Contents lists available at ScienceDirect

Journal of King Saud University – Science

journal homepage: www.sciencedirect.com

Original article

Biresponse nonparametric regression model in principal component analysis with truncated spline estimator

Anna Islamiyati [a,*], Anisa Kalondeng [a], Nurtiti Sunusi [a], Muhammad Zakir [b], Amir Kamal Amir [b]

[a] Department of Statistics, Faculty of Mathematics and Natural Sciences, Hasanuddin University, Makassar 90245, Indonesia
[b] Department of Mathematics, Faculty of Mathematics and Natural Sciences, Hasanuddin University, Makassar 90245, Indonesia

ARTICLE INFO

Article history:
Received 29 November 2020
Revised 1 February 2022
Accepted 3 February 2022
Available online 9 February 2022

Keywords:
Biresponse
Diabetes
Principal component
Spline truncated

ABSTRACT

Objectives: This study aims to model data that contain two correlated responses, multicollinearity in predictors, and has a pattern that does not follow a parametric form.
Methods: We propose the use of principal component analysis of truncated splines in a biresponse model. The use of principal components to overcome correlations between predictors, and biresponse to overcome correlations between responses by involving weighted estimates from the covariance matrix. In the PCA spline contains the optimal knot points which control the accuracy of the regression curve. The knot point chosen is the point which has the smallest GCV value among all knot points. In addition, we also consider the value of MSE in showing the model's ability.
Results: We demonstrated the ability of this method through simulation studies and obtained smaller GCV and MSE values compared to parametric regression and PCA. Furthermore, the data for type 2 diabetes mellitus, obtained two main components with different patterns of change. Based on the analysis, it was found that LDL cholesterol, total cholesterol, and triglycerides had a greater effect on changes in the pattern of fasting blood sugar and HbA1C.
Conclusions: The small errors of the simulation data indicate the accurate capabilities of the biresponse spline PCA model. The diabetes data analysis, it shows that patients need to pay attention to their cholesterol and triglyceride levels within normal limits.

1. Introduction

At this time, we have entered the era of big data on the number of samples, responses, and predictor variables. What concerns us here is that the larger the data, the greater the likelihood of assumptions for error correlation and multicollinearity in the predictors. One popular statistical approach to addressing this problem is principal component analysis (PCA). Several researchers who have studied PCA include Jolliffe and Cadima (2016) have developed PCA, which can reduce predictor variables through eigenvalues so that the components are mutually independent. The ability of PCA has been demonstrated by Bouwmans and Zahzah (2014) in image data analysis. Ghasemi et al. (2013) have classified the mineral composition of water samples, Vichi and Saporta (2009) have classified economic problems, and Hannachi et al. (2006) on climate issues. All of these PCA studies used a parametric approach that was limited to constructing the major components for a single response.

Another problem that can occur is that there are multicollinearity data that have an irregular pattern or do not follow a parametric pattern so that it is difficult to model it with the PCA parametric regression approach. Therefore, researchers developed nonparametric regression research, including Durand (1993) who has worked on instrumental variables with spline transformations. Wang et al. (2016) used PCA local polynomials and Shiokawa et al. (2018) with the PCA kernel. The use of another estimator by Lavado and Calapez (2011) have developed PCA with M Spline. For the spline estimator, there is a spline that contains a penalty function in its estimation criteria that can be used to overcome multicollinearity, namely spline smoothing by Lestari et al. (2010) and spline penalized by Islamiyati et al. (2020a). However, there is also another spline estimator that does not contain a penalty function, namely the truncated spline which cannot overcome

[*] Corresponding author.
E-mail address: annaislamiyati701@gmail.com (A. Islamiyati).

Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.jksus.2022.101892
1018-3647/

the multicollinearity of the predictor. Therefore, in this article, we are developing a study on spline truncated PCA for two responses.

On a larger response dimension in nonparametric regression studies, Soo and Bates (1996) have developed a multi-response spline estimator using the Generalized Gauss-Newton algorithm. Wang, et al. (2000) have analyzed the bivariate data with the smoothing spline estimator. Furthermore, Chamidah et al. (2012) examined the use of local polynomial estimators in nonparametric regression. Zahra and Mhlawy (2013) made a numerical study on an exponential spline. Khan and Shahna (2019) used a quadratic spline. Tohari and Chamidah (2020) used a negative bi-response binomial regression with a linear local estimator. Furthermore, Islamiyati et al. (2018) developed a penalized spline estimator in the longitudinal biresponse case. However, all these studies have not considered the multicollinearity cases that can occur in large predictor data dimensions. They only consider the correlations that occur in responses that are overcome by weight in the estimation criteria, such as using weight in the variance–covariance matrix.

We demonstrated the capabilities of the method through simulation data and compared it with the parametric regression model approach, PCA, and the nonparametric spline regression model. Next, we applied it to real data, namely data on type 2 diabetes mellitus that we obtained from the Hasanuddin University Teaching Hospital. Islamiyati et al. (2020b) has examined the effect of treatment time on blood sugar through a longitudinal penalized spline. Islamiyati et al. (2020c) examined the pattern of changes in blood sugar based on the diet of diabetic patients through a biresponse approach Islamiyati et al. (2020c); Zahra and Mhlawy (2013). Furthermore, Islamiyati (2022) obtained several segments of changes in blood sugar based on lifestyle factors of diabetic patients. All of them indicate that the blood sugar fits the spline approach because there are changes at certain intervals.

2. Spline truncated function in the PCA

Given the pairs of observation data $(t_{i1}, t_{i2}, \ldots t_{ip}, y_{1.i}, y_{2.i})$, the predictor variable t as many as p and the response variable y as many as two which follow the nonparametric pattern in $i = 1, 2, \ldots, n$. If it is assumed that the predictor variables are strongly correlated, then multicollinearity occurs and must first be resolved. In a statistical approach, one method of handling multicollinearity is principal component analysis (PCA) which has been widely used in many applications. Jolliffe and Cadima (2016) explain that PCA reduces a group of predictor variables into a group of new variables as much as predictors called principal component. It is a linear combination of predictor variables in which the number of principal components formed is as many as predictors. The assumption is that the components are orthogonal so that they are not correlated and it is believed that the information provided does not overlap.

It is known that $\mathbf{\Sigma}$ is the variance matrix of the predictor variable $t_1, t_2, \ldots, t_p$ which is used as the basis for selecting the number of main components. If c is the main component, then the equation for each component can be stated as follows:

$$\left.\begin{aligned} c_{1i} &= \gamma_{11} t_{1i} + \gamma_{12} t_{2i} + \ldots + \gamma_{1p} t_{pi} \\ c_{2i} &= \gamma_{21} t_{1i} + \gamma_{22} t_{2i} + \ldots + \gamma_{2p} t_{pi} \\ &\vdots \\ c_{pi} &= \gamma_{p1} t_{1i} + \gamma_{p2} t_{2i} + \ldots + \gamma_{pp} t_{pi} \end{aligned}\right\} \tag{1}$$

Eq. (1) can also be expressed in vector form, namely:

$$\underset{\sim}{c}_1 = \boldsymbol{T}\underset{\sim}{\gamma}_1,\ \underset{\sim}{c}_2 = \boldsymbol{T}\underset{\sim}{\gamma}_2,\ \ldots,\ \underset{\sim}{c}_p = \boldsymbol{T}\underset{\sim}{\gamma}_p$$

where $c_1, c_2, \ldots, c_p$ is called the principal component 1, 2,…, p and each has a variance of $\lambda_1, \lambda_2, \ldots, \lambda_p$, $\mathbf{T}$ is the predictor matrix and $\underset{\sim}{\gamma}$ is the principal component coefficient vector. The order of the main components is taken based on the large variety so that the largest variance is in the 1st component and the smallest variance is in the p-component with $\underset{\sim}{\gamma}_1 = \left(\gamma_{1.1}, \gamma_{1.2}, \ldots, \gamma_{1.p}\right)^T$ and $\underset{\sim}{\gamma}_1^T \underset{\sim}{\gamma}_1 = 1$. Suppose $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_p$ is the characteristic root corresponding to the feature vector $\underset{\sim}{\gamma}_1, \underset{\sim}{\gamma}_2, \ldots, \underset{\sim}{\gamma}_p$ of the matrix $\mathbf{\Sigma}$ and $\underset{\sim}{\gamma}_1^T \underset{\sim}{\gamma}_1 = 1$ for $j = 1, 2, \ldots, p$, then $\underset{\sim}{c}_1 = \boldsymbol{T}\underset{\sim}{\gamma}_1, \underset{\sim}{c}_2 = \boldsymbol{T}\underset{\sim}{\gamma}_2, \ldots, \underset{\sim}{c}_p = \boldsymbol{T}\underset{\sim}{\gamma}_p$ is the 1st, 2nd,…, p^{th} principal component of t. For data applications, the number of principal components is selected based on the cumulative variance described by the components.

In many multivariate studies, the principal component problem only comes to Eq. (1), which describes the principal components that are formed based on their total variety. However, the problem is different when our data is nonparametric. To model the data, the principal components obtained in Eq. (1) are then connected to the predictors through an estimator function in nonparametric regression, namely the truncated spline.

If the principal component selected is m from p component and is symbolized by c_j, $j = 1, 2, \ldots, m, \ldots, p$ for $m \leqslant p$, then the principal component function of the truncated spline based on the predictor can be stated as follows:

$$\left.\begin{aligned} c_1 &= f_1(t_1) + f_1(t_2) + \ldots + f_1(t_p) + \xi_1 \\ c_2 &= f_2(t_1) + f_2(t_2) + \ldots + f_2(t_p) + \xi_2 \\ &\vdots \\ c_m &= f_m(t_1) + f_m(t_2) + \ldots + f_m(t_p) + \xi_m \\ &\vdots \\ c_p &= f_p(t_1) + f_p(t_2) + \ldots + f_p(t_p) + \xi_p \end{aligned}\right\} \tag{2}$$

where $c_1, c_2, \ldots, c_m, \ldots, c_p$ is called the 1st, 2nd,…, m^{th},…, p^{th} principal component, $(f_j(t_1), f_j(t_2), \ldots, f_j(t_p))$ is the spline function in the predictors t_1, t_2, …t_{p} and $\xi_1, \xi_2, \ldots, \xi_m, \ldots, \xi_p$ is the error in the spline function truncated by the 1st, 2nd,…, m^{th},…, p^{th} principal component.

The function of each predictor $(f_j(t_1), f_j(t_2), \ldots, f_j(t_p))$ in (2) is a vector of the spline function of unknown shape for $j = 1, 2, \ldots, m$. It is estimated with a truncated spline in the order q and the point of knots K. The spline function in each predictor for each j^{th} component can be described as follows:

$$\left.\begin{aligned} f_j(t_1) &= \sum_{u_1=0}^{q_1} \beta_{j.u_1} t_1^{u_1} + \sum_{v_1=1}^{d_1} \beta_{j.(q_1+v_1)1} \left(t_1 - K_{j.v_1}\right)_+^{q_1} \\ f_j(t_2) &= \sum_{u_2=0}^{q_2} \beta_{j.u_2} t_2^{u_2} + \sum_{v_2=1}^{d_2} \beta_{j.(q_2+v_2)2} \left(t_2 - K_{j.v_2}\right)_+^{q_2} \\ &\vdots \\ f_j(t_p) &= \sum_{u_p=0}^{q_p} \beta_{j.u_p} t_p^{u_p} + \sum_{v_p=1}^{d_p} \beta_{j.(q_p+v_p)p} \left(t_p - K_{j.v_p}\right)_+^{q_p} \end{aligned}\right\} \tag{3}$$

where q is the degree of spline, β is the feature vector that corresponds to the root of the feature, K is the knot point, and v is the number of knot point. The truncated elements are shown as follows:

$$\left(t_j - K_{j.v_j}\right)_+^{q_j} = \begin{cases} \left(t_j - K_{j.v_j}\right) & ;\quad t_j > K_{j.v_j} \\ 0 & ;\quad t_j \leqslant K_{j.v_j} \end{cases}$$

Eq. (3) can be expressed in vector form, which is as follows:

$$f_j(t_1)+f_j(t_2)+\ldots+f_j(t_p)=\boldsymbol{X}_j\underset{\sim j}{\beta}$$

where $\mathbf{X}$ is the predictor matrix containing the knots point and $\underset{\sim j}{\beta}=\left(\underset{\sim 1}{\beta},\underset{\sim 2}{\beta},\ldots,\underset{\sim p}{\beta}\right)^T$ is the feature vector for each predictor.

Furthermore, the spline function of the first principal component can be stated as follows:

$$\underset{\sim 1}{c}=\boldsymbol{X}_1\underset{\sim 1}{\beta}+\underset{\sim 1}{\xi}$$

where $\underset{\sim 1}{\beta}=(\beta_{1.1},\beta_{1.2},\ldots,\beta_{1.p})^T$.

Furthermore, the spline function of the second main component, up to p, can be stated as follows:

$$\underset{\sim 2}{c}=\boldsymbol{X}_2\underset{\sim 2}{\beta}+\underset{\sim 2}{\xi},\ldots,\underset{\sim p}{c}=\boldsymbol{X}_p\underset{\sim p}{\beta}+\underset{\sim p}{\xi}$$

3. Biresponse nonparametric regression model with spline PCA

The biresponse nonparametric regression model on PCA is a nonparametric regression model that contains two response variables (y_r) with $r=1,2$ and several main component variables (c_j). Suppose that the number of main components selected is m, then the observation data pair $(c_{i1},c_{i2},\ldots c_{im},y_{1.i},y_{2.i})$, with $i=1,2,\ldots,n$, satisfies the biresponse nonparametric regression model as follows:

$$\underset{\sim i}{y}=\underset{\sim}{f}(c_{i1},c_{i2},\ldots,c_{im})+\underset{\sim i}{\varepsilon},\ i=1,2,\ldots,n \tag{4}$$

The model in (4) can be stated as:

$$\underset{\sim}{y}=\underset{\sim}{f}(c_1)+\underset{\sim}{f}(c_2)+\ldots+\underset{\sim}{f}(c_m)+\underset{\sim}{\varepsilon} \tag{5}$$

where $\underset{\sim}{y}$ is the response vector which contains the 1st response vector and the 2nd response, namely $\underset{\sim}{y}=\left(\underset{\sim 1}{y},\underset{\sim 2}{y}\right)^T$. Vector $\underset{\sim}{\varepsilon}$ is the random error vector, namely $\underset{\sim}{\varepsilon}=\left(\underset{\sim 1}{\varepsilon},\underset{\sim 2}{\varepsilon}\right)^T$ with $E\left(\underset{\sim}{\varepsilon}\right)=\underset{\sim}{0}$ and $\mathrm{Var}\left(\underset{\sim}{\varepsilon}\right)=\boldsymbol{V}$. The vector $\underset{\sim i}{\varepsilon}=(\varepsilon_{1.i},\varepsilon_{2.i})^T$ is assumed that:

$$E(\varepsilon_{1,i})=E(\varepsilon_{2,i})=0,E\left(\varepsilon_{1,i}^2\right)=\sigma_{1,i,}^2E\left(\varepsilon_{2,i}^2\right)=\sigma_{2,i}^2 \tag{6}$$

with $\sigma_{12.i}=\sigma_{21.i}$. The assumption in (6) shows that there is a correlation error between the 1st response with the 2nd response on the same i, but the error is not correlated for every i that is different in the response. Therefore, the model involves the weights obtained from the estimation of the covariance matrix, namely $\hat{\boldsymbol{\theta}}^{-1}$ as follows:

$$\hat{\boldsymbol{\theta}}=\begin{bmatrix}\hat{\boldsymbol{\Sigma}}_1 & \hat{\boldsymbol{\Sigma}}_{12}\\ \hat{\boldsymbol{\Sigma}}_{21} & \hat{\boldsymbol{\Sigma}}_2\end{bmatrix}$$

where $\hat{\boldsymbol{\Sigma}}_1$ is the estimate of the variance matrix in the 1st responses, $\hat{\boldsymbol{\Sigma}}_{12}=\hat{\boldsymbol{\Sigma}}_{21}$ is the estimate of the covariance matrix of the 1st and 2nd responses, and $\hat{\boldsymbol{\Sigma}}_2$ is the estimate of the variance matrix in the 2nd response.

Eq. (5) can also be written in matrix form, namely:

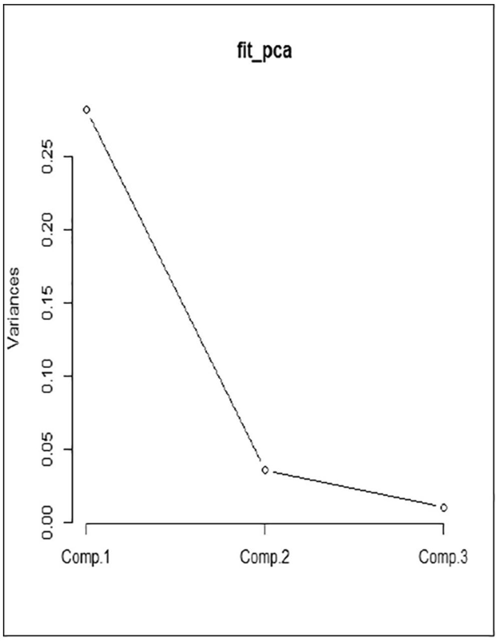


Fig. 2. Scree plot of simulation data.

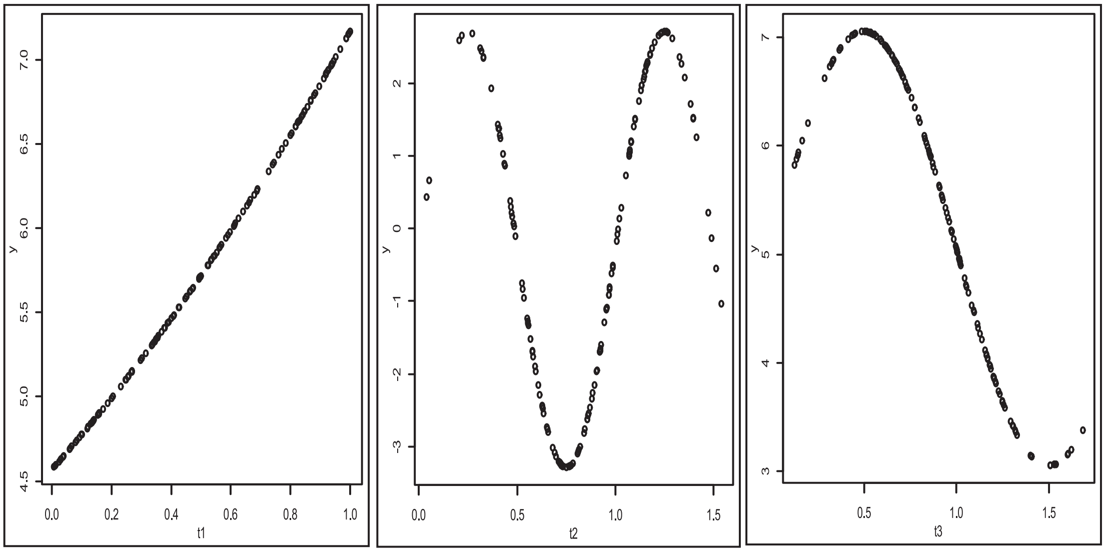


Fig. 1. Plot of data between predictors and responses.

$$\underset{\sim}{y} = \boldsymbol{X}\underset{\sim}{\alpha} + \underset{\sim}{\varepsilon} \qquad (7)$$

Furthermore, the Eq. (7) as a biresponse nonparametric regression model in PCA spline, it was estimated using weighted least square (WLS). The WLS estimator symbolized by P is as follows:

$$\mathrm{P} = \underset{\sim}{\varepsilon}^{T}\hat{\boldsymbol{\theta}}^{-1}\underset{\sim}{\varepsilon}$$

Further obtained:

$$\hat{\underset{\sim}{\alpha}} = \left(\boldsymbol{X}^{T}\hat{\boldsymbol{\theta}}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{T}\hat{\boldsymbol{\theta}}^{\}}-1\underset{\sim}{y} \qquad (8)$$

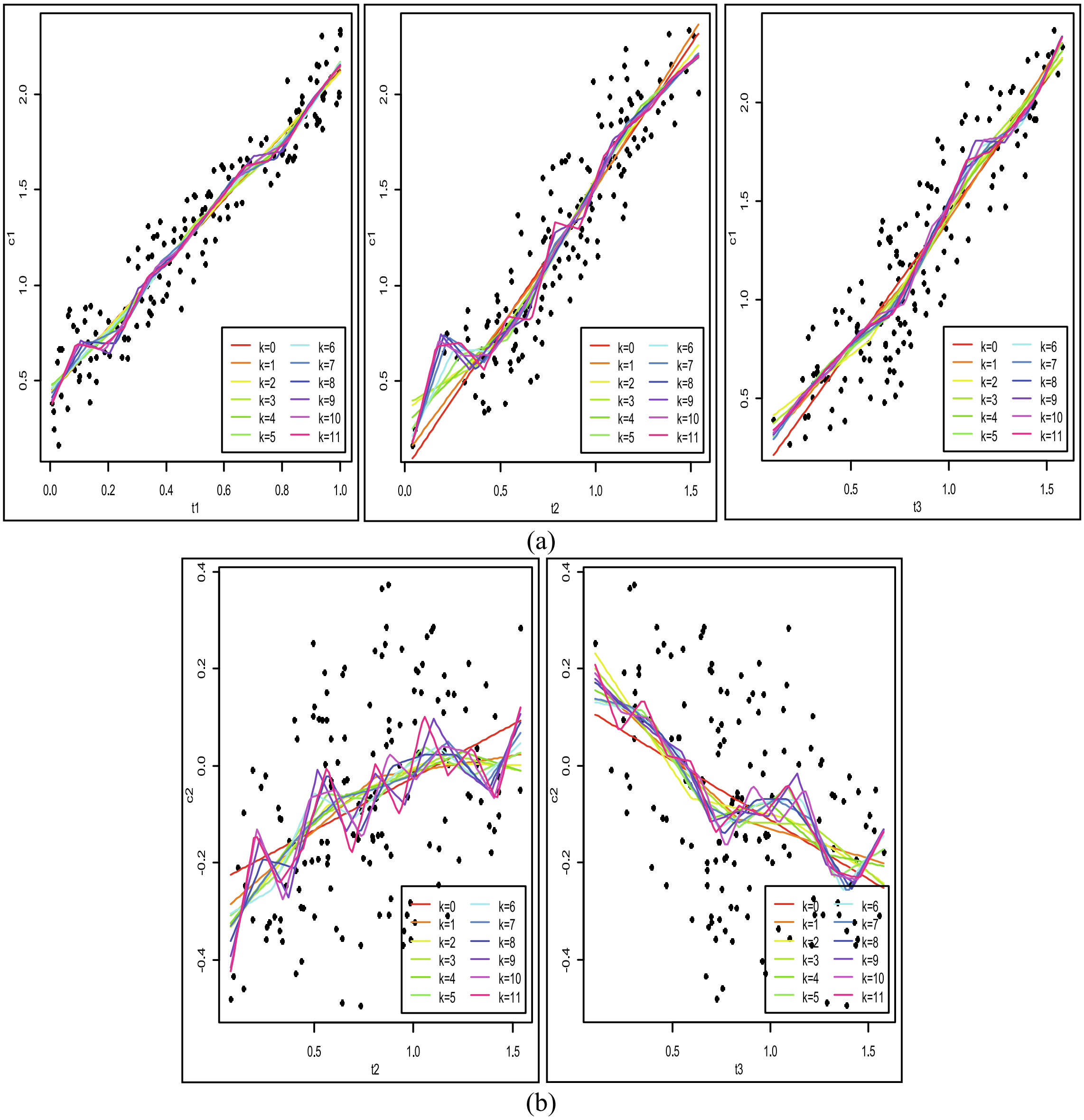
Based on the estimation results of the regression parameters in (8), we get an estimate of the biresponse nonparametric regression model on PCA through a truncated spline estimator as in Eq. (9).

$$\hat{\underset{\sim}{y}} = X\hat{\underset{\sim}{\alpha}} = X\left(X^{T}\hat{\theta}^{-1-1}X\right)^{-1}X^{T}\hat{\theta}^{-1}\underset{\sim}{y} \qquad (9)$$

4. Simulation data

We make different experimental functions on the predictors, namely $f(t_{i1})$ is in the form of polynomial while $f(t_{i2})$ and $f(t_{i3})$ is in the form of trigonometry. The number of subjects tested was n = 10, 30, 50, 100, 150 with correlation between predictors between 0.7 and 0.8. In this study, we choose a positive correlation because it is related to the condition variable to the real data. Simulations are being performed on a single response to demonstrate the ability of the PCA spline to model multicollinearity nonparametric data. The nonparametric regression model follows $y_i = f(t_{i1}, t_{i2}, t_{i3}) + \varepsilon_i$ with $i = 1, 2, \ldots, n$. The functions of the 1st predictor, 2nd predictor, and 3rd predictor are indicated by $f(t_{i1}) = 0.6t_i^2 + 2t_{i1} + 3$, $f(t_{i2}) = 3 \times \sin(2\pi t_{i2})$, and $f(t_{i3}) = 5 + 2\sin(\pi t_{i3})$.

In this section, we present a data plot for a sample size of n = 150 as shown in Fig. 1 for the 1st, 2nd, and 3rd predictors, respectively. The results of the correlation test between the predictors showed that there was multicollinearity in the data where



(a)

(b)

Fig. 3. The estimation results of the PCA spline regression curve at several knots for (a) the first component and (b) the second component.

there was a strong correlation between t_1 and t_2 of 0.86, t_1 and t_3 of 0.82, t_2 and t_3 of 0.71. In this article, the predictors are reduced to independent components via PCA with 3 principal components that correspond to the number of predictors. Based on the value of the cumulative proportion which can also be seen through the scree plot in Fig. 2, we take two principal components to be analyzed because the proportion of variance that can be explained has reached 97%. Furthermore, the predictor variables entered into each component are shown through the loading factor. The first component contains the three predictors, namely t_1, t_2, and t_3, while the second component contains only two predictors, namely t_2 and t_3. This indicates that the simulation data can be made into two independent components with each influencing predictor. There are two different conditions in the data, one is that there is a group of data that is influenced by all the predictors and there is another group that is only affected by two predictors. However, in the data, it is not only multicollinearity that occurs, but the data also has plots that do not follow a parametric pattern. The use of PCA alone has not been able to solve the problems that occur in the data. Therefore, in this study, we estimated the principal component based on the predictor through the truncated spline. Through the loading factor in PCA, it is shown the factors that significantly influence each main component. Significant predictors were then estimated from PC values through the nonparametric regression model of spline truncated PCA.

Fig. 3a shows the first component contains the significant predictors, t_1, t_2 and t_3 and shows an ascending linear pattern. The second component contains the t_2 and t_3 predictors shown in Fig. 3b. Furthermore, the two main components were modeled based on significant predictors through truncated spline PCA. We model it using knot points of 1 to 11 knots. Based on the truncated spline PCA, we obtain a spline regression curve with several optimal knot points. There is a different regression curve for each selected knot point, both for the first and second components. Therefore, we need to select the optimal knot point for each major component through the minimum GCV and MSE values as in Table 1 which corresponds to the knot points in Table 2. The minimum GCV and MSE values obtained at c_1 for t_1, t_2, and t_3 are 11, 8 and 10 knots, respectively. The minimum GCV and MSE values at c_2 for t_2 and t_3 is 11 knots. These results indicate that the minimum GCV and MSE values is obtained at different knot points for each component. Where the knot point is the starting point for a pattern change in the main component.

Furthermore, Fig. 4 shows a box plot of the MSE value which aims to compare the estimated results of the PCA spline with the multiple linear regression model and PCA. The use of the MSE value in the plot is because the model we used as a comparison with the estimated results of the PCA spline is a parametric model. The results in Fig. 4 shows that the PCA spline provides a much smaller MSE value compared to the parametric linear regression and PCA models. Therefore, the Spline PCA nonparametric regression model is very suitable to be used to model data between predictors with responses that do not follow a parametric pattern and correlated variables.

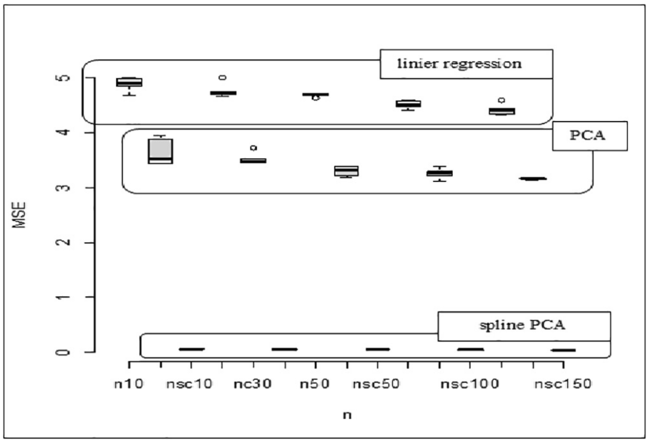


Fig. 4. Boxplot MSE of linear regression, PCA and PCA spline.

Table 1
GCV and MSE values at each knot point.

	GCV					MSE				
	c_1			c_2		c_1			c_2	
	t_1	t_2	t_3	t_2	t_3	t_1	t_2	t_3	t_2	t_3
1 knot	3.4219	5.0511	5.9059	4.1088	4.2035	0.0221	0.0331	0.0341	0.0283	0.0295
2 knots	3.0961	4.8394	5.7727	4.0977	4.1028	0.0216	0.0328	0.0340	0.0279	0.0282
3 knots	3.0955	4.8389	5.9038	4.0558	4.0952	0.0215	0.0327	0.0341	0.0277	0.0281
4 knots	3.0963	4.8305	5.5025	4.0181	4.0051	0.0216	0.0325	0.0339	0.0275	0.0274
5 knots	3.0947	4.8301	5.4450	3.9925	3.9762	0.0214	0.0325	0.0338	0.0269	0.0265
6 knots	3.0910	4.8389	5.5012	3.9807	3.9321	0.0211	0.0327	0.0339	0.0268	0.0258
7 knots	3.0946	4.8390	5.1106	3.9228	3.9588	0.0214	0.0328	0.0336	0.0261	0.0261
8 knots	3.0921	**4.8202**	5.1097	3.9414	3.9579	0.0213	**0.0321**	0.0335	0.0263	0.0261
9 knots	3.0926	4.8413	5.1022	3.9121	3.9554	0.0214	0.0329	0.0334	0.0258	0.0261
10 knots	3.0911	4.8388	**5.0461**	3.9304	3.9021	0.0212	0.0327	**0.0331**	0.0263	0.0258
11 knots	**3.0905**	4.8381	5.0837	**3.8012**	**3.8107**	**0.0203**	0.0327	0.0333	**0.0250**	**0.0253**

Bold numbers indicate the minimum GCV and MSE values.

Table 2
Optimal knot points.

		K_1	K_2	K_3	K_4	K_5	K_6	K_7	K_8	K_9	K_{10}	K_{11}
c_1	t_1	0.104	0.185	0.266	0.347	0.428	0.509	0.590	0.670	0.751	0.832	0.913
	t_2	0.241	0.403	0.565	0.728	0.890	1.052	1.214	1.377			
	t_3	0.241	0.375	0.508	0.642	0.775	0.909	1.042	1.176	1.309	1.443	
c_2	t_2	0.200	0.322	0.444	0.565	0.687	0.809	0.931	1.052	1.174	1.295	1.417
	t_3	0.230	0.352	0.475	0.597	0.719	0.842	0.965	1.087	1.209	1.332	1.454

5. Application on type 2 diabetes mellitus data

The ability of the PCA spline method to be more accurate in the simulation data in the previous section has provided assurance that this method can be applied to diabetes data. The variables studied were fasting blood sugar and HbA1C as the first and second responses, respectively. The factors of age, weight, height, HDL cholesterol, LDL cholesterol, total cholesterol, and triglycerides were the first, second, third, fourth, fifth, sixth, and seventh predictors, respectively. Data plots of fasting blood sugar levels are shown in Fig. 5 and HbA1C in Fig. 6. All figures show that the data plots between fasting blood sugar factors and HbA1C with LDL cholesterol, HDL cholesterol, total cholesterol, and triglyceride factors do not show a parametric plot. Therefore, we use a truncated spline as one of the estimators for non-parametric patterned data. This estimator is able to explain some pattern segmentation that occurs in the data through knot points. The patient's blood sugar is always changing in a fast time can be interpreted well by spline truncated through the knot point. Next, the correlation $r_{y_1.y_2} = 0.780$, $r_{t_2.t_3} = 0.856$ and $r_{t_3.t_4} = 0.586$. This shows a correlation between responses and multicollinearity in the predictor variables. To overcome these two types of correlation, we used a PCA biresponse model with a truncated spline estimator.

Based on the Scree plot, we can take two main components of the seven main components, because it can explain the variance of 85.7%. Furthermore, we found that the significant predictor variables in the first and second components were the same, namely the variables LDL cholesterol, total cholesterol and triglycerides. These results indicate that the two groups of diabetic patients can be modeled and we only need to consider three factors from the seven factors studied, namely LDL cholesterol, total cholesterol, and triglycerides. From the value of the principal component that corresponds to the predictor, we can model the main component through the spline function truncated with a certain knot point.

The estimation results of the PCA spline regression curve between the first and second components with predictors are shown in Fig. 7. Based on Fig. 7a and b, the spline curve estimation of each component looks different from one another. In the cholesterol factor, namely LDL and total cholesterol, there is an upward trend in the first and second components, but the increase is different from one another. For triglyceride factors, there is an uptrend in the first component and a downtrend in the second component. The trend is indicated by optimal knot points where the points are selected based on the GCV value. In this data, we get 3 knot points which give the minimum GCV value, namely for LDL cholesterol factors are 105.5, 173, and 240.5, for total cholesterol factors are 164, 252, 340, and for triglyceride factors are 133, 219, 305.

The spline equation is truncated on each component corresponding to the knot point are as follows:

$$\left.\begin{array}{c}
c_1 = 547.147 + 247.493t_5 + 323.661(t_5 - 105.5)_+ + 463.159(t_5 - 173)_+ + 532.79(t_5 - 240.5)_+ + \\
229.765t_6 + 344.662(t_6 - 164)_+ + 545.372(t_6 - 252)_+ + 598.609(t_6 - 340)_+ + 236.209t_7 + \\
351.961(t_7 - 133)_+ + 430(t_7 - 219)_+ + 478.244(t_7 - 305)_+ \\
c_2 = 59.437 + 43.86t_5 + 92.908(t_5 - 105.5)_+ + 131.477(t_5 - 173)_+ - 103.674(t_5 - 240.5)_+ + \\
39.642t_6 + 88.338(t_6 - 164)_+ - 77.957(t_6 - 252)_+ + 101.394(t_6 - 340)_+ - 48.583t_7 + \\
43.911(t_7 - 133)_+ - 2.963(t_7 - 219)_+ - 76.364(t_7 - 305)_+
\end{array}\right\} \tag{10}$$

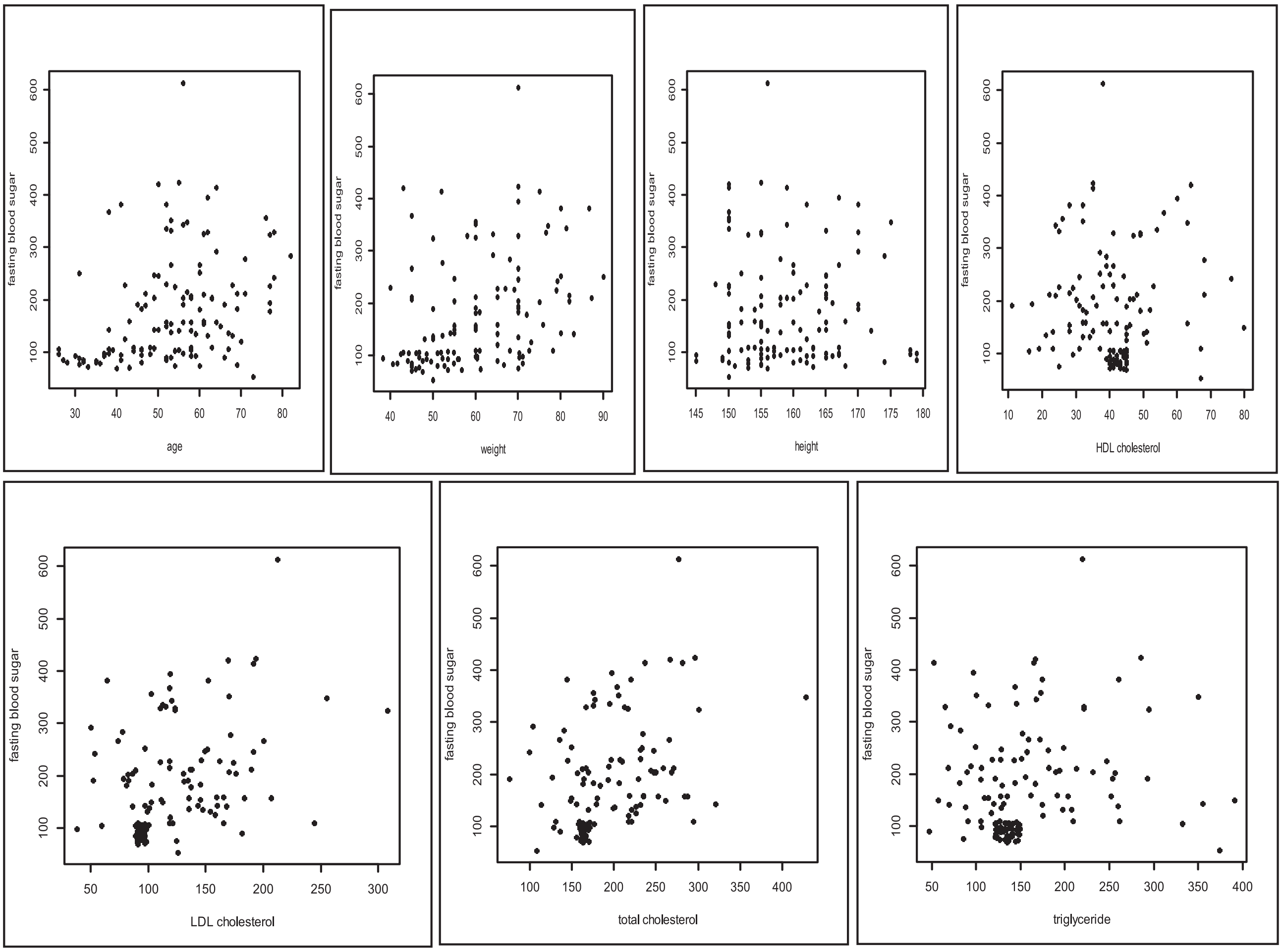


Fig. 5. The plot of fasting blood sugar (y_1) based on predictors.

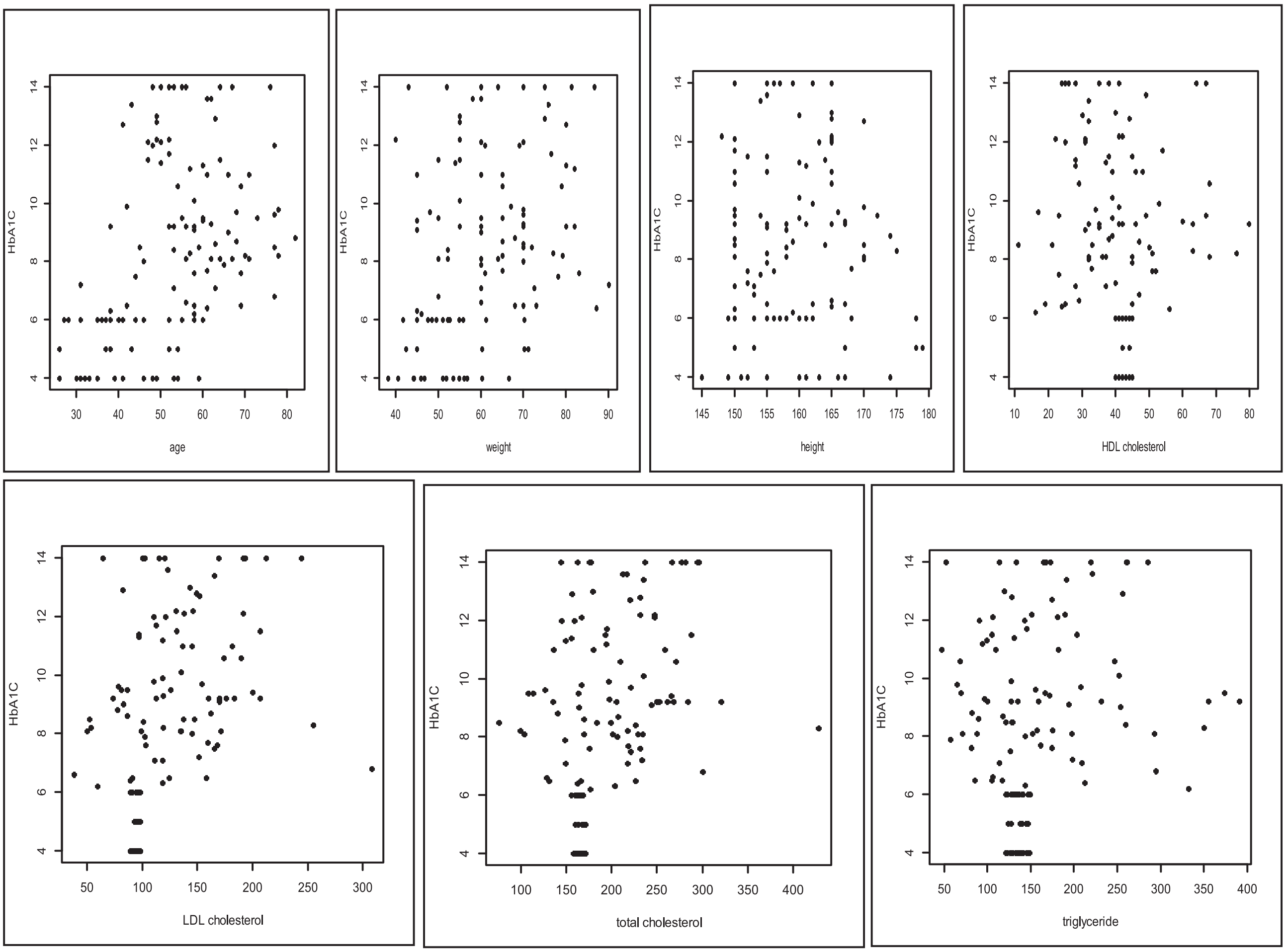


Fig. 6. The plot of HbA1C (y_2) data based on predictors.

Eq. (10) corresponds to Fig. 8 which shows the estimation results of the spline truncated curve for each principal component.

Furthermore, the biresponse PCA spline regression model obtained between the response and the main components of the diabetes data is as follows:

$$y_1 = -37.905 + 0.184c_1 + 0.445c_2$$
$$y_2 = 1.885 + 0.005c_1 + 0.016c_2$$

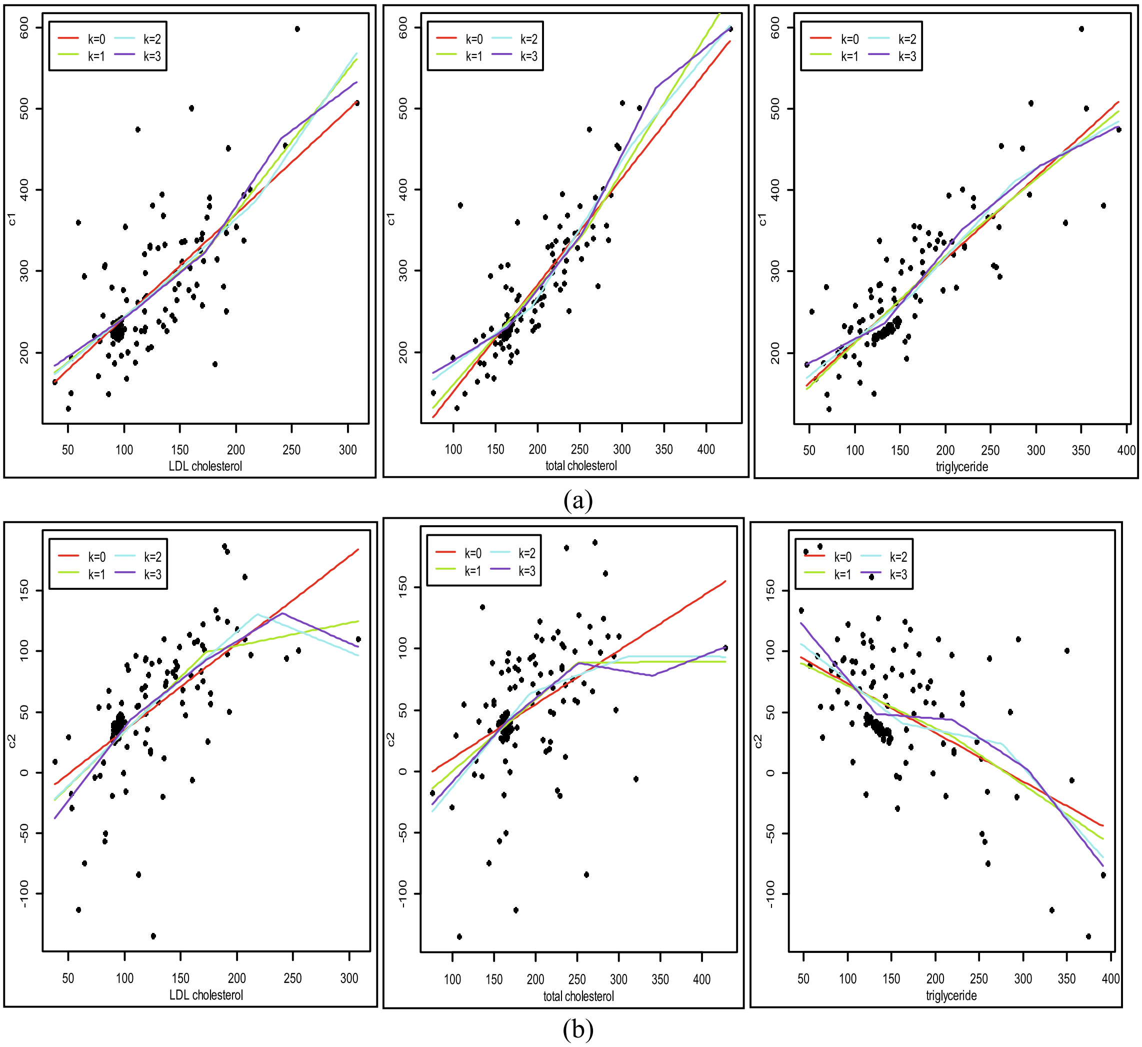
Based on the equation of the principal components in (10), the PCA biresponse spline regression model can be expressed as follows:

$$\begin{aligned}\hat{y}_1 = {} & 89.219 + \{45.538t_5 + 59.553(t_5 - 105.5)_+ + 85.221(t_5 - 173)_+ + 90.033(t_5 - 240.5)_+ + \\ & 42.276t_6 + 63.417(t_6 - 164)_+ + 100.348(t_6 - 252)_+ + 110.144(t_6 - 340)_+ + 43.462t_7 + \\ & 64.761(t_7 - 133)_+ + 79.12(t_7 - 219)_+ + 87.996(t_7 - 305)_+\} + \{19.517t_5 + 41.344(t_5 - 105.5)_+ + \\ & 58.507(t_5 - 173)_+ - 46.134(t_5 - 240.5)_+ + 17.641t_6 + 39.310(t_6 - 164)_+ - 34.690(t_6 - 252)_+ + \\ & 45.123(t_6 - 340)_+ - 21.619t_7 + 19.540(t_7 - 133)_+ - 1.318(t_7 - 219)_+ - 33.982(t_7 - 305)_+\} \\ \hat{y}_2 = {} & 5.571 + \{1.237t_5 + 1.618(t_5 - 105.5)_+ + 2.315(t_5 - 173)_+ + 2.663(t_5 - 240.5)_+ + \\ & 1.148t_6 + 1.723(t_6 - 164)_+ + 2.726(t_6 - 252)_+ + 2.993(t_6 - 340)_+ + 1.181t_7 + \\ & 1.759(t_7 - 133)_+ + 2.15(t_7 - 219)_+ + 2.391(t_7 - 305)_+\} + \{0.701t_5 + 1.486(t_5 - 105.5)_+ + \\ & 2.103(t_5 - 173)_+ - 1.658(t_5 - 240.5)_+ + 0.634t_6 + 1.413(t_6 - 164)_+ - 1.247(t_6 - 252)_+ + \\ & 1.622(t_6 - 340)_+ - 0.777t_7 + 0.702(t_7 - 133)_+ - 0.047(t_7 - 219)_+ - 1.221(t_7 - 305)_+\}\end{aligned} \tag{11}$$

The results of the analysis of the biresponse PCA spline model showed a pattern of changes in fasting blood sugar and HbA1C levels, which were mostly influenced by LDL cholesterol, total cholesterol, and triglycerides. In the first component, fasting blood sugar and HbA1C tend to rise along with the increase in cholesterol and triglycerides. However, the increment varies at certain value intervals. Furthermore, for the second component, fasting blood sugar and HbA1C increased and decreased based on the patient's cholesterol and triglyceride levels in certain intervals. This shows that through spline truncated PCA biresponse, we can identify two conditions that can occur in patients with type 2 diabetes mellitus.

6. Conclusion

A bi-response truncated PCA spline model was developed for data containing multi-dimensional variables in which responses are correlated as well as predictors. The multicollinearity problem in predictors was solved by using PCA spline. The principal component that is formed is modeled with a predictor through a trun-



(a)

(b)

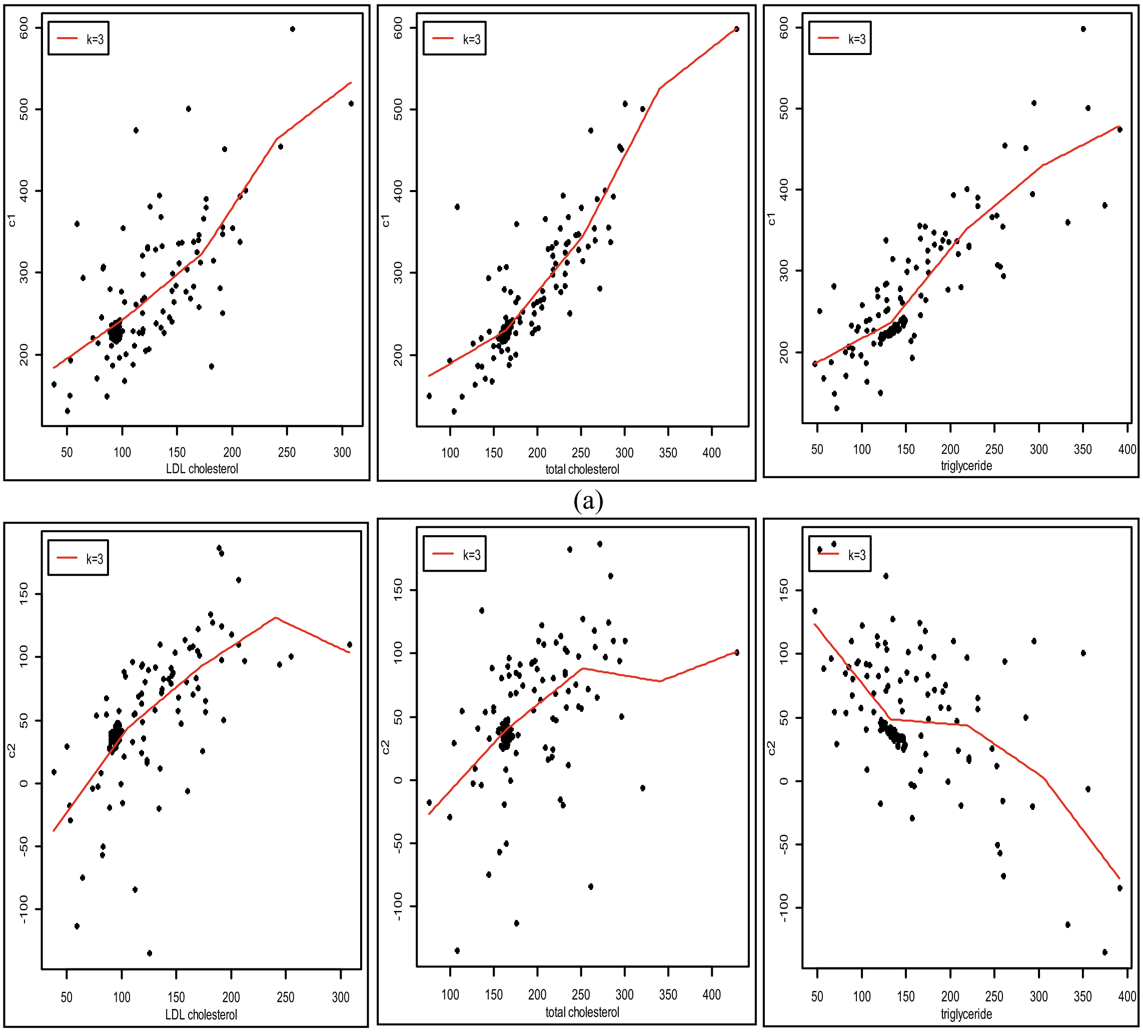
Fig. 7. The estimation results of the truncated spline curve are based on the factors of LDL cholesterol, total cholesterol, and triglycerides on (a) the first component and (b) the second component.

cated spline estimator which considers the knot point. The ability of the method has been demonstrated through simulation data and MSE values were obtained that were smaller than the parametric regression and PCA approaches as shown in Fig. 4. This method is also applied to data on type 2 diabetes mellitus patients. Based on the results of the analysis of the biresponse spline PCA model, it was found that there were two main components which indicated that there were two different groups of type 2 diabetes mellitus patients. The two principal components are equally affected by LDL cholesterol, total cholesterol and triglycerides. What distinguishes these components is the pattern of changes in fasting blood sugar and HbA1C based on these three factors. The pattern can be seen in Fig. 8, and then modeled as in Eq. (10). The condition of the type 2 diabetes mellitus patients described in this article shows that the important factors that the patient should pay attention to are the regulation of LDL cholesterol, total cholesterol, and triglycerides. The shape of their influence on the patient is described in terms of two components. Also, the effect of these three factors shows that there are several patterns of change at certain intervals corresponding to the knot point. This result is one of the advantages of this method that cannot be explained through a parametric approach.

This research is sponsored by Deputy of Research and Development Strengthening, Ministry of Research and Technology/National Agency for Research and Innovation, Republic of Indonesia for the Basic Research, and will continue to be developed on both theory and data applications. There is an obligation for us to publish our research results as a form of review of the development of pre-existing methods. For that matter, no potential conflict will occur associated with this article, neither to funding nor to all authors.

Acknowledgement

Many thanks to the Deputy of Research and Development Strengthening, Ministry of Research and Technology/National Agency for Research and Innovation, The Republic of Indonesia for the Basic Research with the research contract No: 7/AMD/E1/KP.PTNBH/2020 dated 11 May 2020.



(a)

(b)

Fig. 8. The estimation results of the spline truncated PCA curve for biresponse to (a) the fasting blood sugar factor and (b) the HbA1C factor.

References

Bouwmans, T., Zahzah, E., 2014. Robust PCA via principal component pursuit: a review fora comparative evaluation in video surveillance. Comput. Vis. Image Underst. 122, 22–34.

Chamidah, N., Budiantara, I.N., Sunaryo, S., Ismaini, Z., 2012. Designing of child growth chart based on multi response local polynomial modeling. J. Math. Stat. 8 (3), 342–347.

Durand, J.F., 1993. Generalized principal component analysis with respect to instrumental variables via univariate spline transformation. Comput. Stat. Data An. 16 (4), 423–440.

Ghasemi, J.B., Zolfonoun, E., Khosrokhavar, R., 2013. Linear and nonlinear multivariate classification of Iranian bottled mineral waters according to their elemental content determined by ICP-OES. J. Sci. Islam. Repub. Iran 24 (1), 15–22.

Hannachi, A., Jolliffe, I.T., Stephenson, D.B., Trendafilov, N., 2006. In search of simple structures in climate: simplifying EOFs. Int. J. Climatol. 26 (1), 7–28.

Islamiyati, A., Fatmawati, Chamidah, N., 2018. Estimation of covariance matrix on bi-response longitudinal data analysis with penalized spline regression. J. Phys.: Conf. Ser. 979 (012093), 1–8.

Islamiyati, A., Sunusi, N., Kalondeng, A., Fatmawati, F., Chamidah, N., 2020a. Use of two smoothing parameters in penalized spline estimator for bi-variate predictor non-parametric regression model. J. Sci. Islam. Repub. Iran. 31 (2), 175–183.

Islamiyati, A., Fatmawati, Chamidah, N., 2020b. 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.

Islamiyati, A., Fatmawati, Chamidah, N., 2020c. Penalized spline estimator with multi smoothing parameters in biresponse multipredictor regression model for longitudinal data. Songklanakarin J. Sci. Technol. 42 (4), 897–909.

Islamiyati, A. 2022. Spline longitudinal multi-response model for the detection of lifestyle-based changes in blood glucose of diabetic patients. Curr. Diabetes Rev. E-pub Ahead of Print, Published on: 14 January, 2022.

Jolliffe, I.T., Cadima, J., 2016. Principal component analysis: a review and recent developments. Phil. Trans. R. Soc. A. 374 (2065), 20150202. https://doi.org/10.1098/rsta.2015.0202.

Khan, A., Shahna, 2019. Non-polynomial quadratic spline method for solving fourth order singularly perturbed boundary value problems. J. King Saud Univ. Sci. 31 (4), 479–484.

Lavado, N., Calapez, T., 2011. Principal components analysis with spline optimal transformations for continuous data. IAENG Int. J. Appl. Math. 41 (4), 367–375.

Lestari, B., Budiantara, I.N., Sunaryo, S., Mashuri, M., 2010. Spline smoothing for multi-response nonparametric regression model in case of heteroscedasticity of variance. J. Math. Stat. 8 (3), 377–384.

Shiokawa, Y., Date, Y., Kikuchi, J., 2018. Application of kernel principal component analysis and computational machine learning to exploration of metabolites strongly associated with diet. Sci. Rep. 8.

Soo, Y.W., Bates, D.M., 1996. Multiresponse spline regression. Comput. Stat. Data An 22 (6), 619–631.

Tohari, A., Chamidah, N., 2020. Modelling of HIV and AIDS cases in Indonesia using bi-response negative binomial regression approach based on local linear estimator. Ann. Biol 36 (2), 215–219.

Vichi, M., Saporta, G., 2009. Clustering and disjoint principal component analysis. Comput. Stat. Data An. 53 (8), 3194–3208.

Wang, Y., Lu, W., Wang, B., Liu, L., 2016. A robust polynomial principal component analysis for seismic noise attenuation. J. Geophys. Eng. 13 (6), 1002–1009.

Wang, Y., Guo, W., Brown, M.B., 2000. Spline smoothing for bivariate data with application to association between hormones. Stat. Sin. 10, 377–397.

Zahra, W.K., Mhlawy, A.M.E., 2013. Numerical solution of two-parameter singularly perturbed boundary value problems via eksponential spline. J. King Saud Univ. Sci. 25 (3), 201–208.