Statistica Sinica **17**(2007), 1657-1675

TRANSFORMED PARTIAL LEAST SQUARES FOR MULTIVARIATE DATA

Li-Xing Zhu, Li-Ping Zhu and Xin Li

Hong Kong Baptist University, East China Normal University and Beijing Institute of Technology

Abstract: The research described herein is motivated by a study of the relationship between agricultural meteorology and three major yields of crops in a province of China. To build a regression model for this data set with multivariate response and high-dimensional covariates, three issues are of particular interest: reducing the dimension of the covariates, avoiding the collinearity between the components of the covariates, and capturing the nonlinearity structure. To deal with these problems, we propose a method of nonparametric response transformation to build a single-index type model, and use partial least squares to reduce the dimension of covariates and to overcome the problem of collinearity. Our method is an alternative approach to sliced inverse regression when the underlying model is single-index type. To select the transformations, a new criterion based on maximizing the covariance matrix is recommended. The selected transformations are estimated by splines; here B-splines are used for general cases and I-splines with a penalty function are suggested when the transformations are monotonic. A modified BIC selection principle is proposed to determine the dimensionality of the space of spline transformations. The consistency of the estimators is proved and easily implemented algorithms are provided. Application to the agricultural data set is carried out.

Key words and phrases: Agricultural meteorology, canonical correlation analysis, dimension reduction, model selection, partial least squares regression, sliced inverse regression, spline, transformation.

1. Introduction

This paper is motivated by research regarding an agricultural, meteorological disaster in Ji-Lin, a northeastern province of China. Researchers have tried to explore the relationship between meterological conditions and the yields of three crops: soybean, rice, and maize. Ma (1996) reported the yields of 33 years, 1958 to 1990, with 17 measurements of climate change for the different growth periods of crops, including temperature, rain quantity and hours of sunshine. Compared with the sample size, the dimension of the covariates is high. Ma (1996) used an ordinary multivariate linear least squares regression model (OLS) to analyze this data set. However, the multiple correlation coefficients between each of the components of $\boldsymbol{Y}$ and $\boldsymbol{X}$ are only 0.48, 0.47 and 0.33. One suspects nonlinearity and this is supported by our data analysis in Section 6. Furthermore, Li (1999) revealed high collinearity between the components of the covariates. For example, average temperature is highly and positively correlated with sunshine, and sunshine in this region is quite negatively correlated with rain. The covariance matrix of the covariates is nearly singular, with a smallest eigenvalue of only 0.001. In this paper, we tackle three issues: nonlinearity, collinearity and dimensionality when we estimate a regression function.

In a more general setting, we consider the regression of a q-dimensional response $\boldsymbol{Y} = (Y_1, \ldots, Y_q)^T$ on p-dimensional predictors $\boldsymbol{X} = (X_1, \ldots, X_p)^T$. Many recent dimension reduction methods can handle both nonlinearity and high-dimensionality of the data. For instance, single-index models are popularly used when $q = 1$. In these models the response $\boldsymbol{Y}$ is independent of $\boldsymbol{X}$ when a linear combination of $\boldsymbol{X}$, say $\boldsymbol{\beta}_1^T\boldsymbol{X}$, is given, denoted by $Y \perp\!\!\!\perp X | \beta_1^T\boldsymbol{X}$. Projection pursuit regression (PPR) proposed by Friedman and Stuetzle (1981) can be used to estimate $\boldsymbol{\beta}$ by minimizing the least squared distance between Y and $h(\boldsymbol{\beta}^T\boldsymbol{X})$ over all projection directions $\boldsymbol{\beta}$ and all functions $h(\cdot)$. The optimal function $h(\cdot)$ is the conditional expectation $E(\boldsymbol{Y}|\boldsymbol{\beta}_1^T\boldsymbol{X})$, which can also be estimated by any nonparametric smoothing method, splines for example. This method has been investigated by many authors, examples are Huber (1985), Hall (1989) and Zhu and Fang (1992).

Li and Duan (1989) brought in the notion of inverse regression, later developed by Li (1991) into the method of sliced inverse regression (SIR). Instead of studying the relationship between $\boldsymbol{Y}$ and functions of $\boldsymbol{X}$, Li (1991) considered maximizing the correlation between the projected covariates $\boldsymbol{\beta}^T\boldsymbol{X}$ and a transformed response $h(\boldsymbol{Y})$, over all $\boldsymbol{\beta}$ and $h(\cdot)$ the class of square integrable functions with respect to the distribution of X. The optimal $h(\cdot)$ is the inverse regression function of Y, $E(X^T\boldsymbol{\beta}|Y)$, and the projection direction $\boldsymbol{\beta}$ can be determined through an eigen-decomposition of the matrix $\mathrm{Cov}\,(E(\boldsymbol{X}|\boldsymbol{Y}))$. Generally, $\mathrm{Cov}\,(\boldsymbol{U}, \boldsymbol{V}^T)$ stands for the covariance matrix between $\boldsymbol{U} = (U_1, \ldots, U_l)^T$ and $\boldsymbol{V} = (V_1, \ldots, V_m)^T$; when $\boldsymbol{U} = \boldsymbol{V}$, we write $\mathrm{Cov}\,(\boldsymbol{U}) = \mathrm{Cov}\,(\boldsymbol{U}, \boldsymbol{U}^T)$. For identifying β, or more generally the central dimension-reduction subspace (Cook (1998)), Cook and Weisberg (1994) used second moments to develop sliced average variance estimation (SAVE). Under some more conditions, SAVE can also determine the central space. Cook and Li (2002) studied dimension reduction for the conditional mean in regression, Yin and Cook (2002, 2003, 2004) proposed methods using higher moments. A relevant method is principal Hessian directions (pHd) (Li (1992)). Another related method is alternating conditional expectation (ACE), developed by Breiman and Friedman (1985), when we want to simultaneously search a transformation of $\boldsymbol{Y}$ and the transformations of components of the covariates $\boldsymbol{X}$. The iterative algorithm for ACE uses the notion of inverse regression. It can also be used to handle time series models where the transformations are involved in the covariates, see Xia, Li, Tong and Zhu (2000). However, when the central dimension reduction space that is spanned by $\boldsymbol{B}$ is our target, the iterative algorithm is not necessary because we do not transform $\boldsymbol{X}$. Doksum (1987) also investigated estimation for the transformation model where an unknown increasing transformation of the response follows a linear model with p covariates.

For the multivariate response case, that is, when $q > 1$, Li, Aragon, Shedden and Agnan (2003) searched for the most predictable variate of $\boldsymbol{Y}$, which is a convex combination of Y_i, $i = 1, \ldots, q$, and then reduced the dimension of $\boldsymbol{X}$. Cook and Setodji (2003) used a multivariate version of ordinary least squares, Yin and Bura (2006) extended the covariance-based method developed by Yin and Cook (2002) for univariate response to multivariate data. Yin and Zhu (2004) introduced the notion of dual central space to reduce both the dimension of $\boldsymbol{X}$ and of $\boldsymbol{Y}$. Another relevant work is Bura and Cook (2001).

In this article, we focus on modelling multivariate responses against one projected covariate $\boldsymbol{\beta}^T\boldsymbol{X}$. Clearly, the first problem is how to transform the multivariate response Y. Unlike Li, Aragon, Shedden and Agnan (2003), we consider a two-step algorithm to build up the model. The idea is as follows. We first transform every component of $\boldsymbol{Y}$ and then we apply partial least squares (PLS) regression to tackle the collinearity problem, see below for species. Since selecting the transformation does not involve the coefficients $\boldsymbol{\beta}$, PLS can then be employed to build a linear model between $\boldsymbol{H}(\boldsymbol{Y}) = (h_1(y_1), \ldots, h_q(y_q))^T$ and $\boldsymbol{X}$ with conditional mean $E(\boldsymbol{H}(\boldsymbol{Y})|\boldsymbol{X}) = \boldsymbol{\alpha} + \boldsymbol{\beta}^T\boldsymbol{X}$.

Since the transformations h_i's are nonparametric functions, smoothing methods should be used to estimate these functions. Here we adopt splines for ease of computational burden. In the general case, the B-splines are used. When the transformations are monotonic, I-splines with penalty are recommended. The consistency of the spline estimators will be proved; the proof of consistency and convergence rate of spline estimators is somewhat different from the existing approaches because of the use of a different criterion for selecting them.

The paper is organized as follows: In Section 2, we suggest a criterion to search for transformation of the responses. The methodology of PLS after transformation is also described in Section 3. In Section 4, we propose a simple algorithm for obtaining B-splines approximations. When the transformations are monotonic, we propose instead to use I-splines with a penalty function. Although we cannot have an algorithm as simple as that for B-splines, use of I-splines with our proposed penalty can ease the computational burden. For selecting the number of knots, we suggest a BIC type algorithm in Section 5. In Section 6, some simulations for I-spline transformations are carried out, and application to our data set is carried out using B-splines. In Section 7 we discuss a possible extension of our approach to modelling with large dimensional covariates. The technical proofs are given in the Appendix.

2. PLS Regression after Transformation

The basic idea of PLS with multiple transformed response (referred as TPLS hereafter) is as follows. Suppose that transformations of the Y_i's, $\boldsymbol{H}(\boldsymbol{Y}) = (h_1(Y_1), \ldots, h_q(Y_q))^T$, have already been obtained. Our target is to model $\boldsymbol{H}(\boldsymbol{Y})$ against $\boldsymbol{X} = (X_1, \ldots, X_p)^T$ as

$$\boldsymbol{H}(\boldsymbol{Y}) = r_0 + t_1 r_1 + \cdots + t_m \boldsymbol{r}_m + \boldsymbol{F}_m, \tag{2.1}$$

where $\boldsymbol{r}_0$ is the intercept, $t_1, \ldots, t_m$ are linear combinations $\boldsymbol{\beta}_i^T \boldsymbol{X}$, $i = 1, \ldots, m$, and $\boldsymbol{F}_m$ are the residuals. The components t_i are selected by the following procedure, see Tenehaus (1998) for example.

Denote the first component t_1 by $\boldsymbol{\beta}_1^T \boldsymbol{X}$. This t_1 maximizes the covariance of $\boldsymbol{w}^T \boldsymbol{H}(\boldsymbol{Y})$ and $\boldsymbol{X}^T \boldsymbol{\beta}$ over all $\boldsymbol{w}$, $\boldsymbol{\beta}$ with $||\boldsymbol{w}|| = 1$, $||\boldsymbol{\beta}|| = 1$, that is,

$$\mathrm{Cov}\,(\boldsymbol{w}_1^T \boldsymbol{H}(\boldsymbol{Y}), \boldsymbol{\beta}_1^T \boldsymbol{X}) = \max_{\boldsymbol{w}, \boldsymbol{\beta}} \mathrm{Cov}\,(\boldsymbol{w}^T \boldsymbol{H}(\boldsymbol{Y}), \boldsymbol{\beta}^T \boldsymbol{X}). \tag{2.2}$$

Here $\boldsymbol{\beta}_1$ and $\boldsymbol{w}_1$ are, respectively, the eigenvectors associated with the largest eigenvalues of the matrices $\left[\mathrm{Cov}\,(\boldsymbol{X}, \boldsymbol{H}(\boldsymbol{Y})^T)\right] \times \left[\mathrm{Cov}\,(\boldsymbol{H}(\boldsymbol{Y}), \boldsymbol{X}^T)\right]$ and $\left[\mathrm{Cov}\,(\boldsymbol{H}(\boldsymbol{Y}), \boldsymbol{X}^T)\right] \times \left[\mathrm{Cov}\,(\boldsymbol{X}, \boldsymbol{H}(\boldsymbol{Y})^T)\right]$. In other words, we take the first pair of canonical variables $(\boldsymbol{w}_1^T \boldsymbol{H}(\boldsymbol{Y}), \boldsymbol{\beta}_1^T \boldsymbol{X})$ for determining the direction of $\boldsymbol{\beta}_1$, with $||\boldsymbol{\beta}_1|| = 1$. Regressing $\boldsymbol{H}(\boldsymbol{Y})$ and $\boldsymbol{X}$ on t_1, linearly, we obtain

$$\boldsymbol{H}(\boldsymbol{Y}) = \boldsymbol{r}_0 + t_1 \boldsymbol{r}_1 + \boldsymbol{F}_1 \quad \text{and} \quad \boldsymbol{X} = \boldsymbol{p}_0 + t_1 \boldsymbol{p}_1 + \boldsymbol{G}_1, \tag{2.3}$$

where $\boldsymbol{r}_0$ and $\boldsymbol{p}_0$ are intercepts, $\boldsymbol{r}_1$ has the same dimension as does $\boldsymbol{H}(\boldsymbol{Y})$, and $\boldsymbol{p}_1$ that of $\boldsymbol{X}$. Recall that $t_1 \boldsymbol{r}_1$ is similar to an ordinary least squares approximation to $\boldsymbol{H}(\boldsymbol{Y})$ and, when $q = 1$, $\boldsymbol{H}(\boldsymbol{Y}) = t_1 \boldsymbol{r}_1 + \boldsymbol{F}_1$ is almost equivalent to the transformation of the response of He and Shen (1997). The case $q \geq 1$ is a generalized version studied by Fung, He, Liu, and Shi (2002). To extract the information of $\boldsymbol{G}_1$ contained in $\boldsymbol{F}_1$ to the regression part, we further consider the least squares regression of the two residuals $\boldsymbol{F}_1$ and $\boldsymbol{G}_1$ on a second component, as in the following algorithm. The second pair of canonical variables $\boldsymbol{F}_1$ and $t_2 = \boldsymbol{\beta}_2^T \boldsymbol{G}_1$ are obtained by maximizing the covariance $\mathrm{Cov}\,(\boldsymbol{w}_2^T \boldsymbol{F}_1, \boldsymbol{\beta}^T \boldsymbol{G}_1)$ over all $\boldsymbol{w}_2$ and $\boldsymbol{\beta}$, that satisfy $\|\boldsymbol{\beta}\| = \|w\| = 1$, $w \perp\!\!\!\perp w_1$. Regressing $\boldsymbol{F}_1$ and $\boldsymbol{G}_1$ on t_2, we then derive that $\boldsymbol{F}_1 = t_2 \boldsymbol{r}_2 + \boldsymbol{F}_2$ and $\boldsymbol{G}_1 = t_2 \boldsymbol{p}_2 + \boldsymbol{G}_2$ where $\boldsymbol{F}_2$ and $\boldsymbol{G}_2$ are the residuals. This yields

$$\boldsymbol{H}(\boldsymbol{Y}) = \boldsymbol{r}_0 + t_1 \boldsymbol{r}_1 + t_2 \boldsymbol{r}_2 + \boldsymbol{F}_2 \quad \text{and} \quad \boldsymbol{X} = \boldsymbol{p}_0 + t_1 \boldsymbol{p}_1^T + t_2 \boldsymbol{p}_2^T + \boldsymbol{G}_2. \tag{2.4}$$

Performing this procedure m times, say until the predetermined accuracy is achieved in terms of a selection criterion based on generalized cross-validation (GCV) (see Györfi, Kohler, Krzyźak and Walk (2002)), we reach (2.1).

Note that all the t_i's are linear combinations of the components of $\boldsymbol{X}$. Using the iterative equation (2.4), we have $t_h = \beta_h^T \prod_{j=1}^{h-1}(I_p - \beta_j \boldsymbol{p}_j^T)\boldsymbol{X}$, where $\prod$ is the product operator and I_p is the $p \times p$ identity matrix. This $t_1\boldsymbol{r}_1 + \cdots + t_m\boldsymbol{r}_m$ can be rewritten as $\boldsymbol{\beta}^T\boldsymbol{X}$, with

$$\boldsymbol{H}(\boldsymbol{Y}) = \alpha + \boldsymbol{\beta}^T\boldsymbol{X} + \boldsymbol{F}_m. \tag{2.5}$$

This can be viewed as a transformed single-index model when the response is multi-dimensional.

Once the sample $\{(\boldsymbol{x}_i, \boldsymbol{y}_i), i = 1, \ldots, n\}$ is available, some data-driven algorithm should be used to determine the number, m, of components. Generalized cross validation (GCV) has been widely adopted, see Tenehaus (1998, p.2) or the SAS software, Version 6.11. Thus, let $(\boldsymbol{x}^{-i}, \boldsymbol{y}^{-i})$ be of size $n-1$ with the ith observation $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ removed from $\{(\boldsymbol{x}_i, \boldsymbol{y}_i),\ i = 1, \ldots, n\}$. Further, let $\hat{\boldsymbol{H}}_i(\cdot)$ and $\hat{\boldsymbol{\beta}}_{m,i}$ be, respectively, the estimators of $\boldsymbol{H}$ and $\boldsymbol{\beta}_m$, based on $(\boldsymbol{x}^{-i}, \boldsymbol{y}^{-i})$ with m components, and $\hat{\boldsymbol{H}}(\cdot)$ and $\hat{\boldsymbol{\beta}}_m$ be, respectively, the estimators of $\boldsymbol{H}$ and $\boldsymbol{\beta}_m$ based on the whole data set. Define $PRESS_m = (1/n)\sum_{i=1}^n(\hat{\boldsymbol{H}}_m(\boldsymbol{y}_i) - \hat{\boldsymbol{\beta}}_{m,i}^T\boldsymbol{x}_i)^T(\hat{\boldsymbol{H}}_m(\boldsymbol{y}_i) - \hat{\boldsymbol{\beta}}_{m,i}^T\boldsymbol{x}_i)$ and $SS_m = (1/n)\sum_{i=1}^n(\hat{\boldsymbol{H}}(\boldsymbol{y}_i) - \hat{\boldsymbol{\beta}}_{m,i}^T\boldsymbol{x}_i)^T(\hat{\boldsymbol{H}}(\boldsymbol{y}_i) - \hat{\boldsymbol{\beta}}_{m,i}^T\boldsymbol{x}_i)$, with

$$GCV_m = 1 - \frac{PRESS_m}{SS_m}. \tag{2.6}$$

We choose m from 1 to p until $GCV_m \geq 1 - 0.95^2 = 0.0975$.

3. Selection of Transformations

Consider $q = 1$. Clearly, by SIR or the method of He and Shen (1997), the transformation H is only related to $\boldsymbol{X}$ in the direction $\boldsymbol{\beta}_0$. This is similar to TPLS with one component when the transformation is obtained by least squares. However, SIR or He and Shen's method cannot be extended to search for $\boldsymbol{H}$ when the transformed response vector $\boldsymbol{H}(\boldsymbol{Y})$ is related to more than one component $\boldsymbol{\beta}_i^T\boldsymbol{X}, i = 1, \ldots, m$. This then causes the difficulty that the PLS algorithm cannot be used to obtain more than one component. Actually, all existing methods including inverse regression and ACE suffer from this difficulty. To overcome the problem, one can search for a transformation $\boldsymbol{H}(\boldsymbol{Y})$ that is related to all covariables X_i of $\boldsymbol{X} = (X_1, \ldots, X_p)^T$ and having highest correlation in some sense, allowing the algorithm in Section 2 to be performed without difficulty. Based on this consideration, we define a criterion to maximize the square of the covariance between $\boldsymbol{H}(\boldsymbol{Y})$ and all of covariables X_i of $\boldsymbol{X} = (X_1, \ldots, X_p)^T$

over a class of smooth functions. When $q > 1$, for $\boldsymbol{Y} = (Y_1, \ldots, Y_q)^T$, we need to search for q transformations to form a transformation vector $\boldsymbol{H}(\boldsymbol{Y}) = (h_1(Y_1), \ldots, h_q(Y_q))^T$ for smooth functions h_i, $i = 1, \ldots, q$. We use the sum of covariances between h_i and X_l, $1 \le i \le q, 1 \le l \le p$, to measure the correlation between $\boldsymbol{H}(\cdot)$ and $\boldsymbol{X}$. $\boldsymbol{H}_0(\cdot) = (h_{1,0}(\cdot), \ldots, h_{q,0}(\cdot))^T$ is defined as the maximizer of

$$\mathbf{C}(\boldsymbol{H}) = \sum_{i=1}^{q} \sum_{l=1}^{p} \left[\operatorname{Cov}(h_i(Y_i), X_l) \right]^2 = \sum_{i=1}^{q} \left[\operatorname{Cov}(h_i(Y_i), \boldsymbol{X}^T) \right] \left[\operatorname{Cov}(\boldsymbol{X}, h_i(Y_i)) \right] \tag{3.1}$$

over all $\boldsymbol{H}(\cdot) \in \mathcal{H}^q := \mathcal{H} \times \mathcal{H} \cdots \times \mathcal{H}$, where $\mathcal{H}$ is the class of continuous squared integrable functions whose variance is one.

We clarify two points. First note that, because $\mathcal{H}^q$ is a product space, the maximizer $\boldsymbol{H}_0$ of $\mathbf{C}(\boldsymbol{H})$ over $\mathcal{H}^q$ is the vector of transformations $\boldsymbol{H}_0(\cdot) = (h_{1,0}(\cdot), \ldots, h_{q,0}(\cdot))^T$, each being the maximizer $h_{i,0}(\cdot)$ of the covariance between $h_i(Y_i)$ and $\boldsymbol{X}$ over $\mathcal{H}$. In other words, the maximization for this sum is equivalent to the maximization of each term in the sum. For notational convenience, we use the sum as the criterion. Second, we standardize $h_i(Y_i)$ because the use of a non-standardized version artificially inflates $C(H)$. In contrast, $\boldsymbol{X}$ should not be standardized since we want to deal with possible collinearity in $\boldsymbol{X}$ by the partial least squares method.

Unlike SIR, the maximizer $\boldsymbol{H}(\cdot)$ does not have a closed form. To ease the computational burden, we define a spline estimator that has a closed parametric form and is easy to implement.

4. Estimation of The Transformations

4.1. B-splines transformation

Let $\boldsymbol{\theta}_i = (\theta_{i,0}, \theta_{i,1}, \ldots, \theta_{i,J+1})^T$, $\boldsymbol{\pi}(y_i) = (B_0(y_i), \ldots, B_{J+1}(y_i))^T$, be a vector of B-splines basis functions, and $\boldsymbol{H}(\boldsymbol{Y}) = (h_1(y_1), \ldots, h_q(y_q))$ with the ith element $h_i(y_i) = \boldsymbol{\pi}(y_i)^T \theta_i$ an unknown projection vector. The B-splines approximation "linearizes" the smooth function $H(\boldsymbol{Y})$ in (3.1) by $\boldsymbol{\pi}(y_i)^T \theta_i$ for every component. We assume that the range of Y is contained in the q-dimensional cube $[a, b]^q$. In practice, a and b can be, respectively, the minimum and the maximum of the data. See Schumaker (1981, p.124) for more details about B-splines approximation. For each i with $1 \le i \le q$, given a partition $a = r_0 < r_1 < r_2 < r_3 < \cdots < r_J < r_{J+1} = b$, we write $B_i(y) = S_3((y - r_i)J/(b - a))$ $(i = 0, \ldots, J + 1)$, where

$$S_3(z) = 0, \quad |z| \ge 2; \quad S_3(z) = \frac{|z|^3}{2} - z^2 + \frac{2}{3}, \quad |z| \le 1;$$

$$S_3(z) = \frac{-|z|^3}{6} + z^2 - 2|z| + \frac{4}{3}, \quad 1 < |z| < 2.$$

Throughout this article, we write the (j/J)th quantile of the observed response as r_j, $j = 1, \ldots, J$. Denote by $\mathcal{H}_{bs} = \{h(\cdot) : \theta \in R^{J+2}\}$ the space of all splines functions on $[a, b]$, the first order derivatives of which are Lipschitz. It is well known that, as J goes to infinity, any function $h \in \mathcal{H}$ can be uniformly approximated by its projection in the B-splines space $\mathcal{H}_{bs}$ (Schumaker (1981)). Obviously, any function $H(\cdot) \in \mathcal{H}^q$ can be uniformly approximated by its projection in the product of the B-splines $H(\cdot) \in \mathcal{H}^q_{bs}$, $\mathcal{H}^q_{bs} := \mathcal{H}_{bs} \times \mathcal{H}_{bs} \cdots \times \mathcal{H}_{bs}$, whose members have variance one. Thus we can derive the convergence of

$$\max_{\boldsymbol{H} \in \mathcal{H}^q_{bs}} \mathbf{C}_{bs}(H) := \max_{\boldsymbol{H} \in \mathcal{H}^q_{bs}} \sum_{i=1}^{q} \sum_{l=1}^{p} \big[\, \mathrm{Cov}\,(h_i(Y_i), X_l)\big]^2 \tag{4.1}$$

to $\max_{\boldsymbol{H} \in \mathcal{H}^q} \mathbf{C}(\boldsymbol{H})$, see (3.1).

When a data set $\{(\boldsymbol{x}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{x}_n, \boldsymbol{y}_n)\}$ is available, for any $\boldsymbol{H} \in \mathcal{H}^q_{bs}$ the estimator of $\mathbf{C}_{bs}(\boldsymbol{H})$ can be defined as the sum of the sample covariances

$$\hat{\mathbf{C}}_{bs}(\boldsymbol{H}) = \sum_{i=1}^{q} \sum_{l=1}^{p} [\widehat{\mathrm{Cov}}\,(h_i(Y_i), X_l)]^2, \tag{4.2}$$

where $\widehat{\mathrm{Cov}}\,(h_i(Y_i), X_l)$ stands for the sample covariance between $\{h_i(y_{i1}), \ldots, h_i(y_{in})\}$ and $\{x_{l1}, \ldots, x_{ln}\}$. The estimated B-spline $\hat{\boldsymbol{\theta}}^T(\boldsymbol{\pi}_1(\cdot), \ldots, \boldsymbol{\pi}_q(\cdot))$ is defined as a maximizer of $\hat{\mathbf{C}}_{bs}(\boldsymbol{H})$ over $\boldsymbol{H} \in \mathcal{H}^q_{bs}$. This estimator has a closed form as shown in the following theorem.

Theorem 1. *Assume that the covariances between the components of $\boldsymbol{\pi}(\cdot)$ and of $\boldsymbol{X}$ are finite. Then, $h_i(\cdot) = \hat{\boldsymbol{\theta}}_i^T \boldsymbol{\pi}(\cdot)$, $i = 1, \ldots, q$, where*

$$\hat{\boldsymbol{\theta}}_i = \frac{\hat{\boldsymbol{\eta}}_i^T [\widehat{\mathrm{Cov}}\,(\boldsymbol{\pi}(y_i))]^{\frac{1}{2}}}{[|\hat{\boldsymbol{\eta}}_i^T \widehat{\mathrm{Cov}}\,(\boldsymbol{\pi}(y_i))\hat{\boldsymbol{\eta}}_i|]^{\frac{1}{2}}}$$

and $\hat{\boldsymbol{\eta}}_i$ is the eigenvector that is associated with the largest eigenvalue of the matrix $(\mathrm{Cov}\,(\boldsymbol{\pi}(Y_i), \boldsymbol{\pi}(Y_i)^T))^{-1}\big[\,\mathrm{Cov}\,(\boldsymbol{\pi}(Y_i), X^T)\big]\big[\big[\,\mathrm{Cov}\,(\boldsymbol{X}, \boldsymbol{\pi}(Y_i)^T)\big]$.

The following theorem states the convergence of the criterion and the estimator.

Theorem 2. *Assume that $J^{9/4}(pq)^{3/2} = o(\sqrt{n})$ and the fourth moments of $\boldsymbol{X}$ is finite. Then*

$$\max_{\boldsymbol{H} \in \mathcal{H}^q_{bs}} \hat{\mathbf{C}}_{bs}(\boldsymbol{H}) - \max_{\boldsymbol{H} \in \mathcal{H}^q_{bs}} \mathbf{C}_{bs}(\boldsymbol{H}) = O_p\Big(\frac{J^{\frac{9}{4}}(pq)^{\frac{3}{2}}}{\sqrt{n}}\Big) \tag{4.3}$$

and, when the maximizer of $\mathbf{C}_{bs}(\boldsymbol{H})$ is unique, the maximizer of $\hat{\mathbf{C}}_{bs}(\boldsymbol{H})$ is convergent in probability to that of $\mathbf{C}_{bs}(\boldsymbol{H})$ when $(J^{9/4}(pq)^{3/2})/\sqrt{n} \to 0$ as $n \to \infty$. Then the B-splines estimator is convergent to the maximizer of $\mathbf{C}(\boldsymbol{H})$.

Remark 1. The convergence rate seems not to be optimal. However, in this article, the criterion of selecting splines transformation is different from others in the literature even when $q = 1$. This is also applied to the I-splines transformation in the next subsection. Hence, the existing results cannot be directly used to prove an optimal convergence rate. The optimal convergence rate of the splines estimators deserves a further study.

4.2. Monotonic I-splines Transformation

In some circumstances the transformation can be monotonic, see Cook and Weisberg (1994) for an example. In this case, we should restrict our attention to monotonic spline transformations for better approximation. Here, we introduce monotonic I-splines (Xia et al. (2000)) with a penalty function. The algorithm we suggest is a modification of that used in Ramsay (1988). Consider the population version first; when we have a data set, the corresponding formulae can be replaced by the sample version. Consider the I-splines of order 2 based on the knots mesh $\{r_j\}$ with $a = r_0 < r_1 < \cdots < r_J < r_{J+1} = b$. For any i with $1 \le i \le q$, the basis function $\boldsymbol{\pi}(y_i) = (B_0, \ldots, B_{J+1})^T$ is defined through B_k as

$$\begin{aligned}
B_1(y_i) &= \frac{(y_i - r_0)^2}{(r_1 - r_0)^2} I(r_0 \le y_i \le r_1) + I(y_i > r_1),\\
B_k(y_i) &= \frac{(y_i - r_{k-2})^2}{(r_{k-1} - r_{k-2})(r_k - r_{k-2})} I(r_{k-2} \le y_i \le r_{k-1})\\
&\quad + \Big[1 - \frac{(y_i - r_k)^2}{(r_k - r_{k-1})(r_k - r_{k-2})}\Big] I(r_{k-1} \le y_i \le r_k) + I(y_i \ge r_k),\\
B_{J+1}(y_i) &= \frac{(y_i - r_{J-1})^2}{(r_J - r_{J-1})^2} I(r_{J-1} \le y_i \le r_J).
\end{aligned}$$

Let $B_0(y_i) \equiv 1$ and $h_i(y_i) = \boldsymbol{\theta}_i^T \boldsymbol{\pi}(y_i)$. First we consider the transformation for each component y_i, $1 \le i \le q$. Because we want a monotonic transformation, we should consider the maximizer subject to monotonicity on h_i. Monotonicity can be ensured by $\theta_{i,j} \ge 0$ for all $j \ge 1$ (see Ramsay (1988) for details). If we consider the algorithm with this constraint, then solving for $\boldsymbol{\theta}_i$ is not as simple as it was with B-splines. However, the maximization problem is a quadratic problem and can be solved without much difficulty. Note that such a transformation may not be strictly monotonic because some optimal values of the components of $\boldsymbol{\theta}$ may be zero. It is clear that we can restrict the first derivative to be bounded away from 0. To achieve this, the following criterion with a penalty function can be used. For each i with $1 \le i \le q$, let

$$C_{\alpha s,i}(h_i) := \big[\mathrm{Cov}\,(h_i(Y_i), \boldsymbol{X}^T)\big]\big[\mathrm{Cov}\,(\boldsymbol{X}, h_i(Y_i))\big] - \alpha \sum_{j=0}^{J+1} \log\{1 + \theta_{i,j}\}.$$

The criterion is defined as

$$C_{\alpha s}(\boldsymbol{H}) = \sum_{i=1}^{q} C_{\alpha s,i}(h_i), \tag{4.4}$$

and the $(J+2) \times q$ matrix $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_q)$ is the maximizer of $C_{\alpha s}(\boldsymbol{H})$ over all $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_q)$ with $||\boldsymbol{\theta}_i|| = 1$, $i = 1, \ldots, q$. Each $\boldsymbol{\theta}_i$ is the maximizer of the corresponding $C_{\alpha s,i}(h_i)$. Once we have data $\{(\boldsymbol{x}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{x}_n, \boldsymbol{y}_n)\}$, the estimators of $C_{\alpha s,i}(h_i)$ and $C_{\alpha s}(\boldsymbol{H})$ are separately defined by

$$\hat{C}_{\alpha s,i}(h_i) := \big[\widehat{\mathrm{Cov}}\,(h_i(Y_i), \boldsymbol{X}^T)\big]\big[\widehat{\mathrm{Cov}}\,(\boldsymbol{X}, h_i(Y_i))\big] - \alpha \sum_{j=0}^{J+1} \log\{1 + \theta_{i,j}\},$$

$$\hat{C}_{\alpha s}(\boldsymbol{H}) = \sum_{i=1}^{q} \hat{C}_{\alpha s,i}(h_i). \tag{4.5}$$

The estimator $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_q)$ is the maximizer of $\hat{C}_{\alpha s}(\boldsymbol{H})$ over all $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_q)$. Clearly, each $\hat{\boldsymbol{\theta}}_i$ is the maximizer of the corresponding $\hat{C}_{\alpha s,i}(h_i)$.

Adding the penalty function can ease the computational burden of the quadratic problem. We have the following result in this direction.

For each i, note that

$$\frac{d\ C_{\alpha s,i}(\boldsymbol{H}(\boldsymbol{Y}))}{d\ \boldsymbol{\theta}_i} = 2\big[\widehat{\mathrm{Cov}}\,(\boldsymbol{\pi}(Y_i), \boldsymbol{X}^T)\big]\big[\widehat{\mathrm{Cov}}\,(\boldsymbol{X}, \boldsymbol{\pi}(Y_i)^T)\big]\boldsymbol{\theta}_i - \alpha \frac{1}{1 + \boldsymbol{\theta}_i},$$

where $1/(1 + \boldsymbol{\theta}_i) = (1/(1 + \theta_{i,0}), \ldots, 1/(1 + \theta_{i,J+1}))^T$. From this, setting to 0, we can derive the solution of θ_i. Specifically, for any $i = 1, \ldots, q$, let $\boldsymbol{A}_i = \big[\widehat{\mathrm{Cov}}\,(\boldsymbol{\pi}(Y_i), \boldsymbol{X}^T)\big]\big[\widehat{\mathrm{Cov}}\,(\boldsymbol{X}, \boldsymbol{\pi}(Y_i)^T)\big]$ and a_{lm} be the elements of $\boldsymbol{A}_i$. For each l with $0 \le l \le J+1$, $\theta_{i,l}$ is the solution of the following equation:

$$a_{ll}\theta_{i,l}(1 + \theta_{i,l}) + \Big[\sum_{k \neq l} a_{lk}\theta_{i,k}\Big](1 + \theta_{i,l}) - \frac{\alpha}{2} = 0,$$

where $a_{ll} > 0$. Note that the solution of every component $\theta_{i,l}$ is related to all other components $\theta_{i,k}$, $0 \le k \le J+1$. We obtain the final solutions by an iterative algorithm based on the above framework. The following theorem states the convergence of the algorithm.

Theorem 3. *Suppose that all marginal density functions of* X_l, $l = 1, \ldots, p$, *are bounded and positive on [a, b]. Choosing* $\alpha < 2(\lambda_{\min}(\boldsymbol{A}_i\boldsymbol{A}_i))^{1/2}$, *the above algorithm converges in probability, where* $\lambda_{\min}(\boldsymbol{A}_i\boldsymbol{A}_i)$ *is the smallest eigenvalue of the matrix* $\boldsymbol{A}_i\boldsymbol{A}_i$.

Similar to Theorem 2, we can also obtain the convergence of $\max_{\boldsymbol{H} \in \mathcal{H}_{bs}^q} \hat{C}_{\alpha s}(\boldsymbol{H})$ to $\max_{\boldsymbol{H} \in \mathcal{H}_{bs}^q} C_{\alpha s}(\boldsymbol{H})$. The result is as follows.

Theorem 4. *Assume that $J^{9/4}(pq)^{3/2} = o(\sqrt{n})$ and the fourth moments of X is finite. Then*

$$\max_{\boldsymbol{H}\in\mathcal{H}^q_{bs}} \hat{\mathbf{C}}_{\alpha s}(\boldsymbol{H}) - \max_{\boldsymbol{H}\in\mathcal{H}^q_{bs}} \mathbf{C}_{\alpha s}(\boldsymbol{H}) = O_p\Big(\frac{J^{\frac{9}{4}}(pq)^{\frac{3}{2}}}{\sqrt{n}}\Big), \tag{4.6}$$

and the maximizer of $\hat{\mathbf{C}}_{\alpha s}(\boldsymbol{H})$ is convergent in probability to that of $\mathbf{C}_{\alpha s}(\boldsymbol{H})$ when $(J^{9/4}(pq)^{3/2})/\sqrt{n} \to 0$, as $n \to \infty$, if the maximizer of $\mathbf{C}_{\alpha s}(\boldsymbol{H})$ is unique.

5. The Determination of Knots

To perform the above transformations, one must decide on the number of knots J. Because J should depend on the sample size n, we will write $J = k_n$. As a large number of knots will cause under-smoothing, the choice of k_n can naturally be regarded as a model selection problem. We write the transformed response $\hat{\boldsymbol{\theta}}^T\boldsymbol{\pi}(\cdot)$ as $\hat{\boldsymbol{H}}_{k_n}(\cdot)$ to show its dependence on k_n. Let $BIC_m(k_n, \boldsymbol{H})$ be the value of a modified Bayesian Information Criterion (BIC)(Schwarz (1978)) when we use PLS with the transformation $\boldsymbol{H}(\cdot)$, i.e.,

$$BIC_m(k_n, \boldsymbol{H}) = \log(\hat{\sigma}^2(k_n)) + (k_n + p + q + 1)\frac{\max\{\log n, 3\}}{n}, \tag{5.1}$$

where $\hat{\sigma}^2(k_n)$ is the sum of the squares of the residuals $\boldsymbol{F}_m(\boldsymbol{x}_j, \boldsymbol{y}_j)$ that are obtained in (2.1). We choose k_n to be any integer such that $BIC_m(k_n, \boldsymbol{H})$ is minimized. In some sense we have compromised between 2, used in AIC, and $\log n$, used in BIC, to ensure consistency. The rationale is to balance fidelity to data with the complexity of the model. The value 3 is a small sample adjustment and reflects our experience with small to modest data sets. Other model selection criterion could be used. For example, the modified Akaike type information criteria proposed by Fujikoshi and Satoh (1997) is an alternative. McQuarrie and Trai (1998) provides a comprehensive discussion. As He and Shen (1997) pointed out, no single method is the best for all problems, and further research is clearly needed to understand the pros and cons of various possible knot selection rules. However, that issue is beyond the scope of the present work.

6. Simulations and An Application

6.1. Simulations for I-splines transformation

In this subsection, we illustrate the performance of I-splines transformation through simulation.

The model is of a 4-dimensional response. For $i = 1, \ldots, n$,

$$\begin{aligned}
y_{1,i} &= \prod_{\boldsymbol{x},i} + \sigma\epsilon_{1i}, \\
y_{2,i} &= (\Pi_{x,i} + \sigma\epsilon_{2i})^3, \\
y_{3,i} &= \arctan(\prod_{\boldsymbol{x},i} + \sigma\epsilon_{3i}), \\
y_{4,i} &= e^{\prod_{\boldsymbol{x},i} + \sigma\epsilon_{4i}}. \qquad (6.1)
\end{aligned}$$

Here $\prod_{\boldsymbol{x},i} = \prod_{j=1}^{10} x_{ij}$, with $\boldsymbol{x}_i = (x_{1,i}, \ldots, x_{10,i})^T$ drawn from normal $N(0, I_{10})$, $\epsilon_i = (\epsilon_{1,i}, \ldots, \epsilon_{4,i})^T$ being standard normal $N(0, I_4)$. Moreover, $\boldsymbol{x}_i$ is to be independent of ϵ_i. Clearly, this model does not have a single-index, we use it to examine the performance of our modelling. To examine the impact of variance σ^2 on the estimation, we choose $\sigma = 0.5, 1, 2, 4$, and take the sample size at 100. In Figure 1, we plot the fitted model $\boldsymbol{\beta}_m^T X$ against $\boldsymbol{Y}$, and the estimated transform $\boldsymbol{H}(\boldsymbol{Y})$ componentwise with $\sigma = 0.5$, where $m = 2$ is selected by GCV in (2.6). From Figure 1, we can see that even when the model is not single-indexed, the fit is still encouraging.

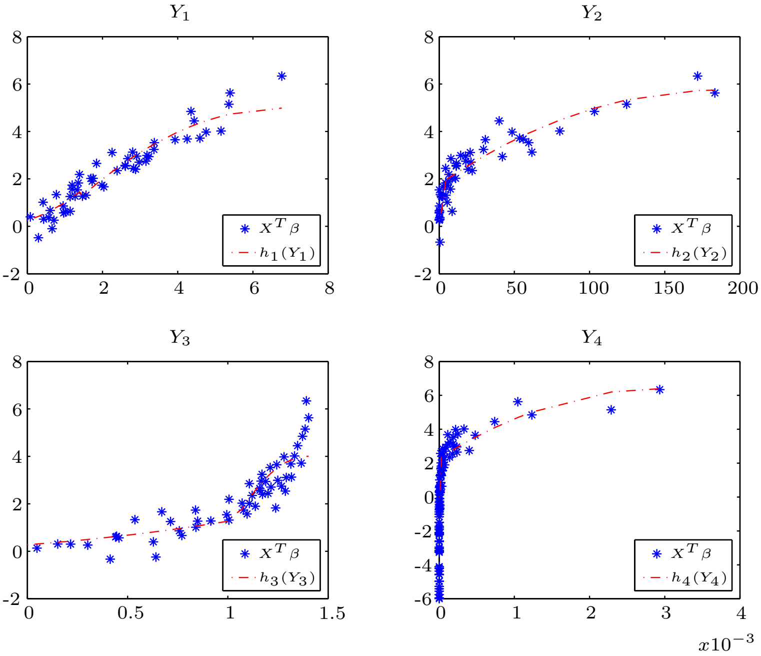

Figure 1. For the model at (6.1), the SE are scatter plots between the I-spline transformed responses $\boldsymbol{H}(\boldsymbol{Y})$ on the vertical axis and the corresponding $\boldsymbol{Y}$ on the horizontal axis.

We also conducted a comparison between PLS and TPLS. The sample size was 100. To give the squared multiple correlation coefficients between the transformed response (TPLS) and the related fitted response and those between the response (PLS) and the related fitted response, we used 100 replicates.

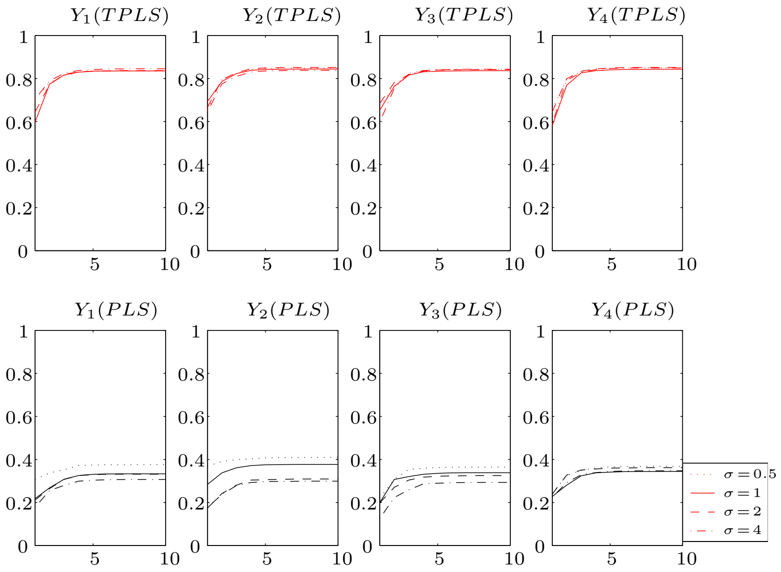

Figure 2. For the model at (6.1), the $Y_i(TPLS)$ plots give the squared multiple correlation coefficients between the I-spline transformed responses $\boldsymbol{H}(\boldsymbol{Y})$ and their corresponding fitted responses against the number of components from 1 to 10; the $Y_i(PLS)$ plots give the mean of 100 simulated squared multiple correlation coefficients between Y and the fitted responses against the number of components.

From Figure 2, we can clearly see the necessity of using transformation to establish a linear model. Because the underlying model does not have an index parameter, a simple one-component model without transformed respones can only make the multivariate correlation coefficients around 0.2, and when transformations are applied these values are around 0.6. When two components are included, TPLS achieves coefficients are around 0.8, but PLS does not work.

6.2. Crop Yields and Agricultural Meteorology

6.2.1. Data Description

We return to the example discussed in the introduction. Due to the development of agricultural techniques and the use of fertilizers, crop yields do not rely only on the meteorological conditions. We consider using ‘residual’ yields, obtained from first fitting with meteorology conditions as the covariates. Let $\boldsymbol{Y} = (Y^s, Y^r, Y^m)$ be the 'residual' yields of soybean, rice, and maize. $\boldsymbol{X} = (X_1, \ldots, X_{17})^T$ is a set of measurements of climate changes to be specified below. The data from 1958 to 1990 were collected by the Institute of Meteorology in Ji-Lin Province. Here $j = 1$ is for the 1958 data and $j = 33$ for the 1990 data, a sample size of $n = 33$. For each $1 \leq j \leq 33$, measurements x_{ij} were collected according to the growth periods of the crops. For each year, in the period from the 11th of May to the 20th of June, $X_{1,j}(t5{-}6), X_{2,j}(lt5{-}6), X_{3,j}(r4{-}6)$ and $X_{4,j}\ (s5{-}6)$ are, respectively, the average temperature of daytime, the lowest temperature of daytime, the rain quantity (note: this variable is the rain quantity from the 11th of April to the 20th of June; see Ma (1996) about this), and the hours of sunshine. Similar measurements, in the period from the 1st of July to the 10th of August are, respectively, $X_{5,j}(t7-8), X_{6,j}(lt7-8), X_{7,j}(r7-8), X_{8,j}(s7-8)$; and the period from the 11th of August to the 10th of September, give us $X_{9,j}(t8-9) X_{10,j}(lt8-9), X_{11,j}(r8-9), X_{12,j}(s8-9)$. The measurements during the major growth period of rice from the 1st of May to the 30th of September are $X_{13,j}(t5-9), X_{14,j}(r5-9), X_{15,j}(s5-9)$. Two other measurements are also obtained; $X_{16,j}(Lastr9-10)$, the rain quantity from the 1st of September to the 31st of October in the previous year; $X_{17,j}$ (r All), the rain quantity in the full year.

6.2.2. Modelling and Analysis Based on TPLS

We consider modelling by TPLS here. The B-splines are used for this example because the "*residual*" yields do not show monotonicity when plotted against the covariates. Figure 3 also verifies this finding. The number of knots is five, determined by the modified BIC in Section 5. The components are also selected by the aforementioned GCV. The number is selected as three. When one component in TPLS is included, the approach is an extension of He and Shen's (1997) approach. The models with one and three components are reported as follows.

$$\hat{h}(y_j) = t_{1,j} r_1 \qquad \text{(I) with one latent variable}$$
$$\hat{h}(y_j) = t_{1,j} r_1 + t_{2,j} r_2 + t_{3,j} r_3 \qquad \text{(II) with three latent variables,}$$

where for each j, $t_{1,j}, t_{2,j}, t_{3,j}$ are also the latent variables.

The plot of $\boldsymbol{\beta}_3^T X$ against $\boldsymbol{Y}$ is presented in Figure 3, which also includes the fitted curve of transformed $\boldsymbol{Y}$, $\boldsymbol{H}(\cdot)$. The fit is satisfactory.

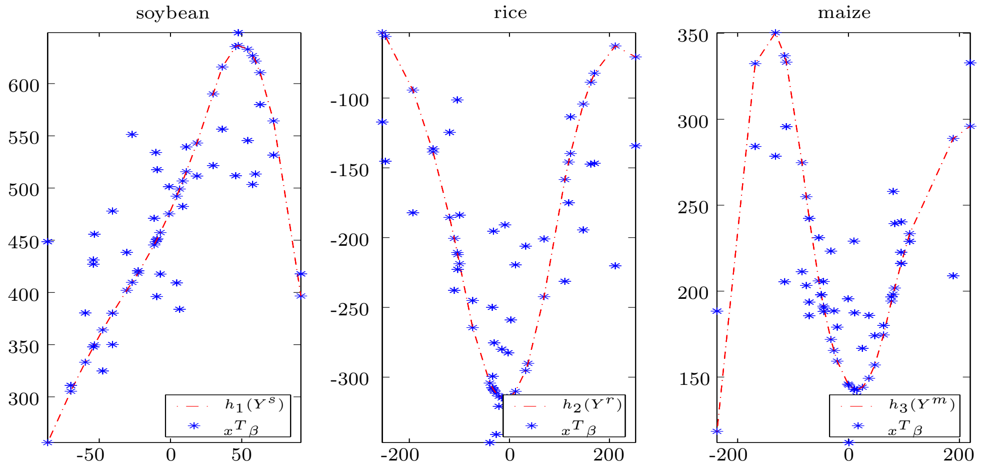

Figure 3. For the data set application, these plots give the scatter plots between the B-spline transformed responses $\boldsymbol{H}(\boldsymbol{Y})$ on the vertical axis and the corresponding $\boldsymbol{Y}$ on the horizontal axis.

To show the performance of TPLS, we compute the multiple correlation coefficients R between $h(Y)$ and X. Table 1 reports the values of R in all cases and the values of R for OLS.

Table 1. The multiple correlation coefficients for OLS and TPLS.

	Soybean	Rice	Maize
	R	R	R
Model (I)	0.59	0.72	0.67
Model (II)	0.78	0.82	0.78
OLS	0.48	0.47	0.33

From the results of Table 1, we see that a transformation produces multiple correlation coefficients that are much larger than those of OLS. PLS is also useful for building a linear model with the transformed responses. The three-component TPLS outperforms the one component TPLS that is an extension of He and Shen's approach to multivariate response data. This indicates that TPLS with a proper number of components can extract more information than a simple least squares linear modelling with transformed response.

7. Further Discussion

PLS can handle problems with high-dimensional response and covariates. Then it is of interest to study the case where the dimensions of both the response and covariates are large. Looking at the condition of Theorem 2 and Theorem 4,

$J^{9/4}(pq)^{3/2} = o(\sqrt{n})$, when the dimensions p and q tend to infinity at some proper rates, we still have the convergence of $\max_{H\in\mathcal{H}_{bs}^q} \hat{\mathbf{C}}_{bs}(H)$ to $\max_{H\in\mathcal{H}^q} \mathbf{C}(H)$ of (3.1). Therefore, our approach may be extended to handle the cases with high-dimensional covariates. This deserves further study.

Acknowledgements

The first manuscript was done when Lixing Zhu worked at the University of Hong Kong. The first author was supported by grants ($HKBU7181/02H$ and $HKBU7060/04P$) from the Research Grants Council of Hong Kong, Hong Kong, P. R. China. The second author was supported by a NSF grant from National Natural Science Foundation of China (No. 10701035), a special fund for young teachers in Shanghai universities (No. 79001320). The authors thank the Institute of Meteorology, Ji-Lin province, China for allowing the authors to use the data, and Dr. Zhang Baoxue for assistance with data analysis. Thanks also go to the Editor, an associate editor and two referees for their constructive comments and suggestions which led to a great improvement in the presentation.

Appendix

Proof of Theorem 1. For $h' \in \mathcal{H}_{bs}$, we take a corresponding squared integrable function h such that $h'(\cdot) = h(\cdot)/\sqrt{\mathrm{Var}\,(h)}$. Hence for any h'_i, we have

$$\widehat{\mathrm{Cov}}\,(h'_i(Y_i), X_l) = \frac{\widehat{\mathrm{Cov}}\,(h_i(Y_i), X_l)}{\sqrt{\mathrm{Var}\,(h_i(Y_i))}} = \frac{\boldsymbol{\theta}_i^T \widehat{\mathrm{Cov}}\,(\boldsymbol{\pi}(Y_i), X_l)}{\sqrt{\boldsymbol{\theta}_i^T \widehat{\mathrm{Cov}}\,(\boldsymbol{\pi}(Y_i))\boldsymbol{\theta}_i}}.$$

Let

$$\boldsymbol{\eta}_i^T = \frac{\boldsymbol{\theta}_i^T \widehat{\mathrm{Cov}}\,(\boldsymbol{\pi}(Y_i))^{\frac{1}{2}}}{\sqrt{\boldsymbol{\theta}_i^T \widehat{\mathrm{Cov}}\,(\boldsymbol{\pi}(Y_i))\hat{\boldsymbol{\theta}}_i}}.$$

We have

$$\begin{aligned}
&\max_{\boldsymbol{H}\in\mathcal{H}_{bs}} \sum_{l=1}^{p} [\widehat{\mathrm{Cov}}\,(h_i(Y_i), X_l)]^2 \\
&= \max_{||\boldsymbol{\theta}_i||=1} \frac{\boldsymbol{\theta}_i^T \widehat{\mathrm{Cov}}\,(\boldsymbol{\pi}(Y_i), \boldsymbol{X}^T)\widehat{\mathrm{Cov}}\,(\boldsymbol{X}, \boldsymbol{\pi}(Y_i))\boldsymbol{\theta}_i}{\boldsymbol{\theta}_i^T \widehat{\mathrm{Cov}}\,(\boldsymbol{\pi}(Y_i))\boldsymbol{\theta}_i} \\
&= \max_{||\boldsymbol{\eta}_i||=1} \boldsymbol{\eta}_i^T (\widehat{\mathrm{Cov}}\,(\boldsymbol{\pi}(Y_i)))^{-\frac{1}{2}} \widehat{\mathrm{Cov}}\,(\boldsymbol{\pi}(Y_i), \boldsymbol{X}^T)\widehat{\mathrm{Cov}}\,(\boldsymbol{X}, \boldsymbol{\pi}(Y_i)^T)(\widehat{\mathrm{Cov}}\,(\boldsymbol{\pi}(Y_i)))^{-\frac{1}{2}}\boldsymbol{\eta}_i.
\end{aligned} \tag{A.1}$$

The maximum of (A.1) is the largest eigenvalue of the matrix $(\widehat{\mathrm{Cov}}\,(\boldsymbol{\pi}(Y_i)))^{-1}$ $\widehat{\mathrm{Cov}}\,(\boldsymbol{\pi}(Y_i), \boldsymbol{X}^T)\widehat{\mathrm{Cov}}\,(\boldsymbol{X}, \boldsymbol{\pi}(Y_i))$. Therefore, the corresponding eigenvector $\hat{\boldsymbol{\eta}}_i$ is the maximizer. From this, we can easily obtain that

$$\hat{\boldsymbol{\theta}}_i^T = \frac{\hat{\boldsymbol{\eta}}_i^T \widehat{\mathrm{Cov}\,(\boldsymbol{\pi}(Y_i))}^{-\frac{1}{2}}}{\sqrt{\hat{\boldsymbol{\eta}}_i^T(\widehat{\mathrm{Cov}}\,(\boldsymbol{\pi}(Y_i)))^{-1}\hat{\boldsymbol{\eta}}_i}}.$$

Proof of Theorem 2. It suffices to prove that

$$\max_{\boldsymbol{H}\in\mathcal{H}_{bs}^q} |\hat{\mathbf{C}}_{bs}(\boldsymbol{H}) - \mathbf{C}_{bs}(\boldsymbol{H})| = O_p(\frac{(J^{\frac{9}{4}}(pq)^{\frac{3}{2}})}{\sqrt{n}}).$$

Rewrite $\mathbf{C}_{bs}(\boldsymbol{H})$ as $\sum_{i=1}^q \boldsymbol{\theta}_i^T \big[\,\mathrm{Cov}\,(\boldsymbol{\pi}_i(Y_i), \boldsymbol{X}^T)\,\mathrm{Cov}\,(\boldsymbol{X}, \boldsymbol{\pi}_i^T(Y_i))\big]\boldsymbol{\theta}_i$, where

$$\begin{aligned}
&\mathrm{Cov}\,(\boldsymbol{\pi}_i(Y), \boldsymbol{X}^T)\,\mathrm{Cov}\,(\boldsymbol{X}, \boldsymbol{\pi}_i(\boldsymbol{Y})^T)\\
&\quad = \Big(\sum_{l=1}^p \mathrm{Cov}\,(B_{ik}(Y_i), X_l) Cov(B_{ik_1}(Y_i), X_l)\Big)_{0\le k,k_1\le J+1}
\end{aligned}$$

is a $(J+2)\times(J+2)$ matrix, and similarly for $\hat{\mathbf{C}}_{bs}(\boldsymbol{H})$. Hence,

$$\begin{aligned}
&\max_{\boldsymbol{H}\in\mathcal{H}_{bs}^q} |\hat{\mathbf{C}}_{bs}(\boldsymbol{H}) - \mathbf{C}_{bs}(\boldsymbol{H})|\\
&\quad \le (J+2)^2 pq \max_{i,k,k_1,l} |\widehat{\mathrm{Cov}}\,(B_{ik}(Y_i), X_l)\widehat{\mathrm{Cov}}\,(B_{ik_1}(Y_i), X_l)\\
&\qquad - \mathrm{Cov}\,(B_{ik}(Y_i), X_l) Cov(B_{ik_1}(Y_i), X_l)|.
\end{aligned}$$

Note that (4.3) is implied by

$$\max_{i,k,l} |\widehat{\mathrm{Cov}}\,(B_{ik}(Y_i), X_l) - \mathrm{Cov}\,(B_{ik}(Y_i), X_l)| = O_p(J^{\frac{1}{4}}\sqrt{\frac{pq}{n}}). \tag{A.2}$$

We now prove (A.2). First, take

$$\begin{aligned}
&\widehat{\mathrm{Cov}}\,(B_{ik}(Y_i), X_l)) - \mathrm{Cov}\,(B_{ik}(Y_i), X_l))\\
&\quad = \frac{1}{n}\sum_{j=1}^n \big(B_{ik}(y_{ij})x_{lj}\big) - E\big(B_{ik}(y_{ij})x_l\big)\\
&\qquad + \frac{1}{n}\sum_{j=1}^n \big(B_{ik}(y_{ij})\big)\frac{1}{n}\sum_{j=1}^n \big(x_{lj}\big) - E\big(B_{ik}(y_{ij})\big)E\big(x_l\big)\\
&\quad = I_{n1,ilk} + I_{n2,ilk}.
\end{aligned} \tag{A.3}$$

For any $1\le i\le q$, $0\le k\le J+1$, $1\le l\le p$, and any $b>0$, by the Markov Inequality,

$$\max_{i,l,k} P\{|I_{n1,ilk}| > b\} \le \max_{i,l,k} \frac{E\big(B_{ik}(Y_i)X_l\big)^2}{nb^2}.$$

Note that $B_{ik}(Y_j) \le cI_{(t_k,t_{k+1})}(Y_j)$ and then $E\big(B_{ik}(Y_j)\big) \le c/J$. By the condition that $\max_l E(X^l)^4 < \infty$, we can obtain that

$$\max_{i,l,k} E\big(B_{i,k}(Y_i)X_l\big)^2 \le \sqrt{c^4 E(I_{(t_k,t_{k+1})}(Y_j))}\sqrt{E(X^l)^4} \le \frac{\mathbf{C}}{\sqrt{J}}.$$

Choosing $b = o(J^{1/4}\sqrt{pq/n})$, we find

$$P\{\max_{i,l,k} |I_{n1,ilk}| > b\} \le \sum_{i,l,k} \max_{i,l,k} P\{|I_{n1,ilk}| > b\} \le \frac{C(\sqrt{J}pq)}{nb^2} = o(1).$$

Similarly, we can derive that

$$\max_{i,k} |\frac{1}{n}\sum_{j=1}^{n} \big(B_{ik}(Y_j)\big) - E\big(B_{ik}(Y_i)\big)| = O_p(\frac{q}{\sqrt{n}})$$
$$\max_{l} |\frac{1}{n}\sum_{j=1}^{n} \big(x_{lj}\big) - E\big(X_l\big)| = O_p(\sqrt{\frac{p}{n}}).$$

A similar argument can be applied to prove the convergence of $I_{n2,ilk}$. By A.2 and A.3 the proof is finished.

Proof of Theorem 3. Let $\theta_i^{(l)}$ result from the kth step of the iterative algorithm. Hence,

$$2\boldsymbol{A}_i\big(\boldsymbol{\theta}_i^{(l+1)} - \boldsymbol{\theta}_i^{(l)}\big) = \alpha\Big(\frac{1}{1+\boldsymbol{\theta}_i^{(l)}} - \frac{1}{1+\boldsymbol{\theta}_i^{(l-1)}}\Big) = \alpha c_i^{(k)}\Big(\boldsymbol{\theta}_i^{(l)} - \boldsymbol{\theta}_i^{(l-1)}\Big),$$

and then

$$\big(\boldsymbol{\theta}_i^{(l+1)} - \boldsymbol{\theta}_i^{(l)}\big) = \frac{\alpha}{2}\boldsymbol{A}_i^{-1}\Big(\frac{1}{1+\boldsymbol{\theta}_i^{(l)}} - \frac{1}{1+\boldsymbol{\theta}_i^{(l-1)}}\Big) = \frac{\alpha}{2}\boldsymbol{A}_i^{-1}\boldsymbol{c}_i^{(k)}\Big(\boldsymbol{\theta}_i^{(l)} - \boldsymbol{\theta}_i^{(l-1)}\Big),$$

where $\boldsymbol{c}_i^{(k)} = 1/[(1+\boldsymbol{\theta}_i^{(l)})(1+\boldsymbol{\theta}_i^{(l-1)})]$. Note that all components of $c_i^{(k)}$ are smaller than or equal to 1. Because $\boldsymbol{A}_i$ is positive definite matrix, so is $\boldsymbol{A}_i^{-1}$. The largest eigenvalue $\lambda_{\max}((\boldsymbol{A}_i\boldsymbol{A}_i)^{-1}) = 1/\lambda_{\min}(\boldsymbol{A}_i\boldsymbol{A}_i) \le C < \infty$. Also note that for any non-negative symmetric matrix $\boldsymbol{B}$ and unitary vector $\boldsymbol{B}_1$, $\|\boldsymbol{B}\boldsymbol{B}_1\| = \boldsymbol{B}_1^T\boldsymbol{B}\boldsymbol{B}\boldsymbol{B}_1 \le \boldsymbol{B}_{1\max}(\boldsymbol{B}\boldsymbol{B})$, where $\boldsymbol{B}_{1\max}$ stands for the largest eigenvalue of the matrix $\boldsymbol{B}\boldsymbol{B}$. Therefore, it is easy to see that

$$\|\boldsymbol{\theta}_i^{(l+1)} - \boldsymbol{\theta}_i^{(l)}\| \le \frac{\alpha}{2(\lambda_{\min}(\boldsymbol{A}_i\boldsymbol{A}_i)))^{\frac{1}{2}}}\|\boldsymbol{\theta}_i^{(l)} - \boldsymbol{\theta}_i^{(-1)}\|.$$

When α is chosen to be smaller than $2(\lambda_{\min}(\boldsymbol{A}_i\boldsymbol{A}_i))^{1/2}$, the algorithm converges.

Proof Theorem 4. Using an argument that is similar to the one used for proving Theorem 2, we only need to prove that, in probability,

$$\max_{\boldsymbol{H}\in\mathcal{H}_{bs}^q} \left|\hat{\mathbf{C}}_{\alpha s}(\boldsymbol{H}) - \mathbf{C}_{\alpha s}(\boldsymbol{H})\right| = O\Big(\frac{J^{\frac{9}{4}}(pq)^{\frac{3}{2}}}{\sqrt{n}}\Big). \tag{A.4}$$

Note that

$$\begin{aligned}&\hat{\mathbf{C}}_{\alpha s,i}(\boldsymbol{H}) - \mathbf{C}_{\alpha s,i}(\boldsymbol{H})\\ &\quad= \big[\widehat{\mathrm{Cov}}\,(h_i(Y_i),\boldsymbol{X}^T)\big]\big[\widehat{\mathrm{Cov}}\,(\boldsymbol{X},h_i(Y_i))\big] - \big[\mathrm{Cov}\,(h_i(Y_i),\boldsymbol{X}^T)\big]\big[\mathrm{Cov}\,(\boldsymbol{X},h_i(Y_i))\big].\end{aligned}$$

The argument that was used for proving Theorem 2 applied and one obtains the bound of (A.4). The details are omited.

References

Breiman, L. and Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *J. Amer. Statist. Assoc.* **80**, 580-598.

Bura, E. and Cook, D. (2001). Estimating the structural dimension of regression via parametric inverse regression. *J. Roy. Statist. Soc. Ser. B* **63**, 393-410.

Cook, R. D. (1998). *Regression Graphics: Ideas for Studying Regressions through Graphics*, Wiley, New York.

Cook, R. D. and Li, B. (2002). Dimension reduction for the conditional mean in regression. *Ann. Statist.* **30**, 455-474.

Cook, R. D. and Setodji, C. M. (2003). A model-free test for reduced rank in multivariate regression. *J. Amer. Statist. Assoc.* **98**, 340-351.

Cook, D. and Weisberg, S. (1994). Transforming a response variable for linearity, *Biometrika* **81**, 731-737.

Doksum, K. A. (1987). An extension of partial likelihood methods for proportional hazard models to general transformation models. *Ann. Statist.* **15**, 325-345

Friedman, J. H. and Stuetzle, W. (1981). Projection pursuit regression, *J. Amer. Statist. Assoc.* **76**, 817-823.

Fujikoshi, Y. and Satoh, K. (1997). Modified AIC and C_p in multivariate linear regression. *Biometrika* **84**, 707-716.

Fung, W. K., He, X., Liu, L. and Shi, P. (2002). Dimension reduction based on canonical correlation. *Statist. Sinica* **12**, 1093-1113.

Györfi, L., Kohler, M., Krzyźak, A. and Walk, H. (2002) *A Distribution-Free Theory of Nonparametric Regression.* Springer, New York.

Hall, P. (1989). On projection pursuit regression. *Ann. Statist.* **17**, 573-588.

He, X. and Shen, L. (1997). Linear regression after spline transformation. *Biometrika* **84**, 474-481.

Huber, P. J. (1985). Projection pursuit (with discussion). *Ann. Statist.* **13**, 435-475.

Li, K. C. (1991). Sliced inverse regression for dimension reduction (with discussion). *J. Amer. Statist. Assoc.* **86**, 316-342.

Li, K. C. (1992). On principal Hessian directions for data visualization and dimension reduction: another application of Stein's lemma. *J. Amer. Statist. Assoc.* **87**, 1025-1039.

Li, K. C. and Duan, N. H. (1989). Regression analysis under link violation. *Ann. Statist.* **17**, 1009-1052.

Li, K. C., Aragon Y., Shedden, K., and Agnan, C. T. (2003). Dimension reduction for multivariate response data. *J. Amer. Statist. Assoc.* **98**, 99-109.

Li, X. (1999). Quantatitive Definition and Division of Meteorological Disaster in Ji-Lin Province. Masters degree Thesis, Northwest Normal University, Ji-Lin province, China.

Ma, S. Q. (1996). *The Study on Agriculture Climate in Ji-Lin Province.* Meteorology Press, Beijing.

McQuarrie, A. D. R. and Trai, C. L. (1998). *Regression and Time Series Model Selection.* World Scientific, Singapore.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461-464.

Ramsay, J. O. (1988). Monotone regression spline in action (with discussion). *Statist. Sci.* **3**, 425-461.

Schumaker, L. L. (1981). *Spline Functions.* Wiley, New York.

Tenehaus, M. (1998). La Régression PLS: Théorie et Pratique. Éditons Technip, Paris. (in French)

Xia, Y., Li, W. K., Tong, H. and Zhu, L. X. (2000). On the estimation of an instantaneous transformation for time series, *J. Roy. Statist. Soc. Ser. B* **62**, 283-297.

Yin, X. and Bura, E. (2006). Moment based dimension reduction for multivariate response regression. *J. Statist. Plann. Inference* **136**, 3675-3688.

Yin, X. and Cook, R. D. (2002). Dimension reduction for the conditional *k*th moment in regression. *J. Roy. Statist. Soc. Ser. B* **64**, 159-175.

Yin, X. and Cook, R. D. (2003). Estimating the central subspace via inverse third moment. *Biometrika* **90**, 113-125.

Yin, X. and Cook, R. D. (2004). Dimension reduction via marginal fourth moments in regression. *J. Comput. Graph. Statist.* **13**, 554-570.

Yin, X. and Zhu, L. X. (2004). On estimation for dual central subspaces in dimension reduction. Submitted.

Zhu, L. X. and Fang, K. T. (1992). On Projection Pursuit approximation for nonparametric regression. In *Proceedings of Order Statistics and Nonparametric: Theory and applications* (Edited by P. K. Sen and A. I. Salama), 455-469. Elsevier Science Publishers.

Department of Mathematics, Hong Kong Baptist University, Hong Kong, China.
E-mail: lzhu@hkbu.edu.hk

Department of Statistics, East China Normal University, Shanghai, China.
E-mail: zhulp1@yahoo.com.cn

Department of Mathematics, Beijing Institute of Technology, Beijing, China.
E-mail: susanli.xin@gmail.com

(Received July 2005; accepted March 2006)