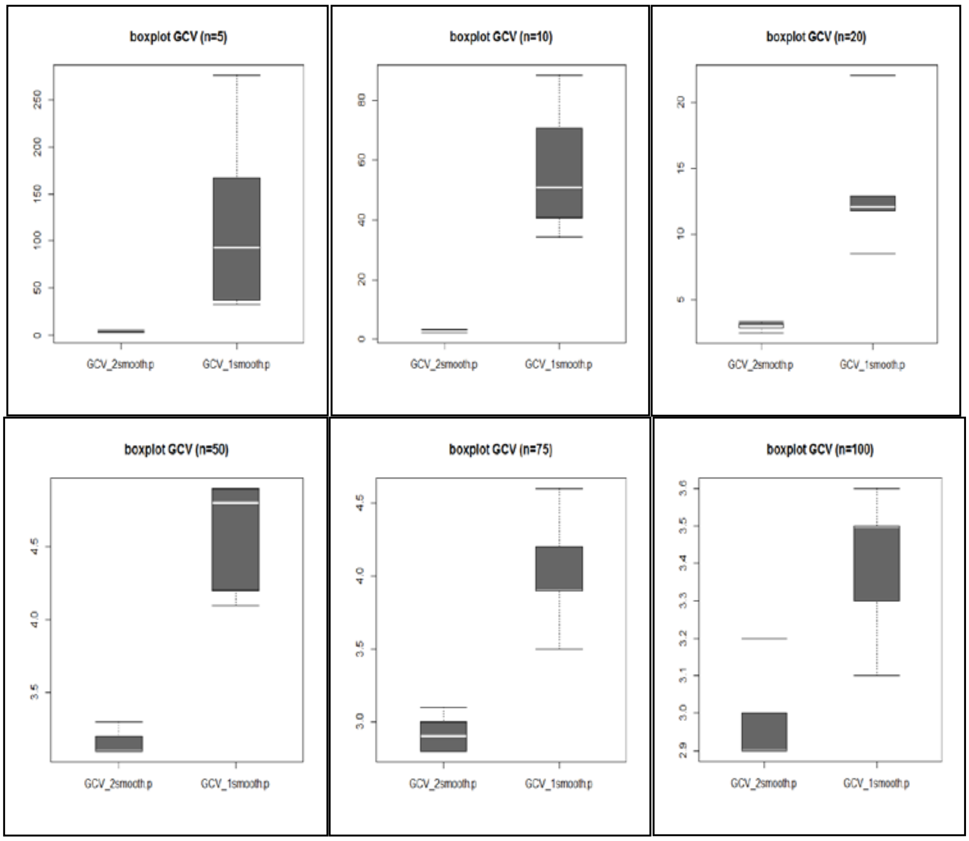

Figure 2. Box plot of GCV values of two and one smoothing parameter for second simulation, n = 5; n = 10; n = 20; n = 50; n = 75; n = 100.

use of two smoothing parameters in the bivariate predictor case is better than one parameter. Next, we can see in the box plot about the range of GCV values for each number of subjects. For n = 5, 10, 20, 50, 75, 100 each has a range of GCV values starting from 0-250, 0-80, 0-20, 0-4.9, 0-4.5, 0-3.0. This shows that as the number of subjects increases, the difference in GCV values for two and one smoothing parameter gets smaller. However, the use of two smoothing parameters still provides a minimum GCV value.

For example, we show the estimation of the regression curve in a plot of the smoothing parameter. That gives a minimum value for each predictor in Figure 3. This shows each predictor the function has a

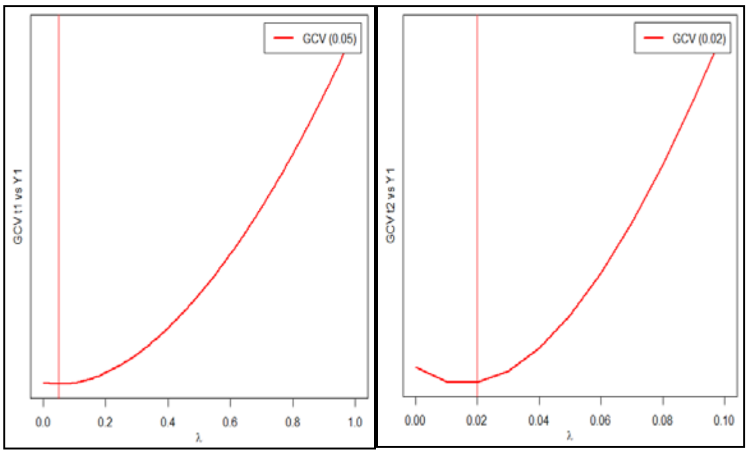

Figure 3. The plot of GCV with smoothing parameters for the 1st and 2nd predictors.

different smoothing parameter value, which is 0.05 and 0.02 for the 1st and the 2nd predictor function. This means that the non-parametric regression models of two predictors get better estimation results when using two smoothing parameters. Next, the estimation of the regression curve is shown in Figure 4 with different smoothing parameters for each predictor function, 0.05 for the first predictor and 0.02 for the second predictor. The bi-variate predictor non-parametric regression model obtained has GCV = 3.567 with R^2 = 90.54% .

Application of model for fasting blood sugar level data based on time of treatment and blood pressure of diabetic patients

In this article, we apply the model to data from diabetic patients. The factor we measured was fasting blood sugar levels based of treatment time and blood pressure. The blood sugar level of a diabetic patient is unpredictable, it can go up or down at a very fast time. Therefore, we model the pattern of data changes with bi-variate predictor non-parametric regression through the penalized spline estimator. Patients as subjects

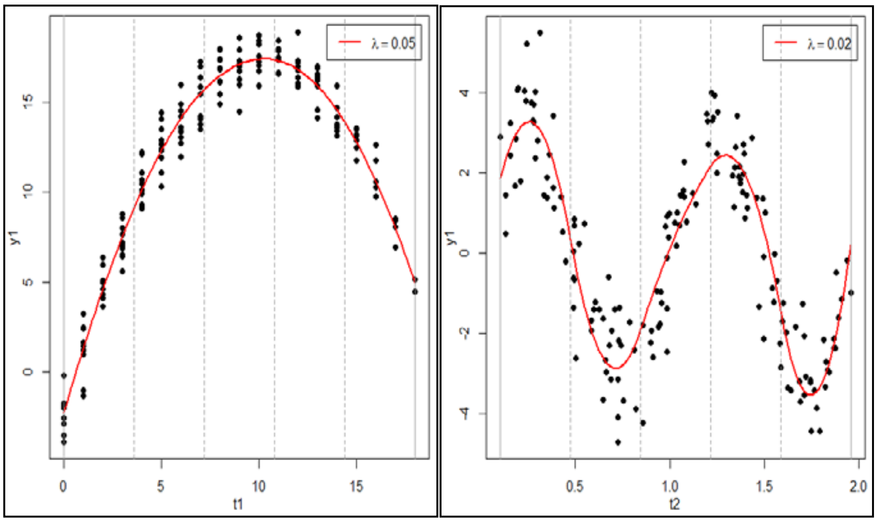

Figure 4. Estimation of bi-variate predictor non-parametric regression curves for the 1st and 2nd predictors.

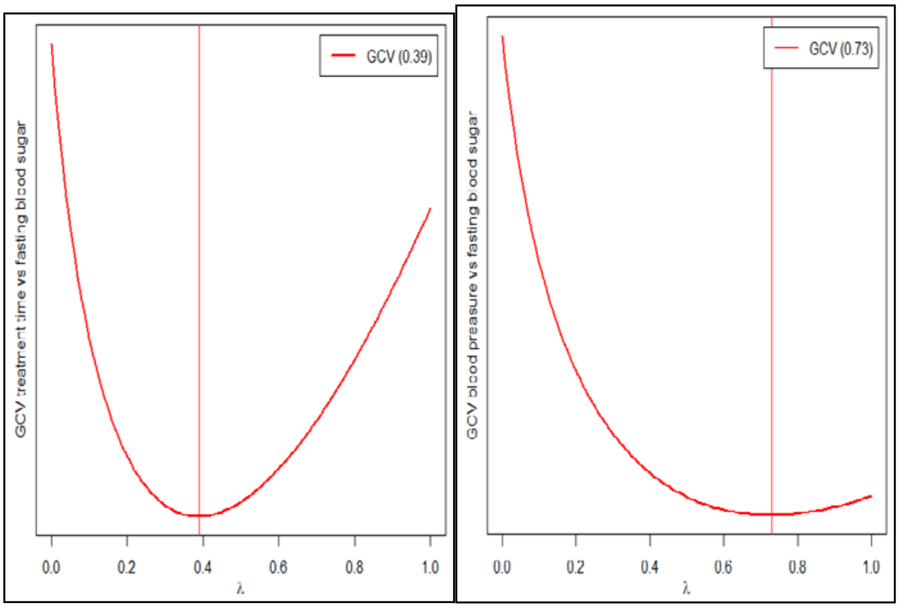

Figure 5. The plot of GCV with smoothing parameters for treatment time and blood pressure predictors for diabetes data.

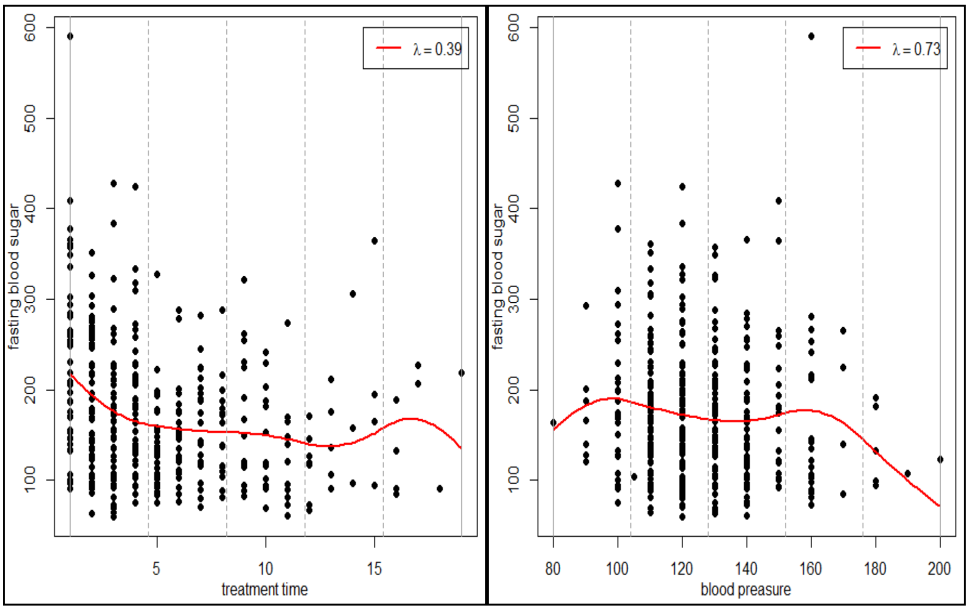

Figure 6. Estimation of bi-variate predictor non-parametric regression curves for treatment time and blood pressure predictors for diabetes data.

numbered 54 people with a total of 374 data. The data plot of treatment time and blood pressure are in Figure 5. Furthermore, the results of data analysis with quadratic penalized spline found the optimal smoothing parameters for treatment time were 0.39 and blood pressure 0.73 as in Figure 5. This provides a minimum GCV value of 5367,156 (5374,169 for one smoothing parameter). The penalized spline bi-variate predictor regression model for blood sugar levels that corresponds to Figure 6 is as follows:

$$\begin{aligned}\hat{y}_{ij} &= 226.86 - 79.04t_{ij1} - 103.34t_{ij1}^2 - 99.20\left(t_{ij1} - 4.6\right)_+^2 - 119.34\left(t_{ij1} - 8.2\right)_+^2 - \\ &+ 69.81\left(t_{ij1} - 11.8\right)_+^2 - 36.49\left(t_{ij1} - 15.4\right)_+^2 + 55.26t_{ij2} + 28.45t_{ij2}^2 + 11.86\left(t_{ij2} - 104\right)_+^2 \\ &+ 35.39\left(t_{ij2} - 128\right)_+^2 - 22.14\left(t_{ij2} - 152\right)_+^2 - 17.93\left(t_{ij2} - 176\right)_+^2\end{aligned}$$

The results of the estimation of fasting blood sugar regression curves based on time of treatment in Figure 6 show that there are five quadratic changes in the pattern of fasting blood sugar in diabetic patients. Of the five patterns, the fourth pattern must receive attention because it shows an increase in blood sugar, which is about 12-16 days of treatment (from the knot point 11.8-15.4). Furthermore, in Figure 6 for blood pressure shows five patterns of change and the third pattern that must receive attention. We see an increase in blood sugar at blood pressure intervals around 128-152 mm Hg.

For a comparison of this real data, the difference in GCV values with the use of one and two smoothing parameters, we analyzed 10 patients with an overall data amount of 73. Optimal smoothing parameters were 0.27 and 0.66 with GCV values of 8102,887 (8120,424 for one smoothing parameter). Next, we analyzed 20 patients with a total of 135 data taken at random. The GCV value of the penalized spline bi-variate predictor regression model for the use of two smoothing parameters is 6156,039 (6178,879 for one smoothing parameter). These results are consistent with the analysis of simulation data that the number of smoothing parameters in the penalized spline estimator for the bi-variate predictor longitudinal case is optimal uses two parameters. Furthermore, for diabetes data problems, we still need to study the case in larger data dimensions and consider other factors that are considered to affect blood sugar.

Conclusion

The penalized spline estimator in the bi-variate predictor non-parametric regression model involves a smoothing parameter in the estimation. The choice of smoothing parameters is very important in the non-parametric regression model through the penalized spline estimator because it is related to the smoothness of the regression curve. In this article, we have simulated the use of two and one smoothing parameter.

The selection of two smoothing parameters is based on different predictor functions from each other. The results of this article have shown the small GCV values found on two smoothing parameters. Therefore, we suggest the use of two smoothing parameters in the bi-variate model in the penalized spline estimator to obtain more accurate estimation results.

Another interesting result of this study is the GCV value for small subjects, where the GCV value is very different from the two and one smoothing parameter. Next, we can see the results of the box plots on the number of subjects 50, 75 and 100, where the GCV values of the two smoothing parameters remain smaller than one parameter. However, the difference in GCV values is not too large compared to the use of small subjects. Therefore, we can also advise to develop the study by adding larger subjects and variables. Our hope is that computationally expensive in a penalized spline can be minimized through the study of parameter smoothing. For the application of fasting blood sugar data from diabetic patients shows two optimal smoothing parameters on the time treatment factor and blood pressure. For diabetes research, we still need further research to get the best recommendations about the care of diabetic patients with an emphasis on patterns of blood sugar levels. The many factors involved medically cause the data to be extended to a larger dimension.

References

1. Ruppert D., and Carrol R. J. Spatially-adaptive penalties for spline fitting. *Aust. N. Z. J. Stat.*., **42**(2) : 205-223 (2000).
2. Claeskens G., Kribovokova T., and Opsomer J. D. Asymptotic properties of penalized spline estimators. *Biometrika*, **96**(3) : 529-544 (2009).
3. Montoya L. E., Ulloa N., and Miller V. A simulation study comparing knot selection methods with equally spaced knot in a penalized regression spline. *Int. J. Stat. Probab.*, **3**(3): 96-110 (2014).
4. Lee T. C. M., and Oh H. S. Robust penalized regression spline fitting with application to additively mixed modeling. *Comput. Stat.*, **22**: 159-171 (2007).
5. Wang B., Shi W., and Miao Z. Comparative analysis for robust penalized spline smoothing methods. *Math. Probl. Eng.*, **2014**(Article ID 642475) : 339-353 (2014).
6. Islamiyati A., Fatmawati, and Chamidah N. Estimation of covariance matrix on bi-response longitudinal data analysis with penalized spline regression. *J. Phys.: Conf. Ser.*, **979:** 012093 (2018).
7. Islamiyati A., Fatmawati, and Chamidah, N. Changes in blood glucose 2 hours after meals in type 2 diabetes patients based on length of treatment at Hasanuddin University Hospital, Indonesia. *Rawal Medical J.*, **45**(1): 31-34 (2020).
8. Zaherzadeh A., Rasekh A., and Babadi B. Diagnostic measures in ridge regression model with AR(1) errors under the stochastic linear restrictions. *J. Sci. I. R. Iran*, **29**: 67-78 (2018).
9. Mohammadi H., and Rasekh A. R. Liu Estimates and Influence Analysis in Regression Models with Stochastic Linear Restrictions and AR (1) Errors. *J. Sci. I. R. Iran*, **30**(3): 271-285 (2019).
10. Heckman N. E., and Ramsay J. O. Penalized regression with model-based penalties. *Can. J. Stat.*, **28**: 241-258 (2000).
11. Yao, and Lee. Penalized spline models for functional principal component analysis. *J. R. Stat. Soc. B.*, **68**(1) : 3-25 (2006).
12. Aydin D., and Yilmaz E. Modified spline regression based on randomly right-censored data: A comparative study. *Commun. Stat-Simul. C.*, **47**(9): 2587-2611 (2018).
13. Chamidah N., and Lestari B. Spline estimator in homoscedastic multi-response nonparametric regression model in case of unbalanced number of observations. *Far East J. Math. Sci.*, **100**(9) : 1433-1453 (2016).
14. Chamidah N., and Rifada M. Local linear estimation in bi-response semiparametric regression model for estimating median growth charts of children. *Far East J. Math. Sci.*, **99**(8): 1233-1244 (2016).
15. Lestari B., Fatmawati, and Budiantara I. N. Estimation of regression function in multi-response nonparametric regression model using smoothing spline and kernel estimators. *J. Phys.: Conf. Ser.*, **1097**: 012091 (2018).
16. Chamidah N., Gusti K. H., Tjahjono E., and Lestari B. Improving of classification accuracy of cyst and tumor using local polynomial estimator. *Telkomnika*, **17**: 1492-1500 (2019).
17. Lin X., and Zhang D. Inference in generalized additive mixed models by using smoothing splines. *J. R. Stat. Soc. B.*, **61**: 381-400 (1999).
18. Ni X., Zhang D., and Zhang H. H. Variable selection for semiparametric mixed models in longitudinal studies. *Biometrics*, **66** : 79-88 (2010).
19. Lai M. J., and Wang L. Bivariate penalized spline for regression. *Stat. Sin.*, **23**: 1399-1417 (2013).