Department of Social Systems and Management

Discussion Paper Series

No. 1217

An Algorithm for Nonlinear Weighted Least Squares Regression

by

Antoni Wibowo

September 2008

UNIVERSITY OF TSUKUBA
Tsukuba, Ibaraki 305-8573
JAPAN

An Algorithm for Nonlinear Weighted Least Squares Regression

Antoni Wibowo*
Graduate School of Systems and Information Engineering,
University of Tsukuba,1-1-1 Tennodai, Tsukuba 305-8573, Japan.

September 18, 2008

Abstract

We consider the regression model with variance of random errors having unequal values in diagonal elements. If variance of random errors has unequal values, the results of ordinary linear regression and kernel principal component regression become inappropriate to be used. The weighted least-squares (WLS) is widely used to handle the limitations. However, WLS on linear regression yields a linear prediction model and has no guarantee that multicollinearity does not exist in the WLS model. In this paper, we propose a method and an algorithm to overcome these difficulties. Then, we compare the proposed model with some other methods.

Keywords: Regression analysis, weighted least squares, kernel principal component analysis, kernel principal component regression, multicollinearity, Gaussian kernel.

1 Introduction

Regression analysis is one of the most widely used techniques for analyzing data. The *multiple linear regression model* with p regressors is given by

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j x_j + \epsilon. \tag{1.1}$$

The parameters β_j, $(j = 0, 1, \ldots, p)$, are called the *regression coefficients* and ϵ is a random variable called the *random error*. It is assumed that the values of $x_1, x_2, \ldots, x_p$ are chosen by an experimenter and β_j's are unknown.

The *ordinary multiple linear regression model* corressponding to Eq. (1.1) is written as

$$\begin{aligned} \mathbf{Y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \\ E(\boldsymbol{\epsilon}) &= \mathbf{0}, \\ Var(\boldsymbol{\epsilon}) &= \sigma^2 \mathbf{I}_N, \end{aligned} \tag{1.2}$$

where $\mathbf{Y} = (Y_1 \quad Y_2 \quad \ldots \quad Y_N)^T$, $\mathbf{x}_i = \begin{pmatrix} x_{i1} & x_{i2} & \ldots & x_{ip} \end{pmatrix}^T$, $\tilde{\mathbf{X}} = (\mathbf{x}_1 \quad \mathbf{x}_2 \quad \ldots \quad \mathbf{x}_N)^T$, $\mathbf{X} = \begin{pmatrix} \mathbf{1}_N & \tilde{\mathbf{X}} \end{pmatrix}$, $\boldsymbol{\beta} = (\beta_0 \quad \beta_1 \quad \ldots \quad \beta_p)^T$, $\boldsymbol{\epsilon} = (\epsilon_1 \quad \epsilon_2 \quad \ldots \quad \epsilon_N)^T$, I_N denotes the

*Email address: wibowo@sk.tsukuba.ac.jp.

$N \times N$ identity matrix, $x_{ij} \in \mathbb{R}$ for $i = 1, 2, \ldots, N; j = 1, 2, \ldots, p$; and $\sigma^2 \in \mathbb{R}$ where $\mathbb{R}$ is the set of real numbers. The sizes of $\mathbf{x}_i$, $\mathbf{Y}$, $\tilde{\mathbf{X}}$, $\mathbf{X}$, $\boldsymbol{\beta}$ and $\boldsymbol{\epsilon}$ are $p \times 1$, $N \times 1$, $N \times p$, $N \times (p+1)$, $(p+1) \times 1$ and $N \times 1$, respectively, and $\mathbf{1}_N = (1 \quad 1 \quad \ldots \quad 1)^T_{N \times 1}$. The vector $\mathbf{x}_i^T$ denotes the transpose of the vector $\mathbf{x}_i$. Matrix $\mathbf{X}$ is called the *regression matrix*.

We assume that the row vectors of $\mathbf{X}$ are linearly independent. The aim of regression analysis is to find the estimator of $\boldsymbol{\beta}$, say $\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\beta}_0 & \hat{\beta}_1 & \ldots & \hat{\beta}_p \end{pmatrix}^T$, such that the least-squares function $\|\boldsymbol{\epsilon}\|^2$ is minimized. The solution can be found by solving the following linear equation

$$\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{Y}. \tag{1.3}$$

Eq. (1.3) is called the *least squares normal equations* and $\hat{\boldsymbol{\beta}}$ is called the *ordinary least-squares (OLS) estimator* of $\boldsymbol{\beta}$. Since the row vectors of $\mathbf{X}$ are linearly independent, $\mathbf{X}^T\mathbf{X}$ is invertible [1]. Hence, eigenvalues of $\mathbf{X}^T\mathbf{X}$ are positive numbers. In addition, we say that *multicollinearity* exists on $\mathbf{X}$ if $\mathbf{X}^T\mathbf{X}$ is a nearly singular matrix, i.e., if some eigenvalues of $\mathbf{X}^T\mathbf{X}$ are close to zero. Since $\mathbf{X}^T\mathbf{X}$ is invertible, we obtain

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}, \tag{1.4}$$

where $(\mathbf{X}^T\mathbf{X})^{-1}$ is the inverse of $\mathbf{X}^T\mathbf{X}$. The variance of $\hat{\beta}_j$ for $j = 0, 1, \ldots, p$, denoted by $Var(\hat{\beta}_j)$, is given by

$$Var(\hat{\beta}_j) = \sigma^2((\mathbf{X}^T\mathbf{X})^{-1})_{j+1,j+1}, \qquad j = 0, 1, \ldots, p. \tag{1.5}$$

If multicollinearity exists on $\mathbf{X}$ then $Var(\hat{\beta}_j)$ can be a large number and under the assumption that ε_i is normally distributed, the tests for inferences β_j $(j = 0, 1, \ldots, p)$ have low power and the confidence interval can be large. Therefore, it will be difficult to decide if a variable x_j makes a significant contribution to the regression. These implications are known as the *effects of multicollinearity* [15].

After the observations are taken, we obtain the observed data corresponding to $\mathbf{Y}$. Let $\mathbf{y} = (y_1 \quad y_2 \quad \ldots \quad y_N)^T \in \mathbb{R}^N$ be the observed data corresponding to $\mathbf{Y}$ and $\hat{\boldsymbol{\beta}}^* = \begin{pmatrix} \hat{\beta}_0^* & \hat{\beta}_1^* & \ldots & \hat{\beta}_p^* \end{pmatrix}^T \in \mathbb{R}^{p+1}$ be the value of $\hat{\boldsymbol{\beta}}$ when $\mathbf{Y}$ is replaced by $\mathbf{y}$ in the Eq. (1.6). Then, we obtain

$$\hat{\boldsymbol{\beta}}^* = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}. \tag{1.6}$$

The *prediction value* of $\mathbf{y}$, say $\hat{\mathbf{y}}$, is given by

$$\hat{\mathbf{y}} = \begin{pmatrix} \hat{y}_1 & \hat{y}_2 & \ldots & \hat{y}_N \end{pmatrix}^T := \mathbf{X}\hat{\boldsymbol{\beta}}^*, \tag{1.7}$$

and the *residual* between $\mathbf{y}$ and $\hat{\mathbf{y}}$ is given by

$$\mathbf{e} = \begin{pmatrix} e_1 & e_2 & \ldots & e_N \end{pmatrix}^T := \mathbf{y} - \hat{\mathbf{y}}. \tag{1.8}$$

The *root mean square error* (RMSE) for the ordinary regression model is given by

$$RMSE_{olr} := \sqrt{\frac{\mathbf{e}^T\mathbf{e}}{N}} \tag{1.9}$$

and the *prediction by the ordinary linear regression* is given by

$$f(\boldsymbol{x}) := \hat{\beta}_0^* + \sum_{j=1}^{p} \hat{\beta}_j^* x_j, \tag{1.10}$$

where f is a function from $\mathbb{R}^p$ to $\mathbb{R}$. We notice that a plot of the residual e_i ($i = 1, 2, \ldots, N$) and its corresponding $\hat{y}_i$ is useful to check the assumption of constant variance [2, 3, 8, 14].

In some cases, we face the regression model with variance of random errors having unequal values in its diagonal elements. If variance of random errors has unequal values, Eq. (1.3) and hypothesis testing based on the ordinary OLS estimator of the variance matrix become invalid [2, 3, 8, 14, 15]. Let the variance of random errors be $Var(\boldsymbol{\epsilon}) = \sigma_1^2\mathbf{V}$ where $\mathbf{V}$ is an $N \times N$ diagonal and nonsingular matrix. Let $\mathbf{L}$ be an $N \times N$ matrix such that $\mathbf{L}\mathbf{L}^T = \mathbf{V}$. Then, the difficulties can be avoided by multiplying the model with $\mathbf{L}^{-1}$. This technique is known as the *weighted least-squares (WLS)* on linear regression. However, the WLS on linear regression yields a linear prediction model. Since the most of real problems are nonlinear, the model has limitations on applications. Beside that, there is no guarantee that multicollinearity does not exist in $\mathbf{L}^{-1}\mathbf{X}$. Although we can use the *kernel principal component regression (KPCR)* to handle the limitations of the linearity and multicollinearity, KPCR can be inappropriate in the regression model. Since KPCR was constructed by the different assumption, that is, the variance of random errors having equal values in its diagonal elements. The KPCR was studied by Rosipal *et al.* [9, 10, 11], Hoegaerts *et al.* [4], Jade *et al.* [5] and Wibowo *et al.* [16].

In this paper, we propose a method and an algorithm to overcome the above difficulties. The procedure to derive a nonlinear prediction model of the proposed method is straightforward as the procedure in WLS on linear regression, except that some mathematical techniques are done to obtain the nonlinear prediction and to avoid the effects of multicollinearity. We refer the proposed technique as *weighted least-squares KPCR (WLS KPCR)*.

This manuscript is organized as follows: Section 2, we review the WLS on linear regression model. In Section 3, the detailed WLS KPCR and its algorithm will be described. In Section 4, we compare the capabilities of the ordinary linear regression, the WLS linear regression, the KPCR and the WLS KPCR. Finally, conclusions are given in Section 5.

2 Weighted Least Squares

Let us consider the following model:

$$\begin{aligned} \mathbf{Y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \\ E(\boldsymbol{\epsilon}) &= \mathbf{0}, \\ Var(\boldsymbol{\epsilon}) &= \sigma_1^2\mathbf{V}, \end{aligned} \tag{2.1}$$

where $\mathbf{V} = diag(1/w_1, 1/w_2, \ldots, 1/w_N)$ and w_i is a positive number for $i = 1, 2, \ldots, N$. The weight w_i is estimated by using the data $\mathbf{y}$ and $\mathbf{X}$, see for example [2, 3, 8]. An implication of the assumption $Var(\boldsymbol{\epsilon}) = \sigma^2\mathbf{V}$ is the OLS estimator $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ be inappropriate. This limitation is avoided by transforming the

model 2.1 to a new model that satisfies the ordinary linear regression model assumption. Afterward, we use the same procedure of the ordinary linear regression model to the transformed model.

Let $\mathbf{L} = diag(1/\sqrt{w_1}, 1/\sqrt{w_2}, \ldots, 1/\sqrt{w_N})$. Hence, $\mathbf{L}^T = \mathbf{L}$, $\mathbf{L}\mathbf{L}^T = \mathbf{V}$ and $\mathbf{L}^{-1} = diag(\sqrt{w_1}, \sqrt{w_2}, \ldots, \sqrt{w_N})$. Then, we have

$$\mathbf{L}^{-1}\mathbf{Y} = \mathbf{L}^{-1}\mathbf{X}\boldsymbol{\beta} + \mathbf{L}^{-1}\boldsymbol{\epsilon}. \tag{2.2}$$

Let $\mathbf{Y_1} = \mathbf{L}^{-1}\mathbf{Y}$, $\mathbf{X_1} = \mathbf{L}^{-1}\mathbf{X}$ and $\boldsymbol{\epsilon_1} = \mathbf{L}^{-1}\boldsymbol{\epsilon}$. It is easy to verify that $E(\boldsymbol{\epsilon_1}) = \mathbf{0}$ and $Var(\boldsymbol{\epsilon_1}) = \sigma^2\mathbf{I}_N$. Hence, model 2.1 becomes

$$\begin{aligned} \mathbf{Y_1} &= \mathbf{X_1}\boldsymbol{\beta} + \boldsymbol{\epsilon_1}, \\ E(\boldsymbol{\epsilon_1}) &= \mathbf{0}, \\ Var(\boldsymbol{\epsilon_1}) &= \sigma^2\mathbf{I}_N. \end{aligned} \tag{2.3}$$

It is evident that the error $\boldsymbol{\epsilon_1}$ in the model 2.3 satisfies the ordinary linear model assumption. The least-squares function is

$$\begin{aligned} \mathcal{S}(\boldsymbol{\beta}) &= \boldsymbol{\epsilon_1}^T\boldsymbol{\epsilon_1}, \\ &= (\mathbf{Y_1} - \mathbf{X_1}\boldsymbol{\beta})^T(\mathbf{Y_1} - \mathbf{X_1}\boldsymbol{\beta}), \\ &= (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}). \end{aligned} \tag{2.4}$$

To obtain the estimator of $\boldsymbol{\beta}$ in model 2.3, we solve

$$\min \quad (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}). \tag{2.5}$$

with respect to $\boldsymbol{\beta}$. Let $\hat{\boldsymbol{\beta}}_\mathbf{1}$ be the solution of the problem 2.5. Hence, $\hat{\boldsymbol{\beta}}_\mathbf{1}$ satisfies the least-squares normal equations

$$(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})\hat{\boldsymbol{\beta}}_\mathbf{1} = \mathbf{X}^T\mathbf{V}^{-1}\mathbf{Y}. \tag{2.6}$$

It is evident that if the row vectors of $\mathbf{X}$ are linearly independent, then the row vectors of $\mathbf{X_1}$ are also linearly independent. Hence, $\mathbf{X_1}^T\mathbf{X_1} = \mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}$ is invertible and we obtain

$$\hat{\boldsymbol{\beta}}_\mathbf{1} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{Y}. \tag{2.7}$$

Here, $\hat{\boldsymbol{\beta}}_\mathbf{1}$ is called the *WLS estimator* of $\boldsymbol{\beta}$. The covariance matrix of $\hat{\boldsymbol{\beta}}_\mathbf{1}$ is

$$Var(\hat{\boldsymbol{\beta}}_\mathbf{1}) = \sigma^2(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}. \tag{2.8}$$

Note that, elements of $\mathbf{X}$ can be chosen such that multicollinearity does not exist in $\mathbf{X}$. Unfortunately, eigenvalues of $\mathbf{X}^T\mathbf{X}$ are not equal to eigenvalues of $\mathbf{X_1}^T\mathbf{X_1}$. Hence, there is no guarantee that multicollinearity does not exist in $\mathbf{X_1}$.

Let $\hat{\boldsymbol{\beta}}^*_\mathbf{1} = \begin{pmatrix} \hat{\beta}^*_{10} & \hat{\beta}^*_{11} & \ldots & \hat{\beta}^*_{1p} \end{pmatrix}^T \in \mathbb{R}^{p+1}$ be the value of $\hat{\boldsymbol{\beta}}_\mathbf{1}$ when $\mathbf{Y}$ is replaced by $\mathbf{y}$ in the Eq. (2.7). The prediction value of $\mathbf{y_1}(= \mathbf{L}^{-1}\mathbf{y})$, say $\hat{\mathbf{y}}_\mathbf{1}$, is given by

$$\hat{\mathbf{y}}_\mathbf{1} := \begin{pmatrix} \hat{y}_{11} & \hat{y}_{12} & \ldots & \hat{y}_{1N} \end{pmatrix}^T = \mathbf{X}_1\hat{\boldsymbol{\beta}}^*_\mathbf{1}, \tag{2.9}$$

and the residual between $\mathbf{y_1}$ and $\hat{\mathbf{y}}_\mathbf{1}$ is given by

$$\mathbf{e_1} := \begin{pmatrix} e_{11} & e_{12} & \ldots & e_{1N} \end{pmatrix}^T = \mathbf{y_1} - \hat{\mathbf{y}}_\mathbf{1}. \tag{2.10}$$

The RMSE for the WLS regression model is given by

$$RMSE_{wls} := \sqrt{\frac{\mathbf{e_1}^T\mathbf{e_1}}{N}} \tag{2.11}$$

and the *prediction by the WLS regression* is given by

$$f_1(\boldsymbol{x}) := \hat{\beta}^*_{\mathbf{10}} + \sum_{j=1}^{p} \hat{\beta}^*_{\mathbf{1}j} x_j, \tag{2.12}$$

where f_1 is a function from $\mathbb{R}^p$ to $\mathbb{R}$.

3 Weighted Least Squares on Kernel Principal Component Regression

3.1 Regression Model in Feature Space

Assume we have a function $\psi : \mathbb{R}^p \to \mathcal{F}$, where $\mathcal{F}$ is called the *feature space* which we assume is an Euclidean space of higher dimension than p, say p_F. Then, we define $\boldsymbol{\Psi} := \begin{pmatrix}\psi(\mathbf{x}_1) & \dots & \psi(\mathbf{x}_N)\end{pmatrix}^T$, $\mathbf{C} := \frac{1}{N}\boldsymbol{\Psi}^T\boldsymbol{\Psi} = \frac{1}{N}\sum_{i=1}^{N}\psi(\mathbf{x}_i)\psi(\mathbf{x}_i)^T$ and $\mathbf{K} := \boldsymbol{\Psi}\boldsymbol{\Psi}^T$, where sizes of $\boldsymbol{\Psi}$, $\mathbf{C}$ and $\mathbf{K}$ are $N \times p_F$, $p_F \times p_F$ and $N \times N$, respectively. We assume that $\sum_{i=1}^{N}\psi(\mathbf{x}_i) = \mathbf{0}$. If $\mathcal{F}$ is infinite-dimensional, we consider the linear operator $\psi(\mathbf{x}_i)\psi(\mathbf{x}_i)^T$ instead of the matrix $\mathbf{C}$ [13].

The *multiple linear regression model in the feature space* is given by

$$\begin{aligned} \mathbf{Y}_o &= \boldsymbol{\Psi}\boldsymbol{\gamma} + \boldsymbol{\epsilon}_2, \\ E(\boldsymbol{\epsilon}_2) &= \mathbf{0}, \\ Var(\boldsymbol{\epsilon}_2) &= \sigma_2^2\tilde{\mathbf{V}}, \end{aligned} \tag{3.1}$$

where $\boldsymbol{\gamma} = \begin{pmatrix}\gamma_1 & \gamma_2 & \dots & \gamma_{p_F}\end{pmatrix}^T$ is a vector of regression coefficients in the feature space, $\boldsymbol{\epsilon}_2$ is a vector of random errors in the feature space, $\mathbf{Y}_o = (I_N - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^T)\boldsymbol{Y}$ and $\tilde{\mathbf{V}} = diag(1/\tilde{w}_1, 1/\tilde{w}_2, \dots, 1/\tilde{w}_N)$ and $\tilde{w}_i$ is a positive number for $i = 1, 2, \dots, N$. The weight $\tilde{w}_i$ is estimated by using the data $\mathbf{y_o} = (I_N - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^T)\boldsymbol{y}$ and $\mathbf{X}$. Here, we cannot use the generalized inverse matrix to obtain the estimator of $\boldsymbol{\gamma}$ since $\boldsymbol{\Psi}$ is not known explicitly. Let $\tilde{\mathbf{L}} = diag(1/\sqrt{\tilde{w}_1}, 1/\sqrt{\tilde{w}_2}, \dots, 1/\sqrt{\tilde{w}_N})$. Hence, $\tilde{\mathbf{L}}^T = \tilde{\mathbf{L}}$, $\tilde{\mathbf{L}}\tilde{\mathbf{L}}^T = \tilde{\mathbf{V}}$ and $\tilde{\mathbf{L}}^{-1} = diag(\sqrt{\tilde{w}_1}, \sqrt{\tilde{w}_2}, \dots, \sqrt{\tilde{w}_N})$. Then, we have

$$\begin{aligned} \mathbf{Z}_o &= \boldsymbol{\theta}\boldsymbol{\gamma} + \tilde{\boldsymbol{\epsilon}}_2, \\ E(\tilde{\boldsymbol{\epsilon}}_2) &= \mathbf{0}, \\ Var(\tilde{\boldsymbol{\epsilon}}_2) &= \sigma_2^2\mathbf{I}_N, \end{aligned} \tag{3.2}$$

where $\mathbf{Z}_o = \tilde{\mathbf{L}}^{-1}\mathbf{Y}_o$, $\boldsymbol{\theta} = \tilde{\mathbf{L}}^{-1}\boldsymbol{\Psi}$ and $\tilde{\boldsymbol{\epsilon}}_2 = \tilde{\mathbf{L}}^{-1}\boldsymbol{\epsilon}_2$. Furthermore, we define two matrices $\tilde{\mathbf{K}} := \boldsymbol{\theta}\boldsymbol{\theta}^T = \tilde{\mathbf{L}}^{-1}\mathbf{K}\tilde{\mathbf{L}}^{-1}$ and $\tilde{\mathbf{C}} := \frac{1}{N}\boldsymbol{\theta}^T\boldsymbol{\theta}$. The relation of eigenvalues and eigenvectors of the matrices $\tilde{\mathbf{C}}$ and $\tilde{\mathbf{K}}$ are related in the following theorem.

Theorem 3.1. *[16] Suppose $\hat{\lambda} \neq 0$ and $\hat{\mathbf{a}} \in \mathcal{F} \setminus \{\mathbf{0}\}$. The following statements are equivalent:*

1. $\hat{\lambda}$ *and* $\hat{\mathbf{a}}$ *satisfy* $\lambda\mathbf{a} = \tilde{\mathbf{C}}\mathbf{a}$.

2. $\hat{\lambda}$ *and* $\hat{\mathbf{a}}$ *satisfy* $\lambda N\tilde{\mathbf{K}}\boldsymbol{b} = \tilde{\mathbf{K}}^2\boldsymbol{b}$ *and* $\mathbf{a} = \sum_{i=1}^{N} b_i\psi(\mathbf{x}_i)$, *for some* $\mathbf{b} = \begin{pmatrix} b_1 & b_2 & \dots & b_N \end{pmatrix}^T \in \mathbb{R}^N \setminus \{\mathbf{0}\}$.

3. $\hat{\lambda}$ *and* $\hat{\mathbf{a}}$ *satisfy* $\lambda N\tilde{\mathbf{b}} = \tilde{\mathbf{K}}\tilde{\mathbf{b}}$ *and* $\mathbf{a} = \sum_{i=1}^{N} \tilde{b}_i\psi(\mathbf{x}_i)$, *for some* $\tilde{\mathbf{b}} = \begin{pmatrix} \tilde{b}_1 & \tilde{b}_2 & \dots & \tilde{b}_N \end{pmatrix}^T \in \mathbb{R}^N \setminus \{\mathbf{0}\}$.

Let $\hat{p}_F$ be the rank of $\boldsymbol{\theta}$ where $\hat{p}_F \leq \min(N, p_F)$. Since the rank of $\boldsymbol{\theta}$ is equal to the rank of $\tilde{\mathbf{K}}$ and the rank of $\boldsymbol{\theta}^T\boldsymbol{\theta}$, then the rank of $\tilde{\mathbf{K}}$ and the rank of $\boldsymbol{\theta}^T\boldsymbol{\theta}$ are equal to $\hat{p}_F$. Note that, $\tilde{\mathbf{K}}$ is symmetric and positive semidefinite. This implies that the eigenvalues of $\tilde{\mathbf{K}}$ are nonnegative real numbers. Let $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_r \geq \lambda_{r+1} \geq \dots \geq \lambda_{\hat{p}_F} > \lambda_{\hat{p}_F+1} = \dots = \lambda_N = 0$ be the eigenvalues of $\tilde{\mathbf{K}}$ and $\mathbf{B} = (\mathbf{b}_1 \quad \mathbf{b}_2 \quad \dots \quad \mathbf{b}_N)$ be the matrix of the corresponding normalized eigenvectors $\mathbf{b}_i$ $(i = 1, 2, \dots, N)$ of $\tilde{\mathbf{K}}$. Then, let $\boldsymbol{\alpha}_l = \frac{\mathbf{b}_l}{\sqrt{\lambda_l}}$ and $\mathbf{a}_l = \boldsymbol{\theta}^T\boldsymbol{\alpha}_l$ for $l = 1, 2, \dots, \hat{p}_F$. By Theorem 3.1 we obtain

$$\begin{aligned} \tfrac{\lambda_l}{N}\mathbf{a}_l &= \tilde{\mathbf{C}}\mathbf{a}_l \qquad \text{for } l = 1, 2, \dots, \hat{p}_F \\ \mathbf{a}_i^T\mathbf{a}_j &= \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{otherwise,} \end{cases} \qquad \text{for } i, j = 1, 2, \dots, \hat{p}_F, \end{aligned}$$

or equivalent to

$$\begin{aligned} \lambda_l\mathbf{a}_l &= \boldsymbol{\theta}^T\boldsymbol{\theta}\mathbf{a}_l \qquad \text{for } l = 1, 2, \dots, \hat{p}_F \\ \mathbf{a}_i^T\mathbf{a}_j &= \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{otherwise.} \end{cases} \qquad \text{for } i, j = 1, 2, \dots, \hat{p}_F, \end{aligned}$$

Since the rank of $\boldsymbol{\theta}^T\boldsymbol{\theta}$ is equal to $\hat{p}_F$, then the remaining $(p_F - \hat{p}_F)$ eigenvalues of $\boldsymbol{\theta}^T\boldsymbol{\theta}$ are zero eigenvalues. Let λ_k, $(k = \hat{p}_F + 1, \hat{p}_F + 2, \dots, p_F)$, be the zero eigenvalues of $\boldsymbol{\theta}^T\boldsymbol{\theta}$ and $\mathbf{a}_k$ be the normalized eigenvectors of $\boldsymbol{\theta}^T\boldsymbol{\theta}$ corresponding to λ_k. Hence, we have

$$\begin{aligned} \lambda_l\mathbf{a}_l &= \boldsymbol{\theta}^T\boldsymbol{\theta}\mathbf{a}_l \qquad \text{for } l = 1, 2, \dots, p_F \\ \mathbf{a}_i^T\mathbf{a}_j &= \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{otherwise.} \end{cases} \qquad \text{for } i, j = 1, 2, \dots, p_F, \end{aligned}$$

Furthermore, we define $\mathbf{A} = (\mathbf{a}_1 \quad \mathbf{a}_2 \quad \dots \quad \mathbf{a}_{p_F})$. It is evident that $\mathbf{A}$ is an orthogonal matrix, that is, $\mathbf{A}^T = \mathbf{A}^{-1}$. It is not difficult to verify that

$$\mathbf{A}^T\boldsymbol{\theta}^T\boldsymbol{\theta}\mathbf{A} = \mathbf{D},$$

where

$$\mathbf{D} = \begin{pmatrix} \mathbf{D}_{(\hat{p}_F)} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{pmatrix},$$

$$\mathbf{D}_{(\hat{p_F})} = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \lambda_{\hat{p}_F} \end{pmatrix}.$$

and $\mathbf{O}$ is a zero matrix.

By using $\mathbf{A}\mathbf{A}^T = \mathbf{I}_{p_F}$, we can rewrite the model (3.2) as

$$\begin{aligned} \mathbf{Z}_o &= \mathbf{U}\boldsymbol{\vartheta} + \tilde{\boldsymbol{\epsilon}}_{\mathbf{2}}, \\ E(\tilde{\boldsymbol{\epsilon}}_2) &= \mathbf{0}, \\ Var(\tilde{\boldsymbol{\epsilon}}_2) &= \sigma_2^2 \mathbf{I}_N, \end{aligned} \tag{3.3}$$

where $\mathbf{U} = \boldsymbol{\theta}\mathbf{A}$ and $\boldsymbol{\vartheta} = \mathbf{A}^T\boldsymbol{\gamma}$. Let

$$\mathbf{U} = \begin{pmatrix} \mathbf{U}_{(\hat{p}_F)} & \mathbf{U}_{(p_F - \hat{p}_F)} \end{pmatrix} \text{ and } \boldsymbol{\vartheta} = \begin{pmatrix} \boldsymbol{\vartheta}_{(\hat{p}_F)}^T & \boldsymbol{\vartheta}_{(p_F-\hat{p}_F)}^T \end{pmatrix}^T,$$

where sizes of $\mathbf{U}_{(\hat{p}_F)}$, $\mathbf{U}_{(p_F-\hat{p}_F)}$, $\boldsymbol{\vartheta}_{(\hat{p}_F)}$, and $\boldsymbol{\vartheta}_{(p_F-\hat{p}_F)}$ are $N \times \hat{p}_F$, $N \times (p_F - \hat{p}_F)$, $\hat{p_F} \times 1$ and $(p_F - \hat{p}_F) \times 1$, respectively. The model (3.3) can be written as

$$\begin{aligned} \mathbf{Z}_o &= \mathbf{U}_{(\hat{p}_F)}\boldsymbol{\vartheta}_{(\hat{p}_F)} + \mathbf{U}_{(p_F-\hat{p}_F)}\boldsymbol{\vartheta}_{(p_F-\hat{p}_F)} + \tilde{\boldsymbol{\epsilon}}_2, \\ E(\tilde{\boldsymbol{\epsilon}}_2) &= \mathbf{0}, \\ Var(\tilde{\boldsymbol{\epsilon}}_2) &= \sigma_2^2 \mathbf{I}_N. \end{aligned} \tag{3.4}$$

As we see that $\mathbf{D} = \mathbf{A}^T\boldsymbol{\theta}^T\boldsymbol{\theta}\mathbf{A} = \mathbf{U}^T\mathbf{U}$, and we obtain

$$\begin{aligned} \mathbf{U}_{(\hat{p}_F)}^T\mathbf{U}_{(\hat{p}_F)} &= \mathbf{D}_{(\hat{p}_F)}, \\ \mathbf{U}_{(p_F-\hat{p}_F)}^T\mathbf{U}_{(p_F-\hat{p}_F)} &= \mathbf{O}, \end{aligned}$$

and

$$\mathbf{U}_{(\hat{p}_F)}^T\mathbf{U}_{(p_F-\hat{p}_F)} = \mathbf{O}.$$

Since $(\mathbf{U}_{(p_F-\hat{p}_F)}\boldsymbol{\vartheta}_{(p_F-\hat{p}_F)})^T(\mathbf{U}_{(p_F-\hat{p}_F)}\boldsymbol{\vartheta}_{(p_F-\hat{p}_F)})$ is equal to zero, we see that $\mathbf{U}_{(p_F-\hat{p}_F)}\boldsymbol{\vartheta}_{(p_F-\hat{p}_F)}$ is equal to $\mathbf{0}$. Consequently, the model (3.4) is simplified to

$$\begin{aligned} \mathbf{Z}_o &= \mathbf{U}_{(\hat{p}_F)}\boldsymbol{\vartheta}_{(\hat{p}_F)} + \tilde{\boldsymbol{\epsilon}}_2, \\ E(\tilde{\boldsymbol{\epsilon}}_2) &= \mathbf{0}, \\ Var(\tilde{\boldsymbol{\epsilon}}_2) &= \sigma_2^2 \mathbf{I}_N. \end{aligned} \tag{3.5}$$

Let us assume that $\lambda_{r+1}, \lambda_{r+2}, \ldots, \lambda_{\hat{p}_F}$ are close to zero. Let

$$\mathbf{U}_{(\hat{p}_F)} = \begin{pmatrix} \mathbf{U}_{(r)} & \mathbf{U}_{(\hat{p}_F - r)} \end{pmatrix}, \boldsymbol{\vartheta}_{(\hat{p}_F)} = \begin{pmatrix} \boldsymbol{\vartheta}_{(r)}^T & \boldsymbol{\vartheta}_{(\hat{p}_F-r)}^T \end{pmatrix}^T$$

and

$$\mathbf{D}_{(\hat{p}_F)} = \begin{pmatrix} \mathbf{D}_{(r)} & \mathbf{O} \\ \mathbf{O} & \mathbf{D}_{(\hat{p}_F - r)} \end{pmatrix},$$

where

$$\mathbf{D}_{(r)} = \begin{pmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \lambda_r \end{pmatrix},$$

$$\mathbf{D}_{(\hat{p}_F-r)} = \begin{pmatrix} \lambda_{r+1} & 0 & \ldots & 0 \\ 0 & \lambda_{r+2} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \lambda_{\hat{p}_F} \end{pmatrix},$$

and sizes of $\mathbf{U}_{(r)}$, $\mathbf{U}_{(\hat{p}_F-r)}$, $\boldsymbol{\vartheta}_{(r)}$, and $\boldsymbol{\vartheta}_{(\hat{p}_F-r)}$ are $N \times r$, $N \times (\hat{p}_F - r)$, $r \times 1$ and $(\hat{p}_F - r) \times 1$, respectively. The model (3.5) can now be written as

$$\begin{aligned} \mathbf{Z}_o = \mathbf{U}_{(r)}\boldsymbol{\vartheta}_{(r)} + \mathbf{U}_{(\hat{p}_F-r)}\boldsymbol{\vartheta}_{(\hat{p}_F-r)} + \tilde{\boldsymbol{\epsilon}}_2 \\ E(\tilde{\boldsymbol{\epsilon}}_2) = \mathbf{0}, \\ Var(\tilde{\boldsymbol{\epsilon}}_2) = \sigma_2^2\mathbf{I}_N, \end{aligned} \tag{3.6}$$

It is evident that the estimator of $\boldsymbol{\vartheta}_{(\hat{p}_F-r)}$, say $\hat{\boldsymbol{\vartheta}}_{(\hat{p}_F-r)} = (\hat{\vartheta}_{r+1} \quad \hat{\vartheta}_{r+2} \quad \dots \quad \hat{\vartheta}_{\hat{p}_F-r})^T$, is given by

$$\hat{\boldsymbol{\vartheta}}_{(\hat{p}_F-r)} = (\mathbf{U}_{(\hat{p}_F-r)}^T\mathbf{U}_{(\hat{p}_F-r)})^{-1}\mathbf{U}_{(\hat{p}_F-r)}^T\mathbf{Z}_o = \mathbf{D}_{(\hat{p}_F-r)}^{-1}\mathbf{U}_{(\hat{p}_F-r)}^T\mathbf{Z}_o, \tag{3.7}$$

and the variance of $\hat{\vartheta}_j$ $(j = r+1, \dots, \hat{p}_F - r)$ is

$$Var(\hat{\vartheta}_j) = \sigma^2(\mathbf{D}_{(\hat{p}-r)}^{-1})_{jj}. \tag{3.8}$$

Since $\lambda_{r+1}, \lambda_{r+2}, \dots, \lambda_{\hat{p}_F-r}$ are close to zero, the diagonal elements of $\mathbf{D}_{(\hat{p}_F-r)}^{-1}$ and also the variance of $\hat{\vartheta}_j$ $(j = r+1, \dots, \hat{p}_F - r)$ will be very large numbers. Thus, we encounter the ill effect of multicollinearity in the model (3.6). To avoid the effects of multicollinearity, we drop the term $\mathbf{U}_{(\hat{p}_F-r)}\boldsymbol{\vartheta}_{(\hat{p}_F-r)}$ as in [15] and obtain

$$\mathbf{Z}_o = \mathbf{U}_{(r)}\boldsymbol{\vartheta}_{(r)} + \tilde{\tilde{\boldsymbol{\epsilon}}}, \tag{3.9}$$

where $\tilde{\tilde{\boldsymbol{\epsilon}}}$ is a random vector influenced by dropping $\mathbf{U}_{(\hat{p}_F-r)}\boldsymbol{\vartheta}_{(\hat{p}_F-r)}$ in the model (3.9). The model (3.9) shows that the ill effects multicollinearity on $\mathbf{U}_{(\hat{p}_F)}$ are avoided by using the matrix $\mathbf{A}$[1].

Note that $\mathbf{U}_{(r)}^T\mathbf{U}_{(r)} = \mathbf{D}_{(r)}$, which is invertible. Hence, the estimator of $\boldsymbol{\vartheta}_{(\tilde{r})}$, say $\hat{\boldsymbol{\vartheta}}_{(\tilde{r})}$, is given by

$$\hat{\boldsymbol{\vartheta}}_{(r)} = (\mathbf{U}_{(r)}^T\mathbf{U}_{(r)})^{-1}\mathbf{U}_{(r)}^T\mathbf{Z}_o. \tag{3.10}$$

Let $\mathbf{z_o} = \tilde{\mathbf{L}}^{-1}(I_N - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^T)\boldsymbol{y}$ be the observed data corresponding to $\mathbf{Z_o}$ and $\hat{\boldsymbol{\vartheta}}_{(\tilde{r})}^* \in \mathbb{R}^{\tilde{r}}$ be the value of $\hat{\boldsymbol{\vartheta}}_{(\tilde{r})}$ when $\mathbf{Z}_o$ is replaced by $\mathbf{z}_o$ in the Eq. (3.10). Hence

$$\begin{aligned} \hat{\boldsymbol{\vartheta}}_{(r)}^* &= (\mathbf{U}_{(r)}^T\mathbf{U}_{(r)})^{-1}\mathbf{U}_{(r)}^T\mathbf{z}_o, \\ &= \mathbf{D}_{(r)}^{-1}\mathbf{U}_{(r)}^T\mathbf{z}_o. \end{aligned} \tag{3.11}$$

Since

$$\mathbf{U} = \begin{pmatrix} \mathbf{U}_{(r)} & \mathbf{U}_{(\hat{p}_F-r)} & \mathbf{U}_{(p_F-\hat{p}_F)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\theta}\mathbf{A}_{(r)} & \boldsymbol{\theta}\mathbf{A}_{(\hat{p}_F-r)} & \boldsymbol{\theta}\mathbf{A}_{(p_F-\hat{p}_F)} \end{pmatrix},$$

we obtain $\mathbf{U}_{(r)} = \boldsymbol{\theta}\mathbf{A}_{(r)}$. As we see that $\mathbf{A}_{(r)} = \boldsymbol{\theta}^T \begin{pmatrix} \boldsymbol{\alpha}_1 & \boldsymbol{\alpha}_2 & \dots & \boldsymbol{\alpha}_r \end{pmatrix}$. Hence,

$$\mathbf{U}_{(r)} = \boldsymbol{\theta}\boldsymbol{\theta}^T\boldsymbol{\Gamma}_{(r)} = \tilde{\mathbf{K}}\boldsymbol{\Gamma}_{(r)}, \tag{3.12}$$

where $\boldsymbol{\Gamma}_{(r)} = \begin{pmatrix} \boldsymbol{\alpha}_1 & \boldsymbol{\alpha}_2 & \dots & \boldsymbol{\alpha}_r \end{pmatrix}$. However, we do not know $\mathbf{U}_{(r)}$ explicitly yet. Let us consider the following Theorem:

[1]To detect multicollinearity on $\mathbf{U}_{(\hat{p}_F)}$, we use the ratio λ_l/λ_1 for $l = 1, 2, \dots, \hat{p}_F$. If λ_l/λ_1 is smaller than, say $< \frac{1}{1000}$, then we consider that multicollinearity exists on $\mathbf{U}_{(\hat{p}_F)}$ [8].

Theorem 3.2. *(Mercer [7, 12]) For any symmetric, continuous and positive semi-definite function $\xi : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$, there exists a function $\phi : \mathbb{R}^p \to \mathcal{F}$ such that*

$$\xi(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{x})^T \phi(\mathbf{y}).$$

By using Theorem 3.2, if we choose a continuous, symmetric and positive semidefinite function $\kappa : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$, then there exists $\phi : \mathbb{R}^p \to \mathcal{F}$ such that $\kappa(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j)$. The function κ is called the *kernel* function. Instead of choosing ψ explicitly, we choose a kernel κ and employ the corresponding function ϕ as ψ . Let $K_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}_j)$. Hence, we have

$$\mathbf{K} = \begin{pmatrix} K_{11} & K_{12} & \dots & K_{1N} \\ K_{21} & K_{22} & \dots & K_{2N} \\ \dots & \dots & \dots & \dots \\ K_{N1} & K_{N2} & \dots & K_{NN} \end{pmatrix}.$$

Now, we know $\mathbf{K}$ explicitly. This implies that $\tilde{\mathbf{K}} = \tilde{\mathbf{L}}^{-1}\mathbf{K}\tilde{\mathbf{L}}^{-1}$, $\mathbf{\Gamma}_{(r)}$ and $\mathbf{U}_{(r)}$ are also known explicitly.

The *prediction value* of $\mathbf{z}_o$ $(= \tilde{\mathbf{L}}^{-1}(I_N - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^T)\mathbf{y})$, say $\hat{\mathbf{z}}_o$ $(= \tilde{\mathbf{L}}^{-1}(I_N - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^T)\hat{\mathbf{y}})$ is given by

$$\hat{\mathbf{z}}_\mathbf{o} := \begin{pmatrix} \hat{z}_{\mathbf{o}1} & \hat{z}_{\mathbf{o}2} & \dots & \hat{z}_{\mathbf{o}N} \end{pmatrix}^T = \tilde{\mathbf{K}}\mathbf{\Gamma}_{(r)}\hat{\boldsymbol{\vartheta}}^*_{(r)}. \tag{3.13}$$

The residual between $\mathbf{z}_o$ and $\hat{\mathbf{z}}_o$ is given by

$$\mathbf{e_2} := \begin{pmatrix} e_{\mathbf{2}1} & e_{\mathbf{2}2} & \dots & e_{\mathbf{2}N} \end{pmatrix}^T = \mathbf{z}_o - \hat{\mathbf{z}}_o, \tag{3.14}$$

and the *prediction by the WLS KPCR* is given by

$$g(\mathbf{x}) := \bar{y} + \sum_{i=1}^{N} c_i \kappa(\mathbf{x}, \mathbf{x}_i), \tag{3.15}$$

where g is a function from $\mathbb{R}^p$ to $\mathbb{R}$ and $\begin{pmatrix} c_1 & c_2 & \dots & c_N \end{pmatrix}^T = \tilde{\mathbf{L}}^{-1}\mathbf{\Gamma}_{(r)}\hat{\boldsymbol{\vartheta}}^*_{(r)}$. The number r is called the *retained number of nonlinear PCs for the WLS KPCR.*

3.2 Algorithms

3.2.1 The KPCR's Algorithm

The KPCR's algorithm is given by the following steps.

Algorithm:

1. Given $(y_i, x_{i1}, x_{i2}, \dots, x_{ip})$, $i = 1, 2, \dots, N$.
2. Calculate $\bar{y} = \frac{1}{N}\mathbf{1}_N^T\mathbf{y}$.
3. Choose a kernel $\kappa : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$.
4. Construct $K_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}_j)$ and $\mathbf{K} = (K_{ij})$.
5. Diagonalize $\mathbf{K}$.
   Let $\mu_1 \geq \mu_2 \geq \dots \geq \mu_{\hat{r}} \geq \dots \geq \mu_{\acute{p}} > \mu_{\acute{p}+1} = \dots = \mu_N = 0$ be the eigenvalues of $\mathbf{K}$ and $\tilde{\mathbf{b}}_1, \tilde{\mathbf{b}}_2, \dots, \tilde{\mathbf{b}}_N$ be the corresponding normalized eigenvectors of $\mathbf{K}$.

6. Detect collinearity and multicollinearity on $\mathbf{K}$.
   Let $\hat{r}$ be the retained number of nonlinear PCs such that $\hat{r} = \max\{s|\frac{\mu_s}{\mu_1} \geq \frac{1}{1000}\}$.

7. Construct $\boldsymbol{\eta}_l := \frac{\tilde{\mathbf{b}}_l}{\sqrt{\tilde{\lambda}_l}}$ for $l = 1, 2, \ldots, \hat{r}$ and $\hat{\boldsymbol{\Gamma}}_{(\hat{r})} := \begin{pmatrix} \boldsymbol{\eta}_1 & \boldsymbol{\eta}_2 & \ldots & \boldsymbol{\eta}_{\hat{r}} \end{pmatrix}$.

8. Calculate $\mathbf{W}_{(\hat{r})} := \mathbf{K}\hat{\boldsymbol{\Gamma}}_{(\hat{r})}$, $\hat{\boldsymbol{\varpi}}^*_{(\hat{r})} := \hat{\mathbf{D}}^{-1}_{(\hat{r})}\mathbf{W}^T_{(\hat{r})}\mathbf{y}$
   and $\mathbf{d} := \begin{pmatrix} d_1 & d_2 & \ldots & d_N \end{pmatrix}^T = \hat{\boldsymbol{\Gamma}}_{(\hat{r})}\hat{\boldsymbol{\varpi}}^*_{(\hat{r})}$, where $\hat{\mathbf{D}}_{(\hat{r})} = \begin{pmatrix} \mu_1 & 0 & \ldots & 0 \\ 0 & \mu_2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \mu_{\hat{r}} \end{pmatrix}$.

9. Given a vector $\mathbf{x} \in \mathbb{R}^p$, the *prediction by KPCR* is given by

$$h(\mathbf{x}) := \bar{y} + \textstyle\sum_{j=1}^N d_i \kappa(\mathbf{x}, \mathbf{x}_j).$$

Note that the above algorithm works under the assumption $\sum_{i=1}^N \psi(\mathbf{x}_i) = \mathbf{0}$. When $\sum_{i=1}^N \psi(\mathbf{x}_i) \neq \mathbf{0}$, we have only to replace $\mathbf{K}$ by $\mathbf{K}_N := \mathbf{K} - \mathbf{E}\mathbf{K} - \mathbf{K}\mathbf{E} + \mathbf{E}\mathbf{K}\mathbf{E}$ in Step 4, where $\mathbf{E}$ is the $N \times N$ matrix with all elements equal to $\frac{1}{N}$. Further, we diagonalize $\mathbf{K}_N$ in Step 5 and work based on $\mathbf{K}_N$ in the subsequent steps.

3.2.2 The WLS KPCR's Algorithm

We summarize the procedure in Subsection 3.1 to obtain the prediction by WLS KPCR.

Algorithm:

1. Given $(y_i, x_{i1}, x_{i2}, \ldots, x_{ip})$, $i = 1, 2, \ldots, N$.
2. Calculate $\bar{y} = \frac{1}{N}\mathbf{1}_N^T\mathbf{y}$ and $\boldsymbol{y}_o = (I_N - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^T)\boldsymbol{y}$.
3. Estimate $\tilde{\mathbf{V}}$ and find $\tilde{\mathbf{L}}$.
4. Calculate $\mathbf{z}_o = \tilde{\mathbf{L}}^{-1}\mathbf{y}_o$.
5. Choose a kernel $\kappa : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$.
6. Construct $K_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}_j)$, $\mathbf{K} = (K_{ij})$ and $\tilde{\mathbf{K}} = \tilde{\mathbf{L}}^{-1}\mathbf{K}\tilde{\mathbf{L}}^{-1}$.
7. Diagonalize $\tilde{\mathbf{K}}$.
   Let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r \geq \ldots \geq \lambda_{\hat{p}_F} > \lambda_{\hat{p}_F+1} = \ldots = \lambda_N = 0$ be the eigenvalues of $\tilde{\mathbf{K}}$ and $\mathbf{b}_1, \mathbf{b}_2 \ldots, \mathbf{b}_N$ be the corresponding normalized eigenvectors of $\tilde{\mathbf{K}}$.
8. Detect collinearity and multicollinearity on $\tilde{\mathbf{K}}$.
   Let r be the retained number of nonlinear PCs such that $r = \max\{s|\frac{\lambda_s}{\lambda_1} \geq \frac{1}{1000}\}$.
9. Construct $\boldsymbol{\alpha}_l = \frac{\mathbf{b}_l}{\sqrt{\lambda_l}}$ for $l = 1, 2, \ldots, r$ and $\boldsymbol{\Gamma}_{(r)} = \begin{pmatrix} \boldsymbol{\alpha}_1 & \boldsymbol{\alpha}_2 & \ldots & \boldsymbol{\alpha}_r \end{pmatrix}$.

10. Calculate $\mathbf{U}_{(r)} = \tilde{\mathbf{K}}\mathbf{\Gamma}_{(r)}$, $\hat{\boldsymbol{\vartheta}}^*_{(r)} = \mathbf{D}^{-1}_{(r)}\mathbf{U}^T_{(r)}\mathbf{z}_o$ and $\mathbf{c} = \begin{pmatrix} c_1 & c_2 & \dots & c_N \end{pmatrix}^T = \tilde{\mathbf{L}}^{-1}\mathbf{\Gamma}_{(r)}\hat{\boldsymbol{\vartheta}}^*_{(r)}$.

11. Given a vector $\mathbf{x} \in \mathbb{R}^p$, the prediction by WLS KPCR is given by

$$g(\mathbf{x}) = \bar{y} + \textstyle\sum_{j=1}^N c_i \kappa(\mathbf{x}, \mathbf{x}_j).$$

We also notice that the above algorithm works under the assumption $\sum_{i=1}^N \psi(\mathbf{x}_i) = \mathbf{0}$. When $\sum_{i=1}^N \psi(\mathbf{x}_i) \neq \mathbf{0}$, we have only to replace $\mathbf{K}$ by $\mathbf{K}_N$ in Step 6. In addition, the *cross validation (CV)* method can be used to obtain the appropriate of the retained number of nonlinear PCs. The CV technique has a large literature, see for example [3, 6, 8, 14]. Let us consider Eq. (3.5) again. It is evident that the estimator of $\boldsymbol{\vartheta}_{(\hat{p}_F)}$ corresponding to $\mathbf{z}_o$, say $\hat{\boldsymbol{\vartheta}}^*_{(\hat{p}_F)} = (\hat{\vartheta}^*_1 \quad \hat{\vartheta}^*_2 \quad \dots \quad \hat{\vartheta}^*_{\hat{p}_F})^T$, is given by

$$\hat{\boldsymbol{\vartheta}}^*_{(\hat{p}_F)} = (\mathbf{U}^T_{(\hat{p}_F)}\mathbf{U}_{(\hat{p}_F)})^{-1}\mathbf{U}^T_{(\hat{p}_F)}\mathbf{z}_o = \mathbf{D}^{-1}_{(\hat{p}_F)}\mathbf{U}^T_{(\hat{p}_F)}\mathbf{z}_o, \tag{3.16}$$

where $\mathbf{U}_{(\hat{p}_F)} = \tilde{\mathbf{K}}\mathbf{\Gamma}_{(\hat{p}_F)}$ and $\mathbf{\Gamma}_{(\hat{p}_F)} = \begin{pmatrix} \boldsymbol{\alpha}_1 & \boldsymbol{\alpha}_2 & \dots & \boldsymbol{\alpha}_{\hat{p}_F} \end{pmatrix}$. The *prediction by the WLS KPCR with the first $\tilde{r}$ vectors of $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \dots, \boldsymbol{\alpha}_{\hat{p}_F}$* is given by

$$g_{(\tilde{r})}(\mathbf{x}) := \bar{y} + \sum_{i=1}^N \tilde{c}_i \kappa(\mathbf{x}, \mathbf{x}_i), \tag{3.17}$$

where $g_{(\tilde{r})}$ is a function from $\mathbb{R}^p$ to $\mathbb{R}$, $\begin{pmatrix} \tilde{c}_1 & \tilde{c}_2 & \dots & \tilde{c}_N \end{pmatrix}^T = \tilde{\mathbf{L}}^{-1}\mathbf{\Gamma}_{(\tilde{r})}\hat{\boldsymbol{\vartheta}}^*_{(\tilde{r})}$, $\hat{\boldsymbol{\vartheta}}^*_{(\tilde{r})} = (\hat{\vartheta}^*_1 \quad \hat{\vartheta}^*_2 \quad \dots \quad \hat{\vartheta}^*_{(\tilde{r})})^T$ and $\mathbf{\Gamma}_{(\tilde{r})} = \begin{pmatrix} \boldsymbol{\alpha}_1 & \boldsymbol{\alpha}_2 & \dots & \boldsymbol{\alpha}_{\tilde{r}} \end{pmatrix}$.

In the CV technique, the original data are partitioned into L disjoint subsets where L is a positive integer. A subset data, say G_k $(k = 1, 2, \dots, L)$, is chosen as the validation for testing the prediction model and the remaining $L - 1$ subsets data are used to estimate the regression coefficients $\boldsymbol{\vartheta}_{(\tilde{r})}$. The CV technique uses the *prediction error sum of squares (PRESS)* to obtain the appropriate $\tilde{r}$, say $\tilde{r}^*$. The *PRESS* of G_k is given by

$$PRESS(G_k)_{(\tilde{r})} = \sum_{s=1}^{m_k} (y^k_s - g_{(\tilde{r})}(\mathbf{x}^k_s))^2, \tag{3.18}$$

where $\mathbf{x}^k_s$ and y^k_s are contained in G_k and m_k is the cardinality of G_k. Then, $PRESS(G_k)$ is summed over all the subsets data, say,

$$PRESS_{(\tilde{r})} = \sum_{k=1}^{L} PRESS(G_k)_{(\tilde{r})}. \tag{3.19}$$

Then, the number $\tilde{r}^*$ is chosen such that $PRESS_{(\tilde{r}^*)} \leq PRESS_{(\tilde{r})}$ for $\tilde{r} = 1, 2, \dots, \hat{p}_F$. Note that, the number L and G_k $(k = 1, 2, \dots, L)$ are chosen by trial and error. We also note that when m_k is equal to one for $k = 1, 2, \dots, L$ and N is a very large number, the CV technique can be inefficient. The CV technique with m_k equal to one for $k = 1, 2, \dots, L$ is known as the *leave one out cross validation*.

4 Case Study

We fix the number of regressors to one and use the Gaussian kernel $\kappa(\mathbf{x}, \mathbf{y}) = \exp(\frac{-\|\mathbf{x}-\mathbf{y}\|^2}{\rho})$ where ρ is the parameter of the kernel. There are several method to estimate the weight w_i [2, 3, 8, 14]. Here, we use the method based on replication to estimate the weight w_i. First, we arrange the data x in order of increasing y_i. Then, we make some groups, say M ($< N$) groups, of the ordered data. Let the kth group, $k = 1, 2, \ldots, M$, contains $\{(\acute{y}_{\hat{i}k}, \acute{x}_{\hat{i}k})\}$ for some $\hat{i} = 1, 2, \ldots, N$ where $\acute{y}_{\hat{i}k} \in \{y_1, y_2, \ldots, y_N\}$ and $\acute{x}_{\hat{i}k} \in \{x_1, x_2, \ldots, x_N\}$. Let $\bar{\acute{x}}_k$ and s_k^2 be the average of $\{\acute{x}_{\hat{i}k}\}$ and variance of $\{\acute{y}_{\hat{i}k}\}$, respectively. Then, we make the prediction from the set $\{(\bar{\acute{x}}_k, s_k^2)\}$, say $f_2(x) = \hat{c}_0 + \hat{c}_1 x$, where f_3 is a function from $\mathbb{R}$ to $\mathbb{R}$ and $\hat{c}_0, \hat{c}_1 \in \mathbb{R}$. Further, we calculate the estimated variance of y_i by using the predictor $f_2(x_i)$. The weight w_i is chosen inversely from the estimated variance of y_i. For KPCR and WLS KPCR, the procedure to obtain their weights is straightforward as the explained procedure. We just replace y_i by y_{oi} where y_{oi} is the ith element of $\mathbf{y}_o$.

In this case study, we use the average monthly income from food sales (y) and the corresponding annual advertising expenses (x) for 30 restaurants [8]. The data are given in Table 1. The data in Table 1 are called the *training data*. We use some of the data to test the prediction by the ordinary linear regression, WLS regression, KPCR and WLS KPCR. We call the data which is used to test the prediction of those methods the *testing data*. As mentioned in Section 1, the plot of the residual e_i and its the corresponding $\hat{y}_i$ is useful to check the assumption of constant variance. The plot of e_i and $\hat{y}_i$ is shown in Figure 1(a). In Figure 1(a), the variation of the residuals increases significantly as the prediction values increase. Hence, this plot indicates violation of the assumption of constant variance. We can also see that the residual e_i has a relatively large number. This implies that $RMSE_{olr}$ is also a large number. For the shake of comparison, the values of M are chosen to be two, four and five. For instance $M = 2$, it means that the ordered data is divided into two groups where each group contains 50 percentage of the ordered data. The plot of residual $e_{\mathbf{2}i}$ and its corresponding prediction value $\hat{z}_{\mathbf{o}i}$ with $M = 2$ and $\rho = 1$ is shown in Figure 1(b). In comparison to the plot in Figure 1(a), it is much more improved since the residual $e_{\mathbf{2}i}$ has a much smaller number than e_i. Beside that, Figure 1(b) shows a residual plot with no systematic pattern around zero. It seems that the assumption of constant variance are satisfied for the data associated with this residual plot

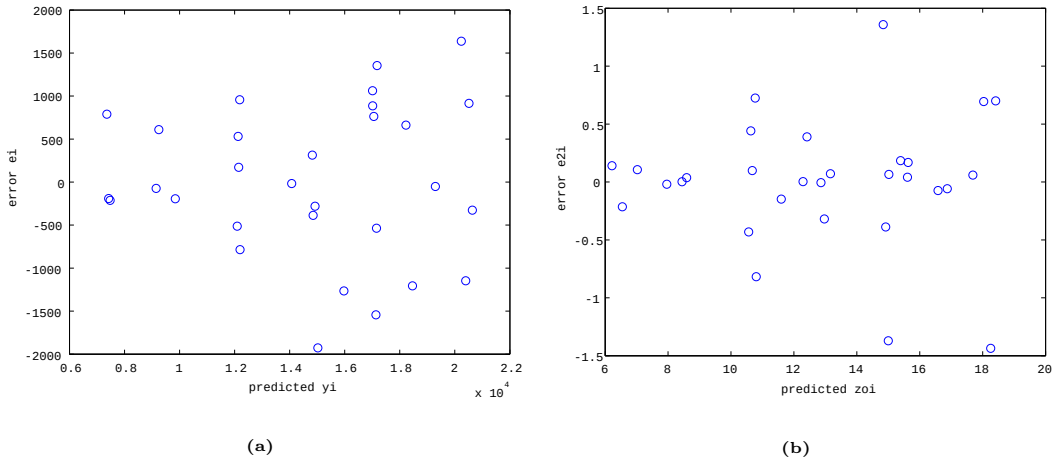
The results of this study are given in Table 2. Note that, multicollinearity exists in the regression matrix for both of the ordinary linear regression model and the WLS regression model. In Table 2, we can see that the WLS KPCR model significantly decreases the RMSEs of the ordinary linear regression and KPCR model. For this data, the choice $M = 2$ yields the better result than that of $M = 4$ and $M = 5$.

5 Conclusion

The WLS regression is a technique to be used in case of the regression model with variance of random errors having unequal values in its diagonal elements. However, this technique yields a linear prediction model and has no guarantee that it can handle the effects of multicollinearity. Although KPCR can be used to handle the

Table 1: The restaurant foods sales data ($y_i \times 100$)

Obs.	1	2	3	4	5	6	7	8	9	10
x	3.00	3.150	3.085	5.225	5.350	6.090	8.925	9.015	8.885	8.950
y	81.464	72.661	72.344	90.743	98.588	96.507	126.574	114.133	115.814	123.181
11	12	13	14	15	16	17	18	19	20	
9.00	11.345	12.275	12.400	12.525	12.310	13.700	15.000	15.175	14.995	
131.434	140.564	151.352	146.426	130.963	144.630	147.041	179.021	166.200	180.732	
21	22	23	24	25	26	27	28	29	30	
15.050	15.200	15.150	16.800	16.500	17.830	19.500	19.200	19.000	19.350	
178.187	185.304	155.931	172.579	188.851	192.424	203.112	192.482	218.715	214.317	



(a) (b)

Figure 1: A plot of the residual and its corresponding predicted value for training data: (a) ordinary linear regression model, (b) WLS KPCR.

Table 2: The RMSE of the ordinary linear regression, WLS linear regression, KPCR and WLS KPCR for the restaurant foods sales data.

	Model	RMSE		
		M=2	M=4	M=5
The training data	ordinary linear regression	869.8845	869.8845	869.8845
	KPCR ($\rho = 0.5$, $\hat{r}$= 18)	601.3838	601.3838	601.3838
	KPCR ($\rho = 1$, $\hat{r}$= 17)	624.2270	624.2270	624.2270
	WLS linear regression	0.3834	0.5471	0.6958
	WLS KPCR ($\rho = 0.5$, r= 18)	0.2769	0.4178	0.5258
	WLS KPCR ($\rho = 1$, r=17)	0.2874	0.4336	0.5457
The testing data	ordinary linear regression	834.9586	834.9586	834.9586
	KPCR ($\rho = 0.5$, $\hat{r}$=18)	689.8944	689.8944	689.8944
	KPCR ($\rho = 1$, $\hat{r}$=17)	721.4072	721.4072	721.4072
	WLS linear regression	0.5380	1.2599	0.5639
	WLS KPCR ($\rho = 0.5$, r=18)	0.3973	0.7893	0.4314
	WLS KPCR ($\rho = 1$, r=17)	0.4229	0.8984	0.4573

limitations of the linearity and the effect of multicollinearity, but KPCR can be inappropriate. Since KPCR was constructed by the assumption that the variance of random errors having equal values in its diagonal elements.

In this paper, we proposed the WLS KPCR to be used in case of the regression model with variance of random errors having unequal values in its diagonal elements. This method yields a nonlinear prediction model and it can avoid the effects of multicollinearity. In our case study, the WLS KPCR yields the better result that of the ordinary linear regression, the WLS linear regression and the KPCR.

Acknowlegdement
The author thank Professor Yoshitsugu Yamamoto, University of Tsukuba, for comments and suggestions. The author also thank the Ministry of Education, Culture, Sports, Science and Technology Japan.

References

[1] Howard Anton. *Elementary Linear Algebra.* John Wiley and Sons, Inc., 2000.

[2] Norman R. Draper and Harry Smith. *Applied Regression Analysis.* John Wiley and Sons, 1998.

[3] Julian J. Faraway. *Linear Models with R.* Chapman and Hall/CRC, 2005.

[4] L. Hoegaerts, J.A.K. Suykens, J. Vandewalle, and B. De Moor. Subset based least squares subspace in reproducing kernel hilbert space. *Neurocomputing*, pages 293–323, 2005.

[5] A.M. Jade, B. Srikanth, B.D Kulkari, J.P Jog, and L. Priya. Feature extraction and denoising using kernel pca. *Chemical Engineering Sciences*, 58:4441–4448, 2003.

[6] I.T. Jolliffe. *Principal Component Analysis.* Springer, 2002.

[7] Ha Quang Minh, Partha Niyogi, and Yuan Yao. Mercer's theorem, feature maps, and smoothing. *Lecture Notes in Computer Science, Springer Berling*, 4005/2006, 20009.

[8] Douglas C. Montgomery, Elizabeth A. Peck, and G. Geoffrey Vining. *Introduction to Linear Regression.* Wiley-Interscience, 2006.

[9] Roman Rosipal, Mark Girolami, Leonard J. Trejo, and Andrzej Cichoki. Kernel pca for feature extraction and de-noising in nonlinear regression. *Neural Computing and Applications*, pages 231–243, 2001.

[10] Roman Rosipal and Leonard J. Trejo. Kernel partial least squares regression in reproducing kernel hilbert space. *Journal of Machine Learning Research 2*, pages 97–123, 2002.

[11] Roman Rosipal, Leonard J. Trejo, and Andrzej Cichoki. Kernel principal component regression with em approach to nonlinear principal component extraction. *Technical Report, University of Paisley, UK*, 2001.

[12] B. Scholkopf, A. Smola, and K.R. Muller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10:1299–1319, 1998.

[13] Bernhard Scholkopf and Alexander J. Smola. *Learning with kernels*. The MIT Press., 2002.

[14] George A.F. Seber and Alan J. Lee. *Linear Regression Analysis*. John Wiley and Sons, Inc., 2003.

[15] M.S. Srivastava. *Methods of Multivariate Statistics*. John Wiley and Sons, Inc., 2002.

[16] Antoni Wibowo and Yoshitsugu Yamamoto. The new approach for kernel principal component regression. *Discussion Paper Series No. 1195,Department of Social Systems and Management, Univ. of Tsukuba*, 2008.