Statistica Sinica Preprint No: SS-2016-0080R3	
Title	Flexible dimension reduction in regression
Manuscript ID	SS-2016-0080R3
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202016.0080
Complete List of Authors	Lixing Zhu and Tao Wang
Corresponding Author	Lixing Zhu
E-mail	lzhu@hkbu.edu.hk

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0080

Flexible dimension reduction in regression*

Tao Wang[1,2] and Lixing Zhu[3]

[1] Department of Bioinformatics and Biostatistics, Shanghai Jiao Tong University, Shanghai, China

[2] SJTU-Yale Joint Center for Biostatistics, Shanghai Jiao Tong University, Shanghai, China

[3] Department of Mathematics, Hong Kong Baptist University, Hong Kong

Abstract

Sliced inverse regression is a valuable tool for dimension reduction. One can replace the predictor vector with a few linear combinations of its components without loss of information on the regression. This paper is about richer nonlinear dimension reduction. Each direction of sliced inverse regression is simply a slope vector of multiple linear regression applied to an optimally transformed response. Using this connection, we propose a nonlinear version of sliced inverse regression by replacing linear function by an additive function of the predictors. Our procedure has a clear interpretation as sliced inverse regression on a set of adaptively chosen transformations of the predictors. The flexibility of our method is illustrated via a simulation study and a data application.

Key words: Canonical correlation; Optimal scoring; Sufficient dimension reduction.

*The research described herewith was supported by National Natural Science Foundation of China (11601326), and the University Grants Council of Hong Kong, Hong Kong. The authors thank the Editor, the Associate Editor, and two anonymous referees for their constructive comments and suggestions that led to an improvement of an early manuscript.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0080

1 Introduction

Transformation is a commonly used technique in statistics. Regarding regression, transformation of the response variable can often simultaneously achieve normally distributed errors with a constant variance and a linear regression function. This methodology, pioneered by Box and Cox (1964), is perhaps the most common. Then too, transformation of the predictors can convert a complex nonlinear regression relation into a simple and often linear one. For example, if it appears in a problem that many predictors are likely to need transformation, iterative fitting of additive models should be considered. More generally, nonlinear multivariate techniques, such as alternating conditional expectations (Breiman and Friedman (1985)) and monotone spline regression (Ramsay (1988)), allow transformation on both the response variable and the predictors. One transformation on the response variable is allowed in these algorithms for the purpose of model fitting.

A fairly common practice is to apply first either marginal transformations or a joint transformation to multivariate data to induce normality, elliptical symmetry, or other appropriate distributions, and then carry out model-based regression, in most cases linear regression, on the transformed data. This often leads to model mis-specification giving biased estimators. In principle, any use of transformations requires that the effects of them on the error structure be understood. The story is different, however, in the context of dimension reduction.

Regarding dimension reduction in regression, transformation of variables plays two roles. Response transformations, such as slicing (Duan and Li (1991), Li (1991)) and spline transformation (He and Shen (1997), Fung et al. (2002)), serve as an intermediate tool for finding interesting patterns in the data, instead of being used to improve the goodness of model fitting in a traditional way. See also Zhu et al. (2010), Wu and Li (2011), Yin and Li (2011), and references therein. Predictor transformations are also useful for reducing the structural dimension of the regression when the data are concentrated on a nonlinear low-dimensional

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0080

space (Cook (1998, Chapter 14)). In many cases, it is preferable to use predictor transformations rather than response transformations, because transformation of the predictors does not change the interpretation of the response variable. This is particularly the case when we are not assuming a model for the regression.

Related to the present work is Wang et al. (2014), who present a framework for dimension reduction in regression that lies between linear and fully nonlinear dimension reduction. Suppose Y is a univariate response variable and $\boldsymbol{X} = (X_1, \ldots, X_p)^\top$ is a p-vector of predictors. The main idea is to first transform each of the raw predictors marginally and monotonically, in the form $\boldsymbol{f}(\boldsymbol{X}) = \{f_1(X_1), \ldots, f_p(X_p)\}^\top$, and then search for a low-dimensional projection in the space defined by the transformed predictors. Toward this end, they assume that, given $\mathbf{B}^\top \boldsymbol{f}(\boldsymbol{X})$, Y is independent of $\boldsymbol{f}(\boldsymbol{X})$, where $\mathbf{B}$ is a $p \times d$ matrix with $d \leq p$. The aim of the analysis is to characterize the subspace spanned by the columns of $\mathbf{B}$, and for this purpose they propose a two-step procedure by combining probability integral transformation and sliced inverse regression (Li (1991)). Although $\boldsymbol{f}(\boldsymbol{X})$ or its distribution is user-specified, probability integral transformed sliced inverse regression suffers from mis-specification of transformations.

We propose a new nonlinear dimension-reduction method to overcome this problem. This is primarily motivated by observing the close connection between sliced inverse regression and multiple linear regression. Our procedure estimates predictor transformations in a data-driven way, and thus can be regarded as an adaptive version of probability integral transformed sliced inverse regression.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0080

2 Methodology

2.1 Sliced inverse regression by optimal scoring

The basic idea of sliced inverse regression by optimal scoring (Wang and Zhu (2013)) is to linearise a response transformation $T(Y)$ by $\boldsymbol{\phi}(Y)^\top\boldsymbol{\theta}$, where $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_K)^\top$ is a K-vector of basis functions, and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)^\top$ is a K-vector of unknown scores. The scored data is then predicted by linear regression on $\boldsymbol{X}$. The simultaneous estimation of the scores and the linear parameters constitutes the optimal scoring problem.

Suppose that $\{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$ is a random sample on $(\boldsymbol{X}, Y)$. Let $\mathbf{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\top$ and $\boldsymbol{\Phi} = \{\boldsymbol{\phi}(y_1), \ldots, \boldsymbol{\phi}(y_n)\}^\top$ be the two data matrices containing the predictor values and the basis function values, respectively. Write $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_p)$ with $\mathbf{X}_k = (x_{1k}, \ldots, x_{nk})^\top$ as its k-th column. Without loss of generality, we assume that the columns of $\mathbf{X}$ are centered. In the sample, the criterion of sliced inverse regression by optimal scoring takes the form

$$\begin{aligned} \underset{\boldsymbol{\theta}_i \in \mathbf{R}^K, \boldsymbol{\beta}_i \in \mathbf{R}^p}{\text{minimize}} \quad & \|\boldsymbol{\Phi}\boldsymbol{\theta}_i - \mathbf{X}\boldsymbol{\beta}_i\|_2^2 \\ \text{subject to} \quad & \boldsymbol{\theta}_i^\top \boldsymbol{\Phi}^\top \boldsymbol{\Phi} \boldsymbol{\theta}_i = n, \boldsymbol{\theta}_i^\top \boldsymbol{\Phi}^\top \boldsymbol{\Phi} \boldsymbol{\theta}_j = 0, \quad j = 1, \ldots, i-1, \end{aligned} \tag{2.1}$$

where $i = 1, \ldots, d \leq K$.

There are various choices for the basis functions. We concentrate on slice indicator functions. Since the columns of $\mathbf{X}$ are centered to have mean zero, one can see that the constant score vector $\mathbf{1}$ of length K is trivial, and hence there are at most $K-1$ nontrivial solutions to (2.1).

Remark 2.1. In the population, sliced inverse regression by optimal scoring sequentially solves

$$\begin{aligned} \underset{T_i, a_i \in \mathbf{R}, \boldsymbol{b}_i \in \mathbf{R}^p}{\text{minimize}} \quad & E(T_i - a_i - \boldsymbol{X}^\top \boldsymbol{b}_i)^2 \\ \text{subject to} \quad & \text{var}\{T_i(Y)\} = 1, \text{cov}\{T_i(Y), T_j(Y)\} = 0, \quad j = 1, \ldots, i-1. \end{aligned} \tag{2.2}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0080

The i-th optimal transformation $T_i(Y)$ is identical up to a scalar multiplication to $E(\boldsymbol{\eta}_i^\top \boldsymbol{X}|Y)$ and $\boldsymbol{b}_i(T_i)$ is proportional to $\boldsymbol{\eta}_i$, where $\boldsymbol{\eta}_i$ is the i-th sliced inverse regression direction such that

$$\text{cov}\{E(\boldsymbol{X}|Y)\}\boldsymbol{\eta}_i = \lambda_i \text{cov}(\boldsymbol{X})\boldsymbol{\eta}_i.$$

Here $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_p$ are eigenvectors satisfying $\boldsymbol{\eta}_i^\top \text{cov}(\boldsymbol{X})\boldsymbol{\eta}_j = 1$ if $i = j$, and 0 if $i \neq j$, and $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$ are the corresponding eigenvalues. Details can be found in Chen and Li (1998) and Wang and Zhu (2013).

To study the relationship between a set of predictors and a categorical response, it is known that Fisher's linear discriminant analysis, canonical correlation analysis, and optimal scoring are equivalent from the dimension-reduction point of view (Hastie et al. (1995)). These data-analytic tools, popular for classification, are also useful in the regression setting. First, the original sliced inverse regression method is formally equivalent to Fisher's linear discriminant analysis (Kent (1991)), and it is tantamount to sliced inverse regression by optimal scoring by using the slice indicator functions obtained from slicing the response variable. Second, Fung et al. (2002) proposed a dimension-reduction method based on canonical correlation, which can be viewed as a variant of sliced inverse regression. The latter estimates the kernel matrix $\text{cov}\{E(\boldsymbol{X}|Y)\}$ using the slice indicator functions, while the former produces a spline-based estimate by using the B-spline basis functions generated for the response variable. Clearly, optimal scoring and canonical correlation analysis are equivalent without regard to classification or regression. Therefore, sliced inverse regression by optimal scoring includes the original sliced inverse regression method and the canonical correlation method as special cases.

In typical applications, the number of basis functions K needed is very small (He and Shen (1997), Fung et al. (2002)), and it is well known that the original sliced inverse regression method is not overly sensitive to the number of slices.

Remark 2.2. Due to the equivalence between optimal scoring and canonical correlation

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0080

analysis, sliced inverse regression by optimal scoring should have the same theoretical properties as those of the canonical correlation method. Further, sliced inverse regression by optimal scoring can be interpreted as a method for estimating the canonical correlations between the columns of $\mathbf{X}$ and the columns of $\mathbf{\Phi}$, and thus the directions found by it are useful in their own right in identifying some important features of the regression of Y on $\boldsymbol{X}$. In particular, if the linearity condition of Li (1991) holds, the directions $\boldsymbol{\eta}_i$ in Remark 2.1 belong to the central dimension-reduction subspace, an essential concept of sufficient dimension reduction (Cook (1998)).

2.2 Flexible dimension reduction

We introduce some notation and definitions. We call span($\mathbf{B}$) a transformed dimension-reduction subspace with respect to $\boldsymbol{f}$ if

$$Y \perp\!\!\!\perp \boldsymbol{X} \mid \mathbf{B}^\top \boldsymbol{f}(\boldsymbol{X}). \tag{2.3}$$

span($\mathbf{B}$) is called a transformed central dimension-reduction subspace if it satisfies (2.3) for some $\boldsymbol{f}$, and at the same time it has the smallest dimension. The following proposition shows that the transformed central dimension-reduction subspace is well defined.

Proposition 2.1. *If there is a p-vector $\boldsymbol{x}^* = (x_1^*, \ldots, x_p^*)^\top$, such that for all $j = 1, \ldots, p$, $f_j'(x_j^*) = 1$ and $f_j'(x_j)$ is continuous at x_j^*, where $f_j'(x_j)$ denotes the derivative of $f_j(x_j)$ with respective to x_j, then the transformed central dimension-reduction subspace is identifiable.*

Remark 2.3. While Wang et al. (2014) requires the transformations f_j to be pre-specified, we propose to simultaneously estimate f_j and perform sufficient dimension reduction. One consequence of this difference is that f_j may not be identifiable, even if the structural dimension is given. For example, if $Y = X_1^3 + \epsilon$, then $f_1(x)$ can be the identity map x or x^3, or any other monotone function. Sufficient dimension reduction is often considered as the first step in statistical analyses. After that, graphical tools or nonparametric methods can be

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0080

used to further investigate the relationship between the response variable and the reduced set of variables. Suppose $\boldsymbol{g} = \{g_1, \ldots, g_p\}$ is an alternative set of transformations such that span($\mathbf{B}$) is a transformed central subspace with respect to $\boldsymbol{g}$. Theoretically, $\mathbf{B}^\top \boldsymbol{f}(\boldsymbol{X})$ and $\mathbf{B}^\top \boldsymbol{g}(\boldsymbol{X})$ can be treated equally, because each of them is a minimal and sufficient reduced predictor; numerically, it is possible that the identifiability issue of transformation functions affects speed and convergence.

Sliced inverse regression by optimal scoring is a linear dimension-reduction method. However, our interest in this approach is that it includes linear regression as a building block. Many techniques exist for generalizing linear regression to more flexible, nonlinear and/or nonparametric forms of regression. This in turn leads to more flexible forms of dimension reduction. One simple and effective approach is to augment the set of predictors to include quadratic and bilinear terms, and then carry out sliced inverse regression by optimal scoring in the enlarged space which, in effect, results in nonlinear dimension reduction in the original predictor space. More flexible approaches use adaptive nonparametric regression or kernels (Hastie et al. (1994, 2009)).

The procedure has a clear interpretation as a linear dimension-reduction method on a set of marginally and adaptively transformed predictors. Specifically, we transform the raw predictors $\boldsymbol{X} = (X_1, \ldots, X_p)^\top$ in the form $\boldsymbol{f}(\boldsymbol{X}) = \{f_1(X_1), \ldots, f_p(X_p)\}^\top$, or simply $\boldsymbol{f} = (f_1, \ldots, f_p)^\top$, for a set of smooth univariate functions $\{f_1, \ldots, f_p\}$ and, at the same time, consider the optimal scoring problem with the transformed data as follows.

$$\begin{aligned} &\underset{\boldsymbol{f},\{\boldsymbol{\theta}_i\},\{\boldsymbol{\beta}_i\}}{\text{minimize}} \quad \sum_{i=1}^{H} \|\boldsymbol{\Phi}\boldsymbol{\theta}_i - \mathbf{X}_{\boldsymbol{f}}\boldsymbol{\beta}_i\|_2^2 \\ &\text{subject to} \quad \boldsymbol{\theta}_i^\top \boldsymbol{\Phi}^\top \boldsymbol{\Phi}\boldsymbol{\theta}_i = n, \boldsymbol{\theta}_i^\top \boldsymbol{\Phi}^\top \boldsymbol{\Phi}\boldsymbol{\theta}_j = 0, \quad j = 1, \ldots, i-1, i = 1, \ldots, H, \end{aligned} \tag{2.4}$$

where $1 \le H \le K$ is a working dimension, $\mathbf{X}_{\boldsymbol{f}} = (\mathbf{X}_{f_1}, \ldots, \mathbf{X}_{f_p})$ with $\mathbf{X}_{f_k} = \{f_k(x_{1k}), \ldots, f_k(x_{nk})\}^\top$ as its k-th column. Since the f_k's in (2.4) are up to scale and shift, we demand that $\sum_{i=1}^n f_k(x_{ik}) = 0$ and $\sum_{i=1}^n \{f_k(x_{ik})\}^2 = n$.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0080

The restriction to component-wise monotone transformations is not always necessary. Our procedure thus permits as much flexibility in the estimated transformations as the data require, and can be interpreted as a method for estimating the canonical correlations between the transformed predictors and the basis functions generated for the response variable. This makes it different from the probability integral transformed sliced inverse regression in Wang et al. (2014).

Remark 2.4. One advantage of the proposed method over the original sliced inverse regression is that it can capture marginal symmetry between $\boldsymbol{X}$ and Y, such as in the model $Y = X_1^2 + X_2^2 + \epsilon$, where ϵ is the error independent of $\boldsymbol{X}$, and $\boldsymbol{X}$ follows a symmetric distribution. This is related to the robustness of sliced inverse regression against violation of the linearity condition, given that $\boldsymbol{f}$ is correctly specified. Our method also inherits one drawback of sliced inverse regression: it fails when the functional relation is jointly symmetric in the sense that $E(\boldsymbol{f} \mid Y)$ is zero, such that in the model $Y = (X_1^3 + X_2^3)^2 + \epsilon$.

The proposed procedure allows multiple transformations on the response variable for the purpose of dimension reduction. It is closely related to the monotone spline canonical correlation of Ramsay (1988), which is a tool for comparison between two sets of marginally and monotonically transformed variables via canonical correlation. The difference is that our procedure is concerned with dimension reduction in regression based on both response transformations and predictor transformations and, in particular, the response transformations (or the set of basis functions) and the predictor transformations are not restricted to be monotone.

2.3 Algorithm

To estimate $\boldsymbol{f}, \{\boldsymbol{\theta}_i\}$, and $\{\boldsymbol{\beta}_i\}$ in (2.4), we use an iterative approach: we first fix $\{\boldsymbol{\theta}_i\}$ and $\{\boldsymbol{\beta}_i\}$ and estimate $\boldsymbol{f}$, then we fix $\boldsymbol{f}$ and estimate $\{\boldsymbol{\theta}_i\}$ and $\{\boldsymbol{\beta}_i\}$, and we iterate between

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0080

these two steps until the algorithm converges.

When $\{\boldsymbol{\theta}_i\}$ and $\{\boldsymbol{\beta}_i\}$ are fixed, minimizing (2.4) with respect to $\boldsymbol{f} = (f_1, \ldots, f_p)^\top$ is similar to fitting an additive model (Wood (2006)). To this end, we use a variant of the back-fitting procedure. Write $\boldsymbol{\beta}_i = (\beta_{i1}, \ldots, \beta_{ip})^\top$. For each $k = 1, \ldots, p$, consider the one-dimensional smoothing problem

$$\underset{f_k}{\text{minimize}} \quad \sum_{i=1}^{H} \left\| \boldsymbol{\Phi}\boldsymbol{\theta}_i - \sum_{l \neq k} \beta_{il} \mathbf{X}_{f_l} - \beta_{ik} \mathbf{X}_{f_k} \right\|_2^2. \tag{2.5}$$

Let $\tilde{\boldsymbol{x}}_k = (\tilde{\boldsymbol{x}}_{1k}^\top, \ldots, \tilde{\boldsymbol{x}}_{Hk}^\top)^\top$, $\boldsymbol{w}_k = (\boldsymbol{w}_{1k}^\top, \ldots, \boldsymbol{w}_{Hk}^\top)^\top$, and $\tilde{\boldsymbol{y}}_k = (\tilde{\boldsymbol{y}}_{1k}^\top, \ldots, \tilde{\boldsymbol{y}}_{Hk}^\top)^\top$, where

$$\tilde{\boldsymbol{x}}_{ik} = (x_{1k}, \ldots, x_{nk})^\top, \boldsymbol{w}_{ik} = (\beta_{ik}^2, \ldots, \beta_{ik}^2)^\top \in \mathbb{R}^n,$$

$$\tilde{\boldsymbol{y}}_{ik} = \frac{\boldsymbol{\Phi}\boldsymbol{\theta}_i - \sum_{l \neq k} \beta_{il} \mathbf{X}_{f_l}}{\beta_{ik}}.$$

Let $\tilde{n} = Hn$ and write $\tilde{\boldsymbol{x}}_k = (\tilde{x}_{1k}, \ldots, \tilde{x}_{\tilde{n}k})^\top$, $\boldsymbol{w}_k = (w_{1k}, \ldots, w_{\tilde{n}k})^\top$, and $\tilde{\boldsymbol{y}}_k = (\tilde{y}_{1k}, \ldots, \tilde{y}_{\tilde{n}k})^\top$. One can show that (2.5) is equivalent to

$$\underset{f_k}{\text{minimize}} \quad \sum_{s=1}^{\tilde{n}} w_{sk} \{\tilde{y}_{sk} - f_k(\tilde{x}_{sk})\}^2. \tag{2.6}$$

The criterion adopted here is slightly different from that used in the back-fitting procedure. Specifically, in each (inner) iteration we identify and update the function that most reduces the objective function, but that function is kept fixed and is not further updated in the next iteration. We stop the algorithm if convergence or when all the functions are updated. In our implementation, each f_k is represented using penalized regression splines with smoothing parameters selected by restricted maximum likelihood. Specifically, we use the function **gam** in the R package **mgcv**.

When $\boldsymbol{f}$ is fixed, (2.4) is the standard optimal scoring problem

$$\begin{aligned} \underset{\boldsymbol{\theta}_i \in \mathbf{R}^K, \boldsymbol{\beta}_i \in \mathbf{R}^p}{\text{minimize}} \quad & \|\boldsymbol{\Phi}\boldsymbol{\theta}_i - \mathbf{X}_{\boldsymbol{f}}\boldsymbol{\beta}_i\|_2^2 \\ \text{subject to} \quad & \boldsymbol{\theta}_i^\top \mathbf{D} \boldsymbol{\theta}_i = 1, \boldsymbol{\theta}_i^\top \mathbf{D} \boldsymbol{\theta}_j = 0, \quad j = 1, \ldots, i-1, \end{aligned} \tag{2.7}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0080

where $\mathbf{D} = n^{-1}\mathbf{\Phi}^\top\mathbf{\Phi}$ and $i = 1, \ldots, H$. The standard way to solve an optimal scoring problem is to use a suitable eigenvalue decomposition. However, we propose to update $\{\boldsymbol{\theta}_i\}$ and $\{\boldsymbol{\beta}_i\}$ separately, as follows. For fixed $\{\boldsymbol{\beta}_i\}$, the optimal scores $\{\boldsymbol{\theta}_i\}$ sequentially solve

$$\begin{aligned} &\underset{\boldsymbol{\theta}_i \in \mathbf{R}^K}{\text{minimize}} \quad \|\mathbf{\Phi}\boldsymbol{\theta}_i - \mathbf{X}_{\boldsymbol{f}}\boldsymbol{\beta}_i\|_2^2 \\ &\text{subject to} \quad \boldsymbol{\theta}_i^\top \mathbf{D}\boldsymbol{\theta}_i = 1, \boldsymbol{\theta}_i^\top \mathbf{D}\boldsymbol{\theta}_j = 0, \quad j = 1, \ldots, i-1. \end{aligned} \tag{2.8}$$

Let $\mathbf{Q}_i = (\mathbf{1}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{i-1})$ denote the $K \times i$ matrix consisting of the previous $i-1$ solutions including the trivial score vector of all ones. One can show that the i-th solution is given by

$$\boldsymbol{\theta}_i = c_i(\mathbf{I}_K - \mathbf{Q}_i\mathbf{Q}_i^\top\mathbf{D})\mathbf{D}^{-1}\mathbf{\Phi}^\top\mathbf{X}_{\boldsymbol{f}}\boldsymbol{\beta}_i,$$

where $\mathbf{I}_K$ is the $K \times K$ identity matrix and c_i is a constant such that $\boldsymbol{\theta}_i^\top\mathbf{D}\boldsymbol{\theta}_i = 1$. For fixed $\{\boldsymbol{\theta}_i\}$, we obtain H linear least squares problems

$$\underset{\boldsymbol{\beta}_i \in \mathbf{R}^p}{\text{minimize}} \quad \|\mathbf{\Phi}\boldsymbol{\theta}_i - \mathbf{X}_{\boldsymbol{f}}\boldsymbol{\beta}_i\|_2^2, \tag{2.9}$$

with solutions $\boldsymbol{\beta}_i = (\mathbf{X}_{\boldsymbol{f}}^\top\mathbf{X}_{\boldsymbol{f}})^{-1}\mathbf{X}_{\boldsymbol{f}}^\top\mathbf{\Phi}\boldsymbol{\theta}_i$.

Our proposed algorithm for solving (2.4) is as follows.

1. Standardization and initialization: Center and normalize $\mathbf{X}$. Let $\mathbf{X}_{\boldsymbol{f}} = \mathbf{X}$. Initialize $\boldsymbol{\theta}_i$ and $\boldsymbol{\beta}_i$ with some plausible values. Let $\boldsymbol{w}_{ik} = (\beta_{ik}^2, \ldots, \beta_{ik}^2)^\top \in \mathbb{R}^n$ and $\boldsymbol{w}_k = (\boldsymbol{w}_{1k}^\top, \ldots, \boldsymbol{w}_{Hk}^\top)^\top$. Write $\boldsymbol{w}_k = (w_{1k}, \ldots, w_{\tilde{n}k})^\top$.

2. Iterate until convergence in terms of $\{\boldsymbol{\beta}_i\}$ or until a maximum number of iterations is reached.

   2.1. Update $\boldsymbol{f} = (f_1, \ldots, f_p)^\top$: Set $\mathcal{C} = \{1, \ldots, p\}$ and $\breve{\boldsymbol{y}}_i = \mathbf{\Phi}\boldsymbol{\theta}_i$.

      2.1.1. For each $k \in \mathcal{C}$, let

$$\tilde{\boldsymbol{y}}_{ik} = \frac{\breve{\boldsymbol{y}}_i - \sum_{l \in \mathcal{C}, l \neq k} \beta_{il}\mathbf{X}_{f_l}}{\beta_{ik}},$$

$$\tilde{\boldsymbol{y}}_k = (\tilde{\boldsymbol{y}}_{1k}^\top, \ldots, \tilde{\boldsymbol{y}}_{Hk}^\top)^\top = (\tilde{y}_{1k}, \ldots, \tilde{y}_{\tilde{n}k})^\top.$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0080

Solve

$$\underset{f_k}{\text{minimize}} \quad \sum_{s=1}^{\tilde{n}} w_{sk}\{\tilde{y}_{sk} - f_k(\tilde{x}_{sk})\}^2.$$

2.1.2. Choose the function $f_{k'}$, $k' \in \mathcal{C}$, that most reduces the objective function. Update

$$f_{k'} \leftarrow f_{k'}, \quad \breve{\boldsymbol{y}}_i \leftarrow \breve{\boldsymbol{y}}_i - \beta_{ik'}\mathbf{X}_{f_{k'}} \quad \text{and} \quad \mathcal{C} \leftarrow \mathcal{C} \setminus \{k'\}.$$

2.1.3. Continue until the change in the objective function falls below a prespecified threshold, or until all p functions have been updated. Center $\mathbf{X}_{\boldsymbol{f}}$.

2.2. Update $\{\boldsymbol{\theta}_i\}$: Let $\mathbf{Q}_i = (\mathbf{1}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{i-1})$ and

$$\boldsymbol{\theta}_i = (\mathbf{I}_K - \mathbf{Q}_i\mathbf{Q}_i^\top\mathbf{D})\mathbf{D}^{-1}\boldsymbol{\Phi}^\top\mathbf{X}_{\boldsymbol{f}}\boldsymbol{\beta}_i.$$

Normalize $\boldsymbol{\theta}_i$ so that $\boldsymbol{\theta}_i^\top\mathbf{D}\boldsymbol{\theta}_i = 1$. Normalize $\mathbf{X}_{\boldsymbol{f}}$.

2.3. Update $\{\boldsymbol{\beta}_i\}$: Let

$$\boldsymbol{\beta}_i = (\mathbf{X}_{\boldsymbol{f}}^\top\mathbf{X}_{\boldsymbol{f}})^{-1}\mathbf{X}_{\boldsymbol{f}}^\top\boldsymbol{\Phi}\boldsymbol{\theta}_i.$$

2.4. Update $\{\boldsymbol{w}_k\}$: Let $\boldsymbol{w}_{ik} = (\beta_{ik}^2, \ldots, \beta_{ik}^2)^\top \in \mathbb{R}^n$ and $\boldsymbol{w}_k = (\boldsymbol{w}_{1k}^\top, \ldots, \boldsymbol{w}_{Hk}^\top)^\top$. Write $\boldsymbol{w}_k = (w_{1k}, \ldots, w_{\tilde{n}k})^\top$.

Since the value of the objective function decreases at each step, the algorithm converges to a unique solution. However, there is no guarantee that the solution minimizes the objective function, because the overall problem is not convex. This is the price paid for computational efficiency, in view of the difficulty of carrying out the simultaneous minimization of (2.4). Our limited experience in simulation studies shows that the algorithm converges quickly, and that it works well empirically.

2.4 Theoretical properties

We assume that for each j, there is a known finite-dimensional envelope $\{g_{j1}, \ldots, g_{jq_j}\}$ such that f_j is a linear combination of $g_{j1}, \ldots, g_{jq_j}$. Let $\boldsymbol{g}_j(X_j) = \{g_{j1}(X_j), \ldots, g_{jq_j}(X_j)\}^\top$.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0080

Under this assumption, $f_j = \boldsymbol{c}_j^\top \boldsymbol{g}_j$ for some $\boldsymbol{c}_j \in \mathbb{R}^{q_j}$. We can write $\boldsymbol{f} = \mathbf{C}\boldsymbol{g}$, where $\mathbf{C} = \text{diag}(\boldsymbol{c}_1^\top, \ldots, \boldsymbol{c}_p^\top)$ and $\boldsymbol{g} = (\boldsymbol{g}_1^\top, \ldots, \boldsymbol{g}_p^\top)^\top$. Let $\boldsymbol{c}_{0j} \in \mathbb{R}^{q_j}$, $\mathbf{C}_0 = \text{diag}(\boldsymbol{c}_{01}^\top, \ldots, \boldsymbol{c}_{0p}^\top)$, and $\boldsymbol{f}_0 = \mathbf{C}_0\boldsymbol{g}$. To study the asymptotic behavior of our procedure, we require that

$$Y = G(\mathbf{B}_0^\top \boldsymbol{f}_0, \epsilon), \tag{2.10}$$

where $G(\cdot, \cdot)$ is an unknown function, $\mathbf{B}_0 \in \mathbb{R}^{p \times d_0}$ with $d_0 \geq 1$, and ϵ is independent of $\boldsymbol{g}$.

Remark 2.5. $\text{span}(\mathbf{B}_0)$ and $\boldsymbol{f}_0$ are generally not identifiable, since $\mathbf{B}_0^\top \boldsymbol{f}_0 = (\boldsymbol{\Gamma}\mathbf{B}_0)^\top \boldsymbol{\Gamma}\boldsymbol{f}_0$ for any $p \times p$ diagonal matrix $\boldsymbol{\Gamma}$ containing 1 or -1 in its diagonal. This makes our procedure different from sliced inverse regression when $\boldsymbol{f}_0$ is known. If the ultimate goal is the reduction $\mathbf{B}_0^\top \boldsymbol{f}_0$ itself, then it is unnecessary to impose identifiability constraints. In particular, the structural dimension d_0 remains estimable.

Let $\hat{\mathbf{B}} \in \mathbb{R}^{p \times d_0}$, $\hat{\mathbf{C}} \in \mathbb{R}^{p \times q}$, and $\hat{\boldsymbol{f}} = \hat{\mathbf{C}}\boldsymbol{g}$ denote our estimates of $\mathbf{B}_0$, $\mathbf{C}_0$, and $\boldsymbol{f}_0$, respectively.

THEOREM 2.1. *Suppose $d_0 \leq H \leq K$. If (A1) $E(\boldsymbol{g}|\mathbf{B}_0^\top \mathbf{C}_0 \boldsymbol{g})$ is linear in $\mathbf{B}_0^\top \mathbf{C}_0 \boldsymbol{g}$ and (A2) $\text{span}[\text{cov}\{E(\boldsymbol{f}_0|Y)\}] = \text{span}(\mathbf{B}_0)$, then for some $p \times p$ diagonal matrix $\boldsymbol{\Gamma}_n$ containing 1 or -1 in its diagonal, we have $\boldsymbol{\Gamma}_n\hat{\mathbf{C}} - \mathbf{C}_0 = O_P(n^{-1/2})$ and $\text{span}\{\hat{\mathbf{B}}\}$ is a $\sqrt{n}$-consistent estimator of $\text{span}\{\boldsymbol{\Gamma}\mathbf{B}_0\}$.*

The linearity condition (A1) and the coverage condition (A2), assumed to hold at the true parameters, are common in regression studies based on methods such as sliced inverse regression. In particular, the linearity condition holds to a reasonable approximation in many problems (Hall and Li (1993)). Nevertheless, these conditions are restrictive due to the effects of predictor transformations, and are imposed mainly for ease and simplification of theory investigation. Currently, we are not able to develop the theory in the general case. Without the linearity condition, our procedure can be interpreted as a method for estimating the canonical correlations between the transformed predictors and the basis functions generated

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0080

for the response variable, and thus the directions found by it are useful in their own right in identifying some important features of the regression. We plan to conduct theoretical research along this direction.

Let $\{(\hat{\boldsymbol{\theta}}_i, \hat{\boldsymbol{\beta}}_i), i = 1, \ldots, H\}$ denote the solution to (2.4). Our procedure produces two sets of linear combinations: $\{\boldsymbol{\Phi}\hat{\boldsymbol{\theta}}_i\}$ for the response and $\{\mathbf{X}_{\hat{\mathcal{F}}}\hat{\boldsymbol{\beta}}_i\}$ for the predictors. From the dimension-reduction viewpoint, it is always desirable to have a reduced rank solution, although this is in general a challenging task. Fortunately, under (2.10), the problem reduces to that of determining the structural dimension d_0. To this end, we propose a BIC-type criterion: let r_i^2 denote the squared correlation between $\boldsymbol{\Phi}\hat{\boldsymbol{\theta}}_i$ and $\mathbf{X}_{\hat{\mathcal{F}}}\hat{\boldsymbol{\beta}}_i$, with

$$\mathrm{BIC}_d = \frac{\sum_{i=1}^{d} r_i^4}{\sum_{i=1}^{H} r_i^4} - \frac{\log n}{n} \times \frac{d(d+1)}{2},$$

the estimated structural dimension is

$$\hat{d} = \arg\max_{1 \le d \le H} \mathrm{BIC}_d.$$

Remark 2.6. Consider the linear regression of a transformed response $T(Y)$ on $\boldsymbol{X}$. Let $R^2(T) = [\mathrm{cor}\{T(Y), \boldsymbol{X}^\top \boldsymbol{b}(T)\}]^2$, where $\boldsymbol{b}(T)$ is the least squares solution. Theorem 3.2 of Chen and Li (1998) showed that, at the population level, $R^2(T_i) = \lambda_i, i = 1, \ldots, d$. The same applies to the regression of Y on $\boldsymbol{f}_0$. Consequently, r_i^2 as defined above is an estimator of δ_i, the i-th largest eigenvalue of $\mathrm{cov}\{E(\boldsymbol{f}_0|Y)\}$. Using the equivalence of the optimal scoring problem and the eigen-decomposition problem, our BIC-type criterion is equivalent to that of Zhu et al. (2010) when $\boldsymbol{f}_0$ is assumed to be known. The degrees of freedom should be modified accordingly when $\boldsymbol{f}_0$ is unknown. However, the use of the proposed BIC-type criterion remains feasible, since the estimation of $\boldsymbol{f}_0$ is independent of the dimension d.

Corollary 2.1. *Under the conditions of Theorem 2.1, $r_i^2 - \delta_i = O_P(n^{-1/2})$ for $i = 1, \ldots, H$, and $\hat{d}$ converges to d_0 in probability.*

Throughout the numerical studies, we initialize $\boldsymbol{\beta}_i$ and $\boldsymbol{\theta}_i$ with the ordinary sliced inverse regression estimates (2.1). The BIC criterion works well, even in cases where sliced inverse

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0080

regression fails. Nevertheless, taking the identifiability of transformations into account, we recommend to the use of multiple initial values, and choice of the $\hat{\boldsymbol{f}}$ and $\hat{\mathbf{B}}$ that correspond to the smallest $\hat{d}$.

3 Simulation study

In this section, we present some simulation results to illustrate the performance of the proposed nonlinear dimension-reduction method.

Let $\boldsymbol{0}_p$ be a p-vector of zeros and $\boldsymbol{\Sigma} = (\Sigma_{ij})$ with $\Sigma_{ij} = \rho^{|i-j|}$ for $1 \leq i, j \leq p$. Throughout the simulation study, we took $n = 400, p = 10$, and $\rho = 0.5$. Five simulation examples were considered.

EXAMPLE 3.1. *Let* $\boldsymbol{\beta}_{01} = (1, 1, 0, \ldots, 0)^\top \in \mathbb{R}^p$ *and* $\boldsymbol{\beta}_{02} = (0, 0, 1, 1, 0, \ldots, 0)^\top \in \mathbb{R}^p$, *with*

$$Y = \frac{\boldsymbol{f}_0^\top \boldsymbol{\beta}_{01}}{(\boldsymbol{f}_0^\top \boldsymbol{\beta}_{02} + 1.5)^2 + 0.5} + 0.5 \times \epsilon,$$

where $\boldsymbol{f}_0 \sim N(\boldsymbol{0}_p, \boldsymbol{\Sigma}), \epsilon \sim N(0, 1)$, *and* $\boldsymbol{f}_0$ *and* ϵ *are independent. To generate* $\boldsymbol{X}$, *we explored two cases: (i)* $X_j = \text{sign}(f_{0j}) \times f_{0j}^2$ *and (ii)* $X_j = F_C^{-1}\{\Phi(f_{0j})\}$, *where* $F_C(\cdot)$ *is the standard Cauchy distribution function and* $\Phi(\cdot)$ *the standard normal.*

EXAMPLE 3.2. *We proceed as in Example 3.1 except that* $\boldsymbol{f}_0$ *has a multivariate t-distribution with location vector zero, scale matrix* $\boldsymbol{\Sigma}$, *and five degrees of freedom. We explored two cases: (i)* $X_j = \text{sign}(f_{0j}) \times f_{0j}^2$ *and (ii)* $X_j = F_C^{-1}\{F_{t_5}(f_{0j})\}$, *where* $F_{t_5}(\cdot)$ *is the common marginal distribution function of* $\boldsymbol{f}_0$.

EXAMPLE 3.3. *We proceed as in Example 3.1 except that* $\boldsymbol{f}_0 = \boldsymbol{\Sigma}^{1/2}\boldsymbol{U}$, *where* $\boldsymbol{U}$ *is uniformly distributed in the hypercube* $[-\sqrt{3}, \sqrt{3}]^p$. *To generate* $\boldsymbol{X}$, *we set* $X_j = \text{sign}(f_{0j}) \times f_{0j}^2$.

EXAMPLE 3.4. *Let* $\boldsymbol{\beta}_{01} = (1, 0, \ldots, 0)^\top \in \mathbb{R}^p$ *and* $\boldsymbol{\beta}_{02} = (0, 1, 0, \ldots, 0)^\top \in \mathbb{R}^p$. *We first generated* $\boldsymbol{X} \sim N(\boldsymbol{0}_p, \boldsymbol{\Sigma})$ *and then set* $f_{01} = (X_1^2 - 1)/\sqrt{2}$ *and* $f_{0j} = X_j$ *for* $j = 2, \ldots, p$.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0080

Here

$$Y = \frac{\boldsymbol{f}_0^\top \boldsymbol{\beta}_{01}}{(\boldsymbol{f}_0^\top \boldsymbol{\beta}_{02} + 1.5)^2 + 0.5} + 0.5 \times \epsilon,$$

where $\epsilon \sim N(0,1)$ *is independent of* $\boldsymbol{X}$.

EXAMPLE 3.5. *We proceed as in Example 3.4 except that*

$$Y = (\boldsymbol{f}_0^\top \boldsymbol{\beta}_{01} + 1) \times \boldsymbol{f}_0^\top \boldsymbol{\beta}_{02} + 0.5 \times \epsilon.$$

In these examples, $\mathbf{B}_0 = (\boldsymbol{\beta}_{01}, \boldsymbol{\beta}_{02})$ and the structural dimension $d_0 = 2$. In Examples 3.1-3.3, each component $f_{0j}(X_j)$ is a monotone increasing function of X_j; in Examples 3.4 and 3.5, $f_{01}(X_1)$ is symmetric in X_1 and $\mathbf{B}_0$ is a sub-matrix of $\mathbf{I}_p$.

Examples 3.1-3.3 have also been used in Wang et al. (2014), in which probability integral transformed sliced inverse regression (T-SIR) was shown to perform comparatively to the oracle sliced inverse regression (O-SIR) assuming $\boldsymbol{f}_0$ is known. The T-SIR is based on the assumption that $\boldsymbol{f}_0$ is multivariate Gaussian, $f_{0j} = \Phi^{-1}\{F_j(X_j)\}$ with $F_j(\cdot)$ the marginal distribution function of X_j. This assumption holds in Example 3.1, but is violated in Examples 3.2-3.5. Since the probability integral transformation is monotonic, we expect T-SIR to fail in Examples 3.4 and 3.5. The linearity condition on the distribution of $\boldsymbol{f}_0$ holds in Examples 3.1 and 3.2, but is not satisfied in Examples 3.3-3.5.

We also examined the performance of T-SIR and O-SIR. For each competitor, we adopted slice indicator functions as transformation functions of the response and fixed the number of slices, K, at 10. The numerical results in Wang and Zhu (2013) show that slice indicator functions are quite competitive with B-spline basis functions. In our flexible dimension-reduction method (FDR), two values of the working dimension H were explored: $H = 4$ and $H = K$. The corresponding methods are denoted by P-FDR and F-FDR, respectively. F-FDR is computationally more intensive than P-FDR. For each example, we generated 200 replications.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0080

For any matrix $\mathbf{A}$, let $\mathbf{A}_o$ be the orthonormalized version of $\mathbf{A}$. Assume, for the moment, that $d_0 = 2$ is known. To evaluate the accuracy of an estimator $\text{span}(\hat{\mathbf{B}})$ of $\text{span}(\mathbf{B}_0)$, we use both the vector correlation coefficient, defined by $(\prod_{l=1}^{2} \phi_l^2)^{1/2}$, and the trace correlation coefficient, defined by $(2^{-1}\sum_{l=1}^{2} \phi_l^2)^{1/2}$, where $1 \geq \phi_1^2 \geq \phi_2^2 \geq 0$ are the eigenvalues of the matrix $\hat{\mathbf{B}}_o^\top \mathbf{B}_o \mathbf{B}_o^\top \hat{\mathbf{B}}_o$. To assess how well T-SIR and FDR estimate $\boldsymbol{f}_0$, we calculated the sample Pearson's correlation coefficient between $\hat{f}_j$ and f_{0j}. For simplicity, we concentrate on the set of active transformations, $\{f_{01}, f_{02}, f_{03}, f_{04}\}$ in Examples 3.1-3.3, and $\{f_{01}, f_{02}\}$ in Examples 3.4 and 3.5. The absolute value of this measure is reported for f_{01} in Examples 3.4 and 3.5. To assess the accuracy of each method in terms of dimension reduction, we used the multiple correlation coefficient (Li and Dong (2009)). Suppose $\boldsymbol{V}_1$ and $\boldsymbol{V}_2$ are two d-dimensional random vectors. Let $\Sigma_{\boldsymbol{V}_i} = \text{cov}(\boldsymbol{V}_i), i = 1, 2$ and $\Sigma_{\boldsymbol{V}_1\boldsymbol{V}_2} = \text{cov}(\boldsymbol{V}_1, \boldsymbol{V}_2)$. The multiple correlation coefficient between $\boldsymbol{V}_1$ and $\boldsymbol{V}_2$ is defined as $\{d^{-1}\text{trace}\left(\Sigma_{\boldsymbol{V}_1}^{-1}\Sigma_{\boldsymbol{V}_1\boldsymbol{V}_2}\Sigma_{\boldsymbol{V}_2}^{-1}\Sigma_{\boldsymbol{V}_1\boldsymbol{V}_2}^\top\right)\}^{1/2}$. We employed the sample version of this measure based on $\mathbf{X}_{\hat{\boldsymbol{f}}}\hat{\mathbf{B}}$ and $\mathbf{X}_{\boldsymbol{f}}\mathbf{B}_0$.

The means and standard deviations of the various measures are summarized in Tables 1-5. Not surprisingly, O-SIR performs very well in all five examples. The relative insensitivity of sliced inverse regression to the non-linearity of the distribution of the predictors has long been observed in the sufficient dimension-reduction literature. We also see that in terms of subspace estimation, the performance of FDR is similar to that of O-SIR. However, FDR is outperformed by O-SIR because FDR involves the estimation of p component functions, some of which are redundant. Generally, the estimates of the active component functions are more accurate for monotonic functions than for symmetric ones. The T-SIR performs comparably to O-SIR in Examples 3.1-3.3. Here the degree of the similarity between the true transformations and the corresponding probability integral transformations is very high: in Tables 1-3 the average Pearson's correlation coefficients of T-SIR are very close to 1. The T-SIR fails in Examples 3.4 and 3.5, where probability integral transformations severely misspecify transformations that are symmetric. From Tables 4 and 5, the Pearson's correlation coefficients of T-SIR are almost zero for symmetric component functions. The T-SIR inherits

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0080

the drawback of sliced inverse regression by failing to capture any symmetrical structure. Generally, F-FDR outperforms P-FDR, but the differences are often small. Unreported results show that the estimation accuracy of FDR improves as the sample size increases.

The performance of each competitor here relies heavily on the correct determination of the structural dimension d_0. Tables 6-10 shows the results of applying the BIC-type criterion. There, O-SIR-BIC and FDR-BIC are roughly equally powerful, with O-SIR-BIC performing slightly better, while T-SIR-BIC has an alarmingly low rate of correctly identifying the true dimension d_0 in Examples 3.4 and 3.5.

4 Data analysis

We applied T-SIR and F-FDR to the Ozone data used in Breiman and Friedman (1985). The goal is to study the relationship between atmospheric ozone concentration and meteorology in the Los Angeles basin. The dataset, available from the R package **mlbench**, consists of daily measurements of maximum one-hour-average ozone reading (Y) and eight meteorological variables for $n = 330$ days, in 1976. The $p = 8$ predictors used in the study were the Sandburg Air Force Base temperature (X_1), inversion base height (X_2), Daggett pressure gradient (X_3), visibility (X_4), Vandenburg 500 millibar height (X_5), humidity (X_6), inversion base temperature (X_7), and wind speed (X_8). All the predictors were scaled to the unit interval $[0, 1]$. Let $\boldsymbol{X} = (X_1, \ldots, X_8)^\top$.

Using the BIC-type criterion, both T-SIR and F-FDR find two directions. The coefficient estimates from either method, denoted by $\hat{\boldsymbol{\beta}}_i$ for $i = 1, 2$, are shown in Table 11. The relatively important predictors are X_1, X_2, X_3, X_5, X_6, and X_7. Let $Z_i = \hat{\boldsymbol{\beta}}_i^\top \boldsymbol{X}$. The first T-SIR predictor and the first F-FDR predictor are almost identical to the predictor found by the linear least squares fitting (not shown). A linear trend is visible from Figure 1. To check

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0080

whether the second predictor Z_2 from each method has a significant effect on the response Y, we considered

$$E(Y|Z_1, Z_2) = g_1(Z_1) + g_2(Z_2) + g_{12}(Z_1, Z_2),$$

where $g_1(\cdot)$ and $g_2(\cdot)$ are smooth main effect functions, and $g_{12}(\cdot, \cdot)$ is a smooth interaction term. We used the **gam** function from the R package **mgcv** to fit this model, and found that for both T-SIR and F-FDR, $g_1(Z_1)$ is highly significant (p-values $< 10^{-15}$) and $g_2(Z_2)$ is not significant (p-values > 0.1). At the 0.01 significance level, the interaction term $g_{12}(Z_1, Z_2)$ for F-FDR is significant (p-value $= 0.004$), while that for T-SIR is not (p-value $= 0.021$). The adjusted R^2 values were 71.6% and 75% for T-SIR and F-FDR, respectively. Using a full two-dimensional smooth did not improve the fit by much (not shown). The second T-SIR direction is likely to be spurious. For F-FDR, the presence of the interaction term is the key to the estimated structural dimension of 2. The adjusted R^2 value from fitting an additive model is 73.8%. Given the transformations of predictors, additive models are essentially one-dimensional. This is also true for alternating conditional expectations of Breiman and Friedman (1985). As a dimension-reduction method, our method is more flexible. Figure 2 shows the estimated transformations from F-FDR of the six important predictors. We can see that the transformation function of X_5 (or X_6) is symmetric over at least some of the range of observation. This suggests that there may be some symmetric pattern that T-SIR fails to handle. Li (1992) used the same dataset to demonstrate how other dimension-reduction methods are needed to complement the original SIR in symmetric cases. We used $K = 10$ slices for T-SIR and F-FDR, but other choices yield almost identical scenes.

5 Appendix

Proof of Proposition 2.1. Without loss of generality, assume that $\mathbf{B} = (\mathbf{I}_d, \mathbf{B}_l^\top)^\top$, where $\mathbf{B}_l$ is the lower $(p-d) \times d$ sub-matrix of $\mathbf{B}$. By definition, the conditional distribution

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0080

of Y given $\mathbf{B}^\top \boldsymbol{f}(\boldsymbol{X})$ equals that of Y given $\boldsymbol{X}$:

$$F\{y \mid \mathbf{B}^\top \boldsymbol{f}(\boldsymbol{X})\} = F(y \mid \boldsymbol{X}).$$

If span($\mathbf{B}$) is not identifiable, then there is another set of univariate functions $\{g_1, \ldots, g_p\}$ and a $p \times d$ matrix $\mathbf{M}$, such that $F\{y \mid \mathbf{M}^\top \boldsymbol{g}(\boldsymbol{X})\} = F(y \mid \boldsymbol{X})$. Write $\mathbf{B} = (B_{ij})$ and $\mathbf{M} = (M_{ij})$. Then

$$\begin{aligned} F\{y \mid \mathbf{B}^\top \boldsymbol{f}(\boldsymbol{X})\} &= h_B\{f_1(x_1) + B_{(1+d)1} f_{1+d}(x_{1+d}) + \cdots + B_{p1} f_p(x_p), \ldots, \\ &\qquad f_d(x_d) + B_{(1+d)d} f_{1+d}(x_{1+d}) + \cdots + B_{pd} f_p(x_p), y\}, \\ F\{y \mid \mathbf{M}^\top \boldsymbol{g}(\boldsymbol{X})\} &= h_M\{g_1(x_1) + M_{(1+d)1} g_{1+d}(x_{1+d}) + \cdots + M_{p1} g_p(x_p), \ldots, \\ &\qquad g_d(x_d) + M_{(1+d)d} g_{1+d}(x_{1+d}) + \cdots + M_{pd} g_p(x_p), y\}. \end{aligned}$$

Here, h_B and h_M are $(d+1)$-dimensional functions. Consequently,

$$\begin{aligned} & h_B\{f_1(x_1) + B_{(1+d)1} f_{1+d}(x_{1+d}) + \cdots + B_{p1} f_p(x_p), \ldots, \\ & \qquad f_d(x_d) + B_{(1+d)d} f_{1+d}(x_{1+d}) + \cdots + B_{pd} f_p(x_p), y\} \\ = \; & h_M\{g_1(x_1) + M_{(1+d)1} g_{1+d}(x_{1+d}) + \cdots + M_{p1} g_p(x_p), \ldots, \\ & \qquad g_d(x_d) + M_{(1+d)d} g_{1+d}(x_{1+d}) + \cdots + M_{pd} g_p(x_p), y\}. \end{aligned}$$

Taking derivatives with respect to $x_1, \ldots, x_p$ on both sides, we get $h'_{Bj} f'_j = h'_{Mj} g'_j$, for $j = 1, \ldots, d$, and

$$(h'_{B1} B_{(j+d)1} + \cdots + h'_{Bd} B_{(j+d)d}) f'_{j+d} = (h'_{M1} M_{(j+d)1} + \cdots + h'_{Md} M_{(j+d)d}) g'_{j+d},$$

for $j = 1, \ldots, p-d$, where h'_{Bj} stands for the derivative of h_B with respect to the jth argument, and similarly for h'_{Mj}. Write $\boldsymbol{h}'_B = (h'_{B1}, \ldots, h'_{Bd})^\top$ and $\boldsymbol{h}'_M = (h'_{M1}, \ldots, h'_{Md})^\top$. Set $\mathbf{D}_{f1} = \text{diag}\{f'_1, \ldots, f'_d\}, \mathbf{D}_{f2} = \text{diag}\{f'_{1+d}, \ldots, f'_p\}, \mathbf{D}_{g1} = \text{diag}\{g'_1, \ldots, g'_d\}$, and $\mathbf{D}_{g2} = \text{diag}\{g'_{1+d}, \ldots, g'_p\}$. Then $\mathbf{D}_{f1} \boldsymbol{h}'_B = \mathbf{D}_{g1} \boldsymbol{h}'_M$ and $\mathbf{D}_{f2} \mathbf{B}_l \boldsymbol{h}'_B = \mathbf{D}_{g2} \mathbf{M}_l \boldsymbol{h}'_M$. By the condition of Proposition 2.1, for $\boldsymbol{x}$ in a neighborhood of $\boldsymbol{x}^*$,

$$\mathbf{B}_l \boldsymbol{h}'_B = \mathbf{D}_{f2}^{-1} \mathbf{D}_{g2} \mathbf{M}_l \mathbf{D}_{g1}^{-1} \mathbf{D}_{f1} \boldsymbol{h}'_B.$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0080

Hence,

$$(\mathbf{B}_l - \mathbf{M}_l)\boldsymbol{h}_B' + (\mathbf{M}_l - \mathbf{D}_{f2}^{-1}\mathbf{D}_{g2}\mathbf{M}_l\mathbf{D}_{g1}^{-1}\mathbf{D}_{f1})\boldsymbol{h}_B' = \mathbf{0}.$$

Again by the condition of Proposition 2.1, as $\boldsymbol{x} \to \boldsymbol{x}^*$, $(\mathbf{M}_l - \mathbf{D}_{f2}^{-1}\mathbf{D}_{g2}\mathbf{M}_l\mathbf{D}_{g1}^{-1}\mathbf{D}_{f1}) \to \mathbf{0}$, and hence $(\mathbf{B}_l - \mathbf{M}_l)\boldsymbol{h}_B' = \mathbf{0}$. Since $d = d_0$, none of the components of $\boldsymbol{h}_B'$ are zero, and there is at most one constant component. This implies that $\mathbf{B}_l = \mathbf{M}_l$. The proof is complete.

Proof of Theorem 2.1. Without loss of generality, assume that $\text{cov}(\boldsymbol{f}) = \mathbf{C}\text{cov}(\boldsymbol{g})\mathbf{C}^\top = \mathbf{I}_p$. By the equivalence of linear discriminant analysis and optimal scoring (Hastie et al. (1995)), our method amounts to estimating $\text{cov}\{E(\boldsymbol{g}|Y)\}$ and then maximizing

$$\text{trace}[\text{cov}\{E(\boldsymbol{f}|Y)\}] = \text{trace}[\mathbf{C}\text{cov}\{E(\boldsymbol{g}|Y)\}\mathbf{C}^\top]$$

with respect to $\mathbf{C}$.

From (2.10) we know that $Y \perp\!\!\!\perp \boldsymbol{g}|\mathbf{B}_0^\top\mathbf{C}_0\boldsymbol{g}$. Then

$$E(\boldsymbol{f}|Y) = E\{E(\boldsymbol{f}|Y, \mathbf{B}_0^\top\mathbf{C}_0\boldsymbol{g})|Y\} = \mathbf{C}E\{E(\boldsymbol{g}|\mathbf{B}_0^\top\mathbf{C}_0\boldsymbol{g})|Y\}.$$

Under the linearity condition (A1), $E(\boldsymbol{g}|\mathbf{B}_0^\top\mathbf{C}_0\boldsymbol{g}) = \text{cov}(\boldsymbol{g})\mathbf{C}_0^\top\mathbf{B}_0(\mathbf{B}_0^\top\mathbf{B}_0)^{-1}\mathbf{B}_0^\top\mathbf{C}_0\boldsymbol{g}$. Hence,

$$E(\boldsymbol{f}|Y) = \mathbf{C}\text{cov}(\boldsymbol{g})\mathbf{C}_0^\top\mathbf{B}_0(\mathbf{B}_0^\top\mathbf{B}_0)^{-1}\mathbf{B}_0^\top E(\boldsymbol{f}_0|Y).$$

Setting $\mathbf{C} = \mathbf{C}_0$ yields $E(\boldsymbol{f}_0|Y) = \mathbf{B}_0(\mathbf{B}_0^\top\mathbf{B}_0)^{-1}\mathbf{B}_0^\top E(\boldsymbol{f}_0|Y)$. This implies that

$$\text{cov}\{E(\boldsymbol{f}|Y)\} = \mathbf{C}\text{cov}(\boldsymbol{g})\mathbf{C}_0^\top\text{cov}\{E(\boldsymbol{f}_0|Y)\}\mathbf{C}_0\text{cov}(\boldsymbol{g})\mathbf{C}^\top.$$

It is easy to check that

$$\begin{aligned} & \text{trace}[\mathbf{C}\text{cov}(\boldsymbol{g})\mathbf{C}_0^\top\text{cov}\{E(\boldsymbol{f}_0|Y)\}\mathbf{C}_0\text{cov}(\boldsymbol{g})\mathbf{C}^\top] \\ \leq\ & \text{trace}[\{\text{cov}(\boldsymbol{g})\}^{1/2}\mathbf{C}_0^\top\text{cov}\{E(\boldsymbol{f}_0|Y)\}\mathbf{C}_0\{\text{cov}(\boldsymbol{g})\}^{1/2}] \\ =\ & \text{trace}[\text{cov}\{E(\boldsymbol{f}_0|Y)\}]. \end{aligned}$$

Consequently, one population solution of $\mathbf{C}$ is $\boldsymbol{\Gamma}\mathbf{C}_0$ and, by the coverage condition (A2), the corresponding solution of $\text{span}\{\mathbf{B}\}$ is $\text{span}\{\boldsymbol{\Gamma}\mathbf{B}_0\}$. At the sample level, following Li (1991)

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0080

and Fung et al. (2002), it is easily verified that $\widehat{\text{cov}}\{E(\boldsymbol{g}|Y)\} - \text{cov}\{E(\boldsymbol{g}|Y)\} = O_P(n^{-1/2})$. This completes the proof.

Proof of Corollary 2.1. The first part follows from Theorem 2.1. For a proof of the second part, see Zhu et al. (2010).

REFERENCES

Box, G. E. and Cox, D. R. (1964). An analysis of transformations. *J. Roy. Statist. Soc. Ser. B* **26**, 211-252.

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

Chen, C.-H. and Li, K.-C. (1998). Can SIR be as popular as multiple linear regression? *Statist. Sinica* **8**, 289-316.

Cook, R. D. (1998). *Regression graphics: Ideas for studying regressions through graphics.* John Wiley & Sons, New York.

Duan, N. and Li, K.-C. (1991). Slicing regression: a link-free regression method. *Ann. Statist.* **19**, 505-530.

Fung, W. K., He, X., Liu, L. and Shi, P. (2002). Dimension reduction based on canonical correlation. *Statist. Sinica* **12**, 1093-1114.

Hall, P. and Li, K.-C. (1993). On almost linearity of low dimensional projections from high dimensional data. *Ann. Statist.* **21**, 867-889.

Hastie, T., Buja, A. and Tibshirani, R. (1995). Penalized discriminant analysis. *Ann. Statist.* **23**, 73-102.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0080

Hastie, T., Tibshirani, R. and Buja, A. (1994). Flexible discriminant analysis by optimal scoring. *J. Amer. Statist. Assoc.* **89**, 1255-1270.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer, New York.

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

Kent, J. T. (1991). Comment. *J. Amer. Statist. Assoc.* **86**, 336-337.

Li, B. and Dong, Y. (2009). Dimension reduction for nonelliptically distributed predictors. *Ann. Statist.* **37**, 1272-1298.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *J. Amer. Statist. Assoc.* **86**, 316-327.

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

Ramsay, J. O. (1988). Monotone regression splines in action. *Statist. Sci.* **3**, 425-441.

Wang, T., Guo, X., Xu, P. and Zhu, L. (2014). Transformed sufficient dimension reduction. *Biometrika* **101**, 815-829.

Wang, T. and Zhu, L. (2013). Sparse sufficient dimension reduction using optimal scoring. *Comput. Statist. Data Anal.* **57**, 223-232.

Wood, S. (2006). *Generalized Additive Models: An Introduction with R*. Chapman & Hall, Boca Raton.

Wu, Y. and Li, L. (2011). Asymptotic properties of sufficient dimension reduction with a diverging number of predictors. *Statist. Sinica* **2011**, 707-730.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0080

Yin, X. and Li, B. (2011). Sufficient dimension reduction based on an ensemble of minimum average variance estimators. *Ann. Statist.* **39**, 3392-3416.

Zhu, L., Wang, T., Zhu, L. and Ferré, L. (2010). Sufficient dimension reduction through discretization-expectation estimation. *Biometrika* **97**, 295-304.

Table 1: Means and standard deviations (in parentheses) of the vector correlation coefficient (VCC) and the trace correlation coefficient (TCC) for subspace estimation, the Pearson's correlation coefficient (PCC) for component estimation and the multiple correlation coefficient (MCC) for dimension reduction, based on 200 data replications, are reported for Example 3.1

		VCC	TCC	PCC ($\hat{f}_1$)	PCC ($\hat{f}_2$)	PCC ($\hat{f}_3$)	PCC ($\hat{f}_4$)	MCC
Case (i)	O-SIR	0.817 (0.081)	0.910 (0.037)					0.956 (0.021)
	T-SIR	0.825 (0.077)	0.914 (0.036)	0.998 (0.001)	0.998 (0.001)	0.998 (0.001)	0.997 (0.001)	0.956 (0.020)
	P-FDR	0.815 (0.078)	0.909 (0.036)	0.960 (0.026)	0.932 (0.043)	0.872 (0.104)	0.867 (0.104)	0.878 (0.032)
	F-FDR	0.804 (0.093)	0.904 (0.040)	0.959 (0.030)	0.929 (0.048)	0.889 (0.094)	0.882 (0.093)	0.882 (0.032)
Case (ii)	O-SIR	0.825 (0.083)	0.914 (0.038)					0.959 (0.019)
	T-SIR	0.830 (0.080)	0.916 (0.037)	0.998 (0.001)	0.998 (0.001)	0.998 (0.001)	0.998 (0.001)	0.959 (0.018)
	P-FDR	0.824 (0.086)	0.913 (0.039)	0.961 (0.024)	0.931 (0.042)	0.880 (0.097)	0.858 (0.099)	0.880 (0.032)
	F-FDR	0.822 (0.090)	0.912 (0.041)	0.960 (0.026)	0.929 (0.045)	0.898 (0.087)	0.877 (0.089)	0.887 (0.032)

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0080

Table 2: Means and standard deviations (in parentheses) of the vector correlation coefficient (VCC) and the trace correlation coefficient (TCC) for subspace estimation, the Pearson's correlation coefficient (PCC) for component estimation and the multiple correlation coefficient (MCC) for dimension reduction, based on 200 data replications, are reported for Example 3.2

		VCC	TCC	PCC ($\hat{f}_1$)	PCC ($\hat{f}_2$)	PCC ($\hat{f}_3$)	PCC ($\hat{f}_4$)	MCC
Case (i)	O-SIR	0.709 (0.128)	0.859 (0.056)					0.926 (0.036)
	T-SIR	0.782 (0.095)	0.893 (0.043)	0.979 (0.015)	0.979 (0.018)	0.977 (0.018)	0.979 (0.013)	0.926 (0.029)
	P-FDR	0.788 (0.094)	0.895 (0.043)	0.913 (0.040)	0.890 (0.046)	0.804 (0.111)	0.781 (0.131)	0.813 (0.043)
	F-FDR	0.786 (0.101)	0.895 (0.045)	0.912 (0.041)	0.890 (0.045)	0.827 (0.103)	0.804 (0.117)	0.821 (0.041)
Case (ii)	O-SIR	0.719 (0.127)	0.863 (0.055)					0.924 (0.038)
	T-SIR	0.789 (0.091)	0.896 (0.041)	0.977 (0.015)	0.976 (0.018)	0.978 (0.015)	0.979 (0.014)	0.924 (0.028)
	P-FDR	0.788 (0.089)	0.895 (0.042)	0.909 (0.033)	0.881 (0.052)	0.782 (0.129)	0.803 (0.118)	0.808 (0.040)
	F-FDR	0.784 (0.108)	0.894 (0.048)	0.909 (0.033)	0.884 (0.049)	0.819 (0.117)	0.814 (0.115)	0.818 (0.040)

Table 3: Means and standard deviations (in parentheses) of the vector correlation coefficient (VCC) and the trace correlation coefficient (TCC) for subspace estimation, the Pearson's correlation coefficient (PCC) for component estimation and the multiple correlation coefficient (MCC) for dimension reduction, based on 200 data replications, are reported for Example 3.3

	VCC	TCC	PCC ($\hat{f}_1$)	PCC ($\hat{f}_2$)	PCC ($\hat{f}_3$)	PCC ($\hat{f}_4$)	MCC
O-SIR	0.844 (0.069)	0.923 (0.032)					0.962 (0.017)
T-SIR	0.801 (0.086)	0.902 (0.040)	0.988 (0.002)	0.991 (0.001)	0.991 (0.002)	0.991 (0.002)	0.949 (0.019)
P-FDR	0.836 (0.086)	0.919 (0.038)	0.983 (0.019)	0.947 (0.042)	0.895 (0.084)	0.885 (0.088)	0.899 (0.031)
F-FDR	0.831 (0.094)	0.917 (0.042)	0.982 (0.021)	0.949 (0.043)	0.907 (0.075)	0.904 (0.079)	0.904 (0.031)

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0080

Table 4: Means and standard deviations (in parentheses) of the vector correlation coefficient (VCC) and the trace correlation coefficient (TCC) for subspace estimation, the Pearson's correlation coefficient (PCC) for component estimation and the multiple correlation coefficient (MCC) for dimension reduction, based on 200 data replications, are reported for Example 3.4

	VCC	TCC	PCC ($\hat{f}_1$)	PCC ($\hat{f}_2$)	MCC
O-SIR	0.807 (0.095)	0.906 (0.042)			0.935 (0.030)
T-SIR	0.562 (0.209)	0.787 (0.088)	0.064 (0.048)	0.998 (0.001)	0.643 (0.034)
P-FDR	0.823 (0.099)	0.914 (0.043)	0.742 (0.070)	0.906 (0.051)	0.776 (0.050)
F-FDR	0.834 (0.102)	0.919 (0.044)	0.739 (0.084)	0.908 (0.049)	0.782 (0.049)

Table 5: Means and standard deviations (in parentheses) of the vector correlation coefficient (VCC) and the trace correlation coefficient (TCC) for subspace estimation, the Pearson's correlation coefficient (PCC) for component estimation and the multiple correlation coefficient (MCC) for dimension reduction, based on 200 data replications, are reported for Example 3.5

	VCC	TCC	PCC ($\hat{f}_1$)	PCC ($\hat{f}_2$)	MCC
O-SIR	0.959 (0.017)	0.979 (0.008)			0.987 (0.005)
T-SIR	0.691 (0.216)	0.867 (0.073)	0.073 (0.053)	0.997 (0.001)	0.691 (0.008)
P-FDR	0.929 (0.049)	0.965 (0.022)	0.665 (0.140)	0.945 (0.028)	0.824 (0.043)
F-FDR	0.935 (0.049)	0.967 (0.022)	0.661 (0.155)	0.945 (0.031)	0.826 (0.047)

Table 6: Percentages of estimated structural dimensions $\hat{d}$ smaller than, equal to, and larger than d_0, based on 200 data replications, are reported for Example 3.1

		O-SIR	T-SIR	P-FDR	F-FDR
Case (i)	$\hat{d} < d_0$	0	0	0	0
	$\hat{d} = d_0$	0.99	0.995	0.94	0.98
	$\hat{d} > d_0$	0.01	0.005	0.06	0.02
Case (ii)	$\hat{d} < d_0$	0	0	0	0
	$\hat{d} = d_0$	0.99	0.995	0.93	0.975
	$\hat{d} > d_0$	0.01	0.005	0.07	0.025

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0080

Table 7: Percentages of estimated structural dimensions $\hat{d}$ smaller than, equal to, and larger than d_0, based on 200 data replications, are reported for Example 3.2

		O-SIR	T-SIR	P-FDR	F-FDR
Case (i)	$\hat{d} < d_0$	0.01	0.005	0	0
	$\hat{d} = d_0$	0.98	0.995	0.97	0.995
	$\hat{d} > d_0$	0.01	0	0.03	0.005
Case (ii)	$\hat{d} < d_0$	0.005	0.005	0	0
	$\hat{d} = d_0$	0.985	0.985	0.925	0.98
	$\hat{d} > d_0$	0.010	0.010	0.075	0.02

Table 8: Percentages of estimated structural dimensions $\hat{d}$ smaller than, equal to, and larger than d_0, based on 200 data replications, are reported for Example 3.3

	O-SIR	T-SIR	P-FDR	F-FDR
$\hat{d} < d_0$	0	0	0	0
$\hat{d} = d_0$	1	0.995	0.965	0.98
$\hat{d} > d_0$	0	0.005	0.035	0.02

Table 9: Percentages of estimated structural dimensions $\hat{d}$ smaller than, equal to, and larger than d_0, based on 200 data replications, are reported for Example 3.4

	O-SIR	T-SIR	P-FDR	F-FDR
$\hat{d} < d_0$	0	0	0	0
$\hat{d} = d_0$	0.98	0.265	0.88	0.92
$\hat{d} > d_0$	0.02	0.735	0.12	0.08

Table 10: Percentages of estimated structural dimensions $\hat{d}$ smaller than, equal to, and larger than d_0, based on 200 data replications, are reported for Example 3.5

	O-SIR	T-SIR	P-FDR	F-FDR
$\hat{d} < d_0$	0	0.89	0	0
$\hat{d} = d_0$	1	0.11	0.995	0.995
$\hat{d} > d_0$	0	0	0.005	0.005

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0080

Table 11: Ozone data. Estimated directions by T-SIR and F-FDR

	T-SIR		F-FDR	
	$\hat{\boldsymbol{\beta}}_1$	$\hat{\boldsymbol{\beta}}_2$	$\hat{\boldsymbol{\beta}}_1$	$\hat{\boldsymbol{\beta}}_2$
X_1	0.454	0.472	0.349	0.619
X_2	-0.102	-0.302	-0.230	-0.352
X_3	0.122	0.082	0.231	0.052
X_4	-0.081	-0.043	-0.110	-0.058
X_5	-0.059	-0.093	0.065	-0.173
X_6	0.129	-0.214	0.042	-0.218
X_7	0.242	-0.732	0.188	-0.784
X_8	0.012	-0.049	-0.014	-0.007

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0080

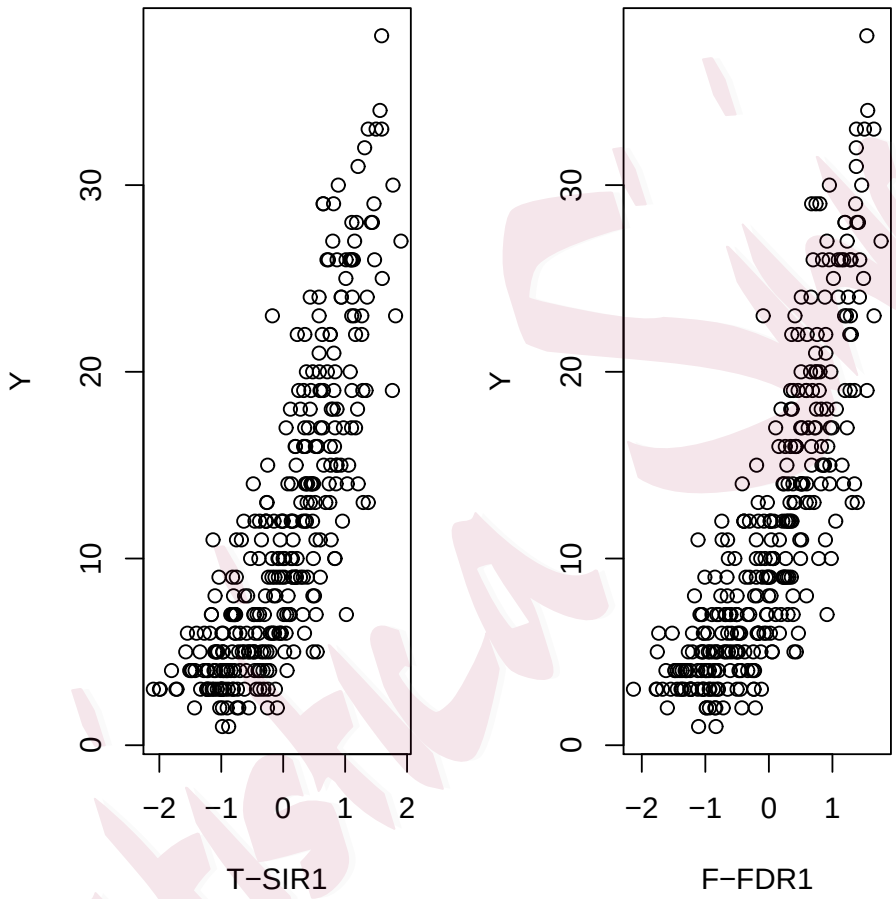

Figure 1: Ozone concentration against the first T-SIR predictor (left panel) and Ozone concentration against the first F-FDR predictor (right panel)

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0080

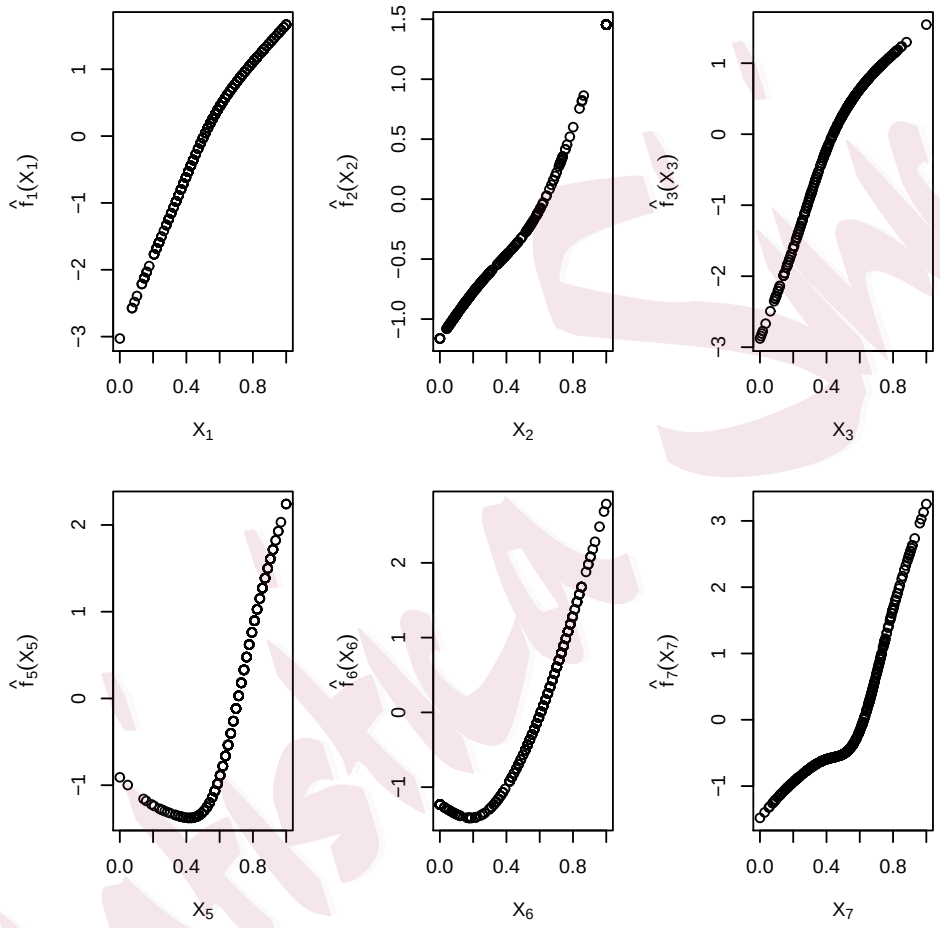

Figure 2: Plots of the estimated transformation functions by F-FDR