Statistica Sinica Preprint No: SS-2017-0532	
Title	On supervised reduction and its dual
Manuscript ID	SS-2017-0532
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202017.0532
Complete List of Authors	Peirong Xu and Tao Wang
Corresponding Author	Richard J. Samworth
E-mail	neowangtao@sjtu.edu.cn
Notice: Accepted version subject to English editing.	

Statistica Sinica

ON SUPERVISED REDUCTION AND ITS DUAL

Peirong Xu and Tao Wang

Shanghai Normal University and Shanghai Jiao Tong University

Abstract: Existing methods for dimension reduction in regression estimate a subspace in the primal predictor-based space, and then obtain the set of reduced predictors by projecting the original predictor vector onto this subspace. We propose a principled method for estimating a sufficient reduction in the dual sample-based space, on the basis of a supervised inverse regression model. Reduction is done without the need to estimate the subspace. Our method extends the duality between principal component analysis and principal coordinate analysis. We study the asymptotic behavior of the proposed method, and demonstrate that it is robust to model misspecification. We present simulation results to support the theoretical conclusion, and illustrate the application of the method in real data analysis.

Key words and phrases: Data visualization, inverse model-based reduction, multidimensional scaling, sufficient dimension reduction, supervised coordinate analysis.

1. Introduction

Dimension reduction is a long-standing and prominent problem in regression analysis (Cook, 2007). Classical methods for dimension reduction in regression transform the predictors and then fit a least squares model using the transformed variables. For example, the widely used principal component regression involves extracting the first few principal components of the predictors, and then using these components as the predictors in a linear model. One concern of principal component regression is that the directions in which the predictors show the most variation are not necessary the directions that are associated with the response. Many methods have been proposed to deal with this issue, such as partial least squares and sliced inverse regression (Li, 1991). A common goal of these methods is to reduce the dimensionality of the predictors without losing any information about the response.

Suppose we have a response $Y \in \mathbb{R}$ and a vector of predictors $\boldsymbol{X} \in \mathbb{R}^p$. Formally, the aim is to estimate a reduction $\mathcal{R} : \mathbb{R}^p \to \mathbb{R}^d, d \leq q$, such that

$$Y \perp\!\!\!\perp \boldsymbol{X} \mid \mathcal{R}(\boldsymbol{X}), \tag{1.1}$$

where $\perp\!\!\!\perp$ indicates independence. $\mathcal{R}(\boldsymbol{X})$ is called a sufficient reduction for the regression of Y onto $\boldsymbol{X}$ (Cook, 1998). Sufficient dimension reduction has been an active research area since the introduction of sliced inverse

regression and sliced average variance estimation (Cook and Weisberg, 1991). In this paper we focus on linear reductions that are linear combinations of the predictors: $\mathcal{R}(\boldsymbol{X}) = \boldsymbol{\eta}^\top \boldsymbol{X}$ for some $p \times d$ matrix $\boldsymbol{\eta}$.

Depending on the stochastic nature of Y and $\boldsymbol{X}$, there are three paradigms for determining a sufficient reduction: forward reduction, inverse reduction, and joint reduction, and they are equivalent when Y and $\boldsymbol{X}$ are jointly distributed (Cook, 2007). Without requiring a pre-specified model for $Y \mid \boldsymbol{X}$, inverse reduction is promising in regressions with many predictors. To estimate a reduction inversely, methods such as sliced inverse regression exploit properties of the conditional moments of $\boldsymbol{X} \mid Y$. These inverse moment-based methods impose constraints on the marginal distribution of $\boldsymbol{X}$. Alternatively, inverse model-based approaches directly specify a model for the inverse regression of $\boldsymbol{X}$ onto Y. Much of existing work relies on normal models. See Adragni and Cook (2009) for a recent review of inverse reduction methods.

Sufficient reduction permits us to restrict attention to a few new predictors $\boldsymbol{\eta}^\top \boldsymbol{X}$, upon which subsequent modeling and prediction can be built. Indeed, the original intent behind (1.1) is to provide a framework for dimension reduction to facilitate graphical analyses (Cook, 1998). Previous studies have largely focused on properties of estimators of the subspace spanned by

the columns of $\boldsymbol{\eta}$. However, the inference object more relevant to subsequent data analyses is not the subspace but the reduction itself. Estimating sufficient reductions is relatively new. Cook et al. (2012) studied the asymptotic behavior of a class of methods for sufficient reduction in large abundant regressions, where most predictors contribute some information on the response. In the modern "small n and large p" setting, Wang et al. (2018) recently proposed a method for estimating sparse reductions using a novel representation of sliced inverse regression. The former is inverse model-based, and the latter is inverse moment-based.

In this article, we propose a new approach for estimating a sufficient reduction. This is motivated by the well-known duality between principal component analysis and principal coordinate analysis (Gower, 1966), also known as classical multidimensional scaling (Hastie et al., 2009). Let $\mathbf{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\top \in \mathbb{R}^{n \times p}$ be the data matrix of predictor values. Without loss of generality, assume that each column of $\mathbf{X}$ has been centered to have mean zero. The singular value decomposition offers a way of expressing principal component analysis. Let $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$ be the singular value decomposition of $\mathbf{X}$; that is, $\mathbf{U} = (\mathbf{U}_1, \ldots, \mathbf{U}_p)$ is $n \times p$ with orthonormal columns, $\mathbf{V} = (\mathbf{V}_1, \ldots, \mathbf{V}_p)$ is $p \times p$ orthogonal, and $\mathbf{D}$ is a $p \times p$ diagonal matrix with diagonal entries $d_1 \geq d_2 \geq \cdots \geq d_p \geq 0$. $\mathbf{V}_j$ is called the jth principal

component direction, and $d_j\mathbf{U}_j$ the jth principal component score vector. In the terminology of dimension reduction in regression, $\mathbf{UD} = \mathbf{XV}$ are linear reductions used in principal component regression. Geometrically, each row of $\mathbf{UD}$ represents the coordinates of the corresponding row of $\mathbf{X}$ with respect to the orthonormal basis $\mathbf{V}$. In this sense, principal component analysis can be viewed as an ordination method. Indeed, it is equivalent to classical multidimensional scaling, and an alternative way of obtaining principal components is to perform an eigen-decomposition of the Gram matrix $\mathbf{XX}^\top = \mathbf{UD}^2\mathbf{U}^\top$.

Instead of estimating the directions, one can directly determine the projection coordinates of the predictor vector onto the subspace spanned by these directions. In the context of moment-based inverse reduction, Zhang et al. (2008) calculated the projection coordinates by applying classical multidimensional scaling to slice means, and then interpolated the projection of a new predictor vector using these coordinates. The method can be thought of as a dual version of sliced inverse regression. At the population level, however, it is not clear what quantity is being taking as the target. In this paper, we propose a principled method for estimating a sufficient reduction under the inverse model-based reduction scheme. Reduction is done without the need to estimate the subspace. For the first time we

study the asymptotics of predictor reduction in terms of prediction and under model misspecification.

To express the projection coordinates explicitly, an inverse regression model is introduced in Section 2, without requiring normal errors. Since the coordinates are unconstrained, sufficient reduction is achieved by classical multidimensional scaling, or principal component analysis by duality. To perform supervised reduction, we model the coordinates in a parametric way in Section 3, and extend the method of Section 2 when the error structure is known. Reduction with an unknown error structure is considered in Section 4. Some theoretical conclusions are provided. Simulation results and a real data application are presented in Section 5. A concluding discussion is given in Section 6. Proofs can be found in the Supplementary Materials.

2. A naive inverse regression model

The subspace spanned by the columns of $\boldsymbol{\eta}$ is called a dimension-reduction subspace. The parsimonious target of sufficient dimension reduction is the central subspace $\mathcal{S}_{Y|\boldsymbol{X}}$, defined as the intersection of all dimension-reduction subspaces (Cook, 1998). Let $\mathbb{Y}$ denote the sample space of Y, and let

$$\mathcal{S}_{\mathrm{E}(\boldsymbol{X}|Y)} = \mathrm{span}\{\mathrm{E}(\boldsymbol{X} \mid Y = y) - \mathrm{E}(\boldsymbol{X}), y \in \mathbb{Y}\}$$

denote the subspace spanned by the centered inverse regression curves. We

have the following proposition.

Proposition 1. *Assume (C1)* $\mathcal{S}_{\mathrm{E}(\boldsymbol{X}|Y)} \subseteq \mathrm{Var}(\boldsymbol{X})\mathcal{S}_{Y|\boldsymbol{X}}$ *and (C2)* $\mathrm{Var}(\boldsymbol{X} \mid Y)$ *is positive definite and is non-random, then*

$$\mathrm{Var}(\boldsymbol{X} \mid Y)\mathcal{S}_{Y|\boldsymbol{X}} = \mathrm{Var}(\boldsymbol{X})\mathcal{S}_{Y|\boldsymbol{X}}.$$

Conditions (C1) and (C2) are generally regarded as mild in the sufficient dimension reduction literature. Condition (C1) holds if $\mathrm{E}(\boldsymbol{X} \mid \boldsymbol{\eta}^\top \boldsymbol{X})$ is a linear function of $\boldsymbol{\eta}^\top \boldsymbol{X}$, where $\boldsymbol{\eta}$ is a basis matrix for $\mathcal{S}_{Y|\boldsymbol{X}}$. A slightly stronger condition is given by (C1') $\mathcal{S}_{\mathrm{E}(\boldsymbol{X}|Y)} = \mathrm{Var}(\boldsymbol{X})\mathcal{S}_{Y|\boldsymbol{X}}$. See Li and Wang (2007) for a good recent discussion of them.

Throughout this paper, conditions (C1') and (C2) are assumed to be true. Then, by Proposition 1,

$$\mathcal{S}_{\mathrm{E}(\boldsymbol{X}|Y)} = \boldsymbol{\Delta}\mathcal{S}_{Y|\boldsymbol{X}},$$

where $\boldsymbol{\Delta} = \mathrm{Var}(\boldsymbol{X} \mid Y)$. This implies that $\mathcal{S}_{Y|\boldsymbol{X}} = \mathrm{span}(\boldsymbol{\Delta}^{-1}\boldsymbol{\Gamma})$, where $\boldsymbol{\Gamma} \in \mathbb{R}^{p\times d}$ is a basis matrix for $\mathcal{S}_{\mathrm{E}(\boldsymbol{X}|Y)}$. Let $\boldsymbol{X}_y$ denote a random vector distributed as $\boldsymbol{X} \mid (Y = y)$. The above argument motivates the inverse regression model

$$\boldsymbol{X}_y = \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{v}_y + \boldsymbol{\epsilon}, \tag{2.2}$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^\top \in \mathbb{R}^p$, $\boldsymbol{v}_y \in \mathbb{R}^d$ is an unknown vector-valued function of y, $\boldsymbol{\epsilon}$ is a p-dimensional random vector with mean vector zero and

covariance matrix $\boldsymbol{\Delta}$, and $\boldsymbol{\epsilon}$ is independent of Y. Since $\boldsymbol{\Gamma}$ is not generally identifiable, we require that $\boldsymbol{\Delta}^{-1/2}\boldsymbol{\Gamma}$ be a $p \times d$ matrix with ortho-normal columns, that is, $\boldsymbol{\Gamma}^\top \boldsymbol{\Delta}^{-1}\boldsymbol{\Gamma}$ is the $d \times d$ identity matrix. Let $\boldsymbol{\mu}_y = \mathrm{E}(\boldsymbol{X}_y) = \mathrm{E}(\boldsymbol{X} \mid Y = y)$. Then $\boldsymbol{v}_y = \boldsymbol{\Gamma}^\top \boldsymbol{\Delta}^{-1}(\boldsymbol{\mu}_y - \boldsymbol{\mu})$. We assume that $\mathrm{Var}(\boldsymbol{v}_Y)$ is positive definite.

2.1 Reduction via classical multidimensional scaling

Assume for the moment that $\boldsymbol{\Delta}$ is known. Without loss of generality, assume that $\boldsymbol{\Delta} = \mathbf{I}_p$, the $p \times p$ identity matrix. This implies that $\boldsymbol{\Gamma}$ is a semi-orthogonal matrix, and $\mathcal{S}_{Y|\boldsymbol{X}} = \mathrm{span}(\boldsymbol{\Gamma})$. Otherwise, multiply both sides of equation (2.2) by $\boldsymbol{\Delta}^{-1/2}$ and replace $(\boldsymbol{X}_y, \boldsymbol{\mu}, \boldsymbol{\Gamma}, \boldsymbol{\epsilon}, \mathcal{S}_{Y|\boldsymbol{X}})$ by $(\boldsymbol{\Delta}^{-1/2}\boldsymbol{X}_y, \boldsymbol{\Delta}^{-1/2}\boldsymbol{\mu}, \boldsymbol{\Delta}^{-1/2}\boldsymbol{\Gamma}, \boldsymbol{\Delta}^{-1/2}\boldsymbol{\epsilon}, \boldsymbol{\Delta}^{1/2}\mathcal{S}_{Y|\boldsymbol{X}})$.

Suppose that the data consist of n independent observations, $\boldsymbol{x}_{y_1}, \ldots, \boldsymbol{x}_{y_n}$. For two observations indexed by y and y', define $d_{yy'} = \|\boldsymbol{\mu}_y - \boldsymbol{\mu}_{y'}\|_2^2$. We have

$$d_{yy'} = \|\boldsymbol{\Gamma}\boldsymbol{v}_y - \boldsymbol{\Gamma}\boldsymbol{v}_{y'}\|_2^2 = \boldsymbol{v}_y^\top \boldsymbol{v}_y - 2\boldsymbol{v}_y^\top \boldsymbol{v}_{y'} + \boldsymbol{v}_{y'}^\top \boldsymbol{v}_{y'}.$$

Let $\mathbf{D} = (d_{yy'}) \in \mathbb{R}^{n\times n}$, $\boldsymbol{w} = (\boldsymbol{v}_{y_1}^\top \boldsymbol{v}_{y_1}, \ldots, \boldsymbol{v}_{y_n}^\top \boldsymbol{v}_{y_n})^\top \in \mathbb{R}^n$, and $\mathbf{V} = (\boldsymbol{v}_{y_1}, \ldots, \boldsymbol{v}_{y_n})^\top \in \mathbb{R}^{n\times d}$. In matrix notation

$$\mathbf{D} = \boldsymbol{w}\mathbf{1}_n^\top + \mathbf{1}_n \boldsymbol{w}^\top - 2\mathbf{V}\mathbf{V}^\top,$$

where $\mathbf{1}_n$ is the n-vector of ones. Let $\mathbf{P}_n = \mathbf{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n^\top$. Then

$$\mathbf{P}_n\mathbf{D}\mathbf{P}_n = -2\mathbf{P}_n\mathbf{V}\mathbf{V}^\top\mathbf{P}_n = -2\mathbf{V}\mathbf{V}^\top,$$

and hence

$$\mathbf{V}\mathbf{V}^\top = -\frac{1}{2}\mathbf{P}_n\mathbf{D}\mathbf{P}_n.$$

Here, without loss of generality, we assume that the columns of $\mathbf{V}$ are centered, that is, $\sum_{i=1}^n \boldsymbol{v}_{y_i}$ is the d-vector of zeros.

Since $d_{yy'}$ is actually unknown, we replace it by $\hat{d}_{yy'} = \|\boldsymbol{x}_y - \boldsymbol{x}_{y'}\|_2^2 - 2p$. It is easy to show that $\hat{d}_{yy'}$ is an unbiased estimate of $d_{yy'}$. Let $\hat{\mathbf{D}} = (\hat{d}_{yy'})$ and $\mathbf{X} = (\boldsymbol{x}_{y_1}, \ldots, \boldsymbol{x}_{y_n})^\top \in \mathbb{R}^{n\times p}$. Then,

$$\mathbf{V}\mathbf{V}^\top \approx -\frac{1}{2}\mathbf{P}_n\hat{\mathbf{D}}\mathbf{P}_n = \mathbf{P}_n\mathbf{X}\mathbf{X}^\top\mathbf{P}_n.$$

Write the eigen-decomposition of $\mathbf{P}_n\mathbf{X}\mathbf{X}^\top\mathbf{P}_n$ as

$$\mathbf{P}_n\mathbf{X}\mathbf{X}^\top\mathbf{P}_n = \sum_{i=1}^n \lambda_i\boldsymbol{\alpha}_i\boldsymbol{\alpha}_i^\top,$$

where $\lambda_1 \geq \cdots \geq \lambda_n \geq 0$ are the eigenvalues, and $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_n$ are the corresponding eigenvectors. By the Eckart–Young theorem, a solution for $\mathbf{V}$ is given by

$$\tilde{\mathbf{V}} = (\lambda_1^{1/2}\boldsymbol{\alpha}_1, \ldots, \lambda_d^{1/2}\boldsymbol{\alpha}_d).$$

Write $\tilde{\mathbf{V}} = (\tilde{\boldsymbol{v}}_{y_1}, \ldots, \tilde{\boldsymbol{v}}_{y_n})^\top$. In the statistics literature, the reduction from $\boldsymbol{x}_y$ to $\tilde{\boldsymbol{v}}_y$ is known as the classical multidimensional scaling or principal

coordinate analysis. We can view $\tilde{\boldsymbol{v}}_y$ as the vector of coordinates of $\boldsymbol{x}_y$ with respect to the ortho-normal basis $\boldsymbol{\Gamma}$. From the viewpoint of dimension reduction in regression, $\tilde{\mathbf{V}}$ then contains all the regression information on the response. In subsequent analyses, graphical displays and regression methods can be exploited to examine the relationship between the response and the vector of coordinates.

As such, it is important to predict the coordinates of a new observation, $\boldsymbol{x}_{y^*}, y^* \in \mathbb{Y}$. This can be done by the classical method of adding a point to vector diagrams (Gower, 1968; Zhang et al., 2008). For each $i \in \{1, \ldots, n\}$, we define $\tilde{s}_i = \|\tilde{\boldsymbol{v}}_{y_i}\|_2^2 - \|\boldsymbol{x}_{y^*} - \boldsymbol{x}_{y_i}\|_2^2$. Let $\tilde{\boldsymbol{s}} = (\tilde{s}_1, \ldots, \tilde{s}_n)^\top \in \mathbb{R}^n$. The predicted coordinates $\tilde{\boldsymbol{v}}_{y^*}$ of $\boldsymbol{x}_{y^*}$ is then

$$\tilde{\boldsymbol{v}}_{y^*} = \frac{1}{2}(\tilde{\mathbf{V}}^\top \tilde{\mathbf{V}})^{-1} \tilde{\mathbf{V}}^\top \tilde{\boldsymbol{s}}. \tag{2.3}$$

In the classical sufficient reduction, one is interested mainly in the matrix $\boldsymbol{\Gamma}$ or the subspace $\mathcal{S}_{Y|\boldsymbol{X}}$ spanned by it. The above procedure operates in the space of coordinates of $\boldsymbol{x}_y$ with respect to the ortho-normal basis $\boldsymbol{\Gamma}$. It achieves dimension reduction while at the same time avoiding the estimation of $\boldsymbol{\Gamma}$, and is thus appealing.

2.2 Subspace estimation

Once $\boldsymbol{v}_y$ is found, it becomes natural to use least squares for estimating $\boldsymbol{\Gamma}$ in model (2.2), if desired. Specifically, we estimate $\boldsymbol{\Gamma}$ by minimizing the residual sum-of-squares

$$\|\mathbf{P}_n\mathbf{X} - \tilde{\mathbf{V}}\boldsymbol{\Gamma}^\top\|_F^2.$$

Here, $\|\cdot\|_F$ denotes the Frobenius matrix norm. The minimizer can be shown to be

$$\tilde{\boldsymbol{\Gamma}} = \mathbf{X}^\top\tilde{\mathbf{V}}(\tilde{\mathbf{V}}^\top\tilde{\mathbf{V}})^{-1}. \tag{2.4}$$

Write $\tilde{\boldsymbol{\Gamma}} = (\tilde{\boldsymbol{\Gamma}}_1, \ldots, \tilde{\boldsymbol{\Gamma}}_d)$. The estimate of $\mathcal{S}_{Y|\boldsymbol{X}}$ is then given by $\text{span}(\tilde{\boldsymbol{\Gamma}})$.

After some further manipulations, $\tilde{\boldsymbol{\Gamma}}_j$ can be shown to equal the jth the principal component direction of $\mathbf{P}_n\mathbf{X}$, and thus the first d principal component score vectors of $\mathbf{P}_n\mathbf{X}$ produces a sufficient reduction. Consequently, our method coincides with that of Cook (2007) under a PC regression model. The PC regression model is the same as the inverse regression model (2.2), except that $\boldsymbol{\epsilon}$ is assumed to be normally distributed, and the estimation under this model is by the method of maximum likelihood.

2.3 A toy example

Before we proceed, let us consider a simple simulation with $p = 5$ and $d = 2$. Observations on $(\boldsymbol{X}, Y)$ were generated from the inverse regression model (2.2) as follows. First, $Y = y$ was sampled from a normal distribution with mean 0 and variance 4. Then, $\boldsymbol{X}_y = \boldsymbol{x}_y$ was generated according to $\boldsymbol{X}_y = \boldsymbol{\Gamma}\boldsymbol{v}_y + \boldsymbol{\epsilon}$, where $\boldsymbol{\Gamma} = (\boldsymbol{\Gamma}_1, \boldsymbol{\Gamma}_2)^\top$ with $\boldsymbol{\Gamma}_1 = (1, 0, 0, 0, 0)^\top$ and $\boldsymbol{\Gamma}_2 = (0, 1, 0, 0, 0)^\top$, $\boldsymbol{v}_y = (y, y^2/3)^\top$, and the error vector was sampled from a normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\boldsymbol{\Delta} = \mathrm{diag}(1, 1, 5, 5, 5)$.

In the upper plot of Figure 1, the 2-dimensional coordinates of 200 samples from classical multidimensional scaling (CMDS, or principal component analysis by duality) are displayed, with each sample colored according to the response value. There appears to be little discernible relationship between the response and the coordinates (i.e., principal component scores). This lack of pattern is to be expected: aside from the subscript y, nothing on the right-hand side of (2.2) is observable. Consequently, dimension reduction under this model is based solely on the predictors, and hence is unsupervised. The lower plot shows the results of applying the supervised reduction method in Section 4. We see that the response increases as we move from the left to the right, and that the middle and extreme of response

values are somewhat separated by the second coordinate. In other words, some proportion of variability in the response can be explained using the new coordinates.

As was the case in this toy example, in many applications the response is expected to play an important role in supervising our reduction. Indeed, this is the main motivation for most modern dimension-reduction methods, including those developed in the framework of sufficient dimension reduction. We elaborate on this in the next section.

3. A supervised inverse regression model

To facilitate supervised reduction, we can model the coordinate vectors as

$$\boldsymbol{v}_y = \boldsymbol{\beta}\boldsymbol{f}_y,$$

where $\boldsymbol{\beta} \in \mathbb{R}^{d\times r}$ has rank $d \leq \min(r, p)$, and $\boldsymbol{f}_y \in \mathbb{R}^r$ is a known vector-valued function of y. Usually, $\boldsymbol{f}_y$ is required to contain a reasonably flexible set of basis functions, like slice indicator functions or B-spline basis functions. This parameterization has been widely used in model-based reduction; see, for example, Cook and Forzani (2008), Cook et al. (2012), and Wang and Zhu (2013). Replacing $\boldsymbol{v}_y$ in model (2.2) by $\boldsymbol{\beta}\boldsymbol{f}_y$, we are led to the following

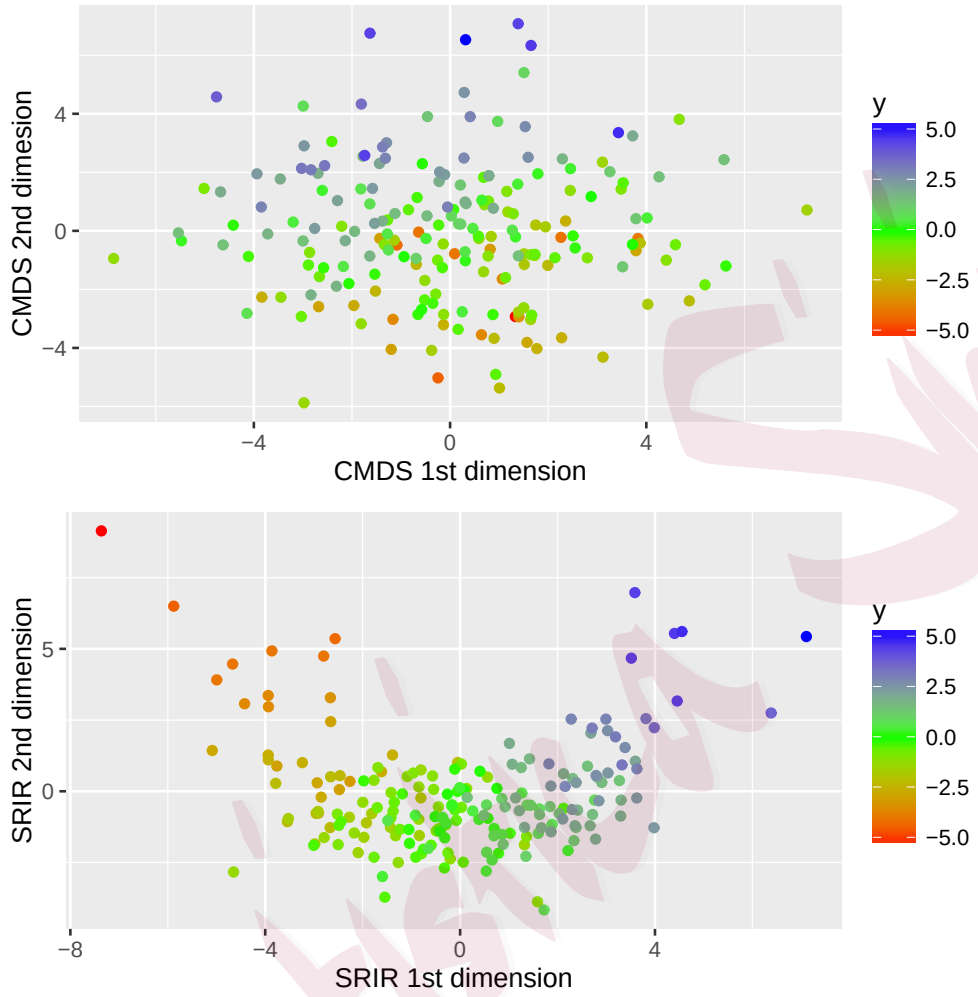

Figure 1: 2-dimensional plots of 200 samples from the inverse regression model (2.2). The simulation setup is described in the text. Top: The axes represent the first and second CMDS coordinates. Bottom: The axes represent the first and second coordinates produced by the supervised reduction method in Section 4.

model

$$\boldsymbol{X}_y = \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\beta}\boldsymbol{f}_y + \boldsymbol{\epsilon}. \tag{3.5}$$

We will refer to this model as a supervised inverse regression model. Without loss of generality, we assume that the sample mean vector of $\boldsymbol{f}_y$ is zero.

The process of dimension reduction based on classical multidimensional scaling is essentially the same as before. Note that, under (3.5),

$$d_{yy'} = \|\boldsymbol{\Gamma}\boldsymbol{\beta}\boldsymbol{f}_y - \boldsymbol{\Gamma}\boldsymbol{\beta}\boldsymbol{f}_{y'}\|_2^2 = \boldsymbol{f}_y^\top\boldsymbol{\beta}^\top\boldsymbol{\beta}\boldsymbol{f}_y - 2\boldsymbol{f}_y^\top\boldsymbol{\beta}^\top\boldsymbol{\beta}\boldsymbol{f}_{y'} + \boldsymbol{f}_{y'}^\top\boldsymbol{\beta}^\top\boldsymbol{\beta}\boldsymbol{f}_{y'}.$$

Let $\boldsymbol{\pi} = (\boldsymbol{f}_{y_1}^\top\boldsymbol{\beta}^\top\boldsymbol{\beta}\boldsymbol{f}_{y_1}, \ldots, \boldsymbol{f}_{y_n}^\top\boldsymbol{\beta}^\top\boldsymbol{\beta}\boldsymbol{f}_{y_n})^\top \in \mathbb{R}^n$ and $\mathbf{F} = (\boldsymbol{f}_{y_1}, \ldots, \boldsymbol{f}_{y_n})^\top \in \mathbb{R}^{n\times r}$. In matrix form

$$\mathbf{D} = \boldsymbol{\pi}\mathbf{1}_n^\top + \mathbf{1}_n\boldsymbol{\pi}^\top - 2\mathbf{F}\boldsymbol{\beta}^\top\boldsymbol{\beta}\mathbf{F}^\top.$$

Since $\mathbf{P}_n\mathbf{F} = \mathbf{F}$, a simple calculation shows that

$$\boldsymbol{\beta}^\top\boldsymbol{\beta} = -\frac{1}{2}(\mathbf{F}^\top\mathbf{F})^{-1}\mathbf{F}^\top\mathbf{D}\mathbf{F}(\mathbf{F}^\top\mathbf{F})^{-1}.$$

Substituting $\mathbf{D}$ by $\hat{\mathbf{D}}$, we arrive at

$$\begin{aligned}(\mathbf{F}^\top\mathbf{F})^{1/2}\boldsymbol{\beta}^\top\boldsymbol{\beta}(\mathbf{F}^\top\mathbf{F})^{1/2} &\approx -\frac{1}{2}(\mathbf{F}^\top\mathbf{F})^{-1/2}\mathbf{F}^\top\hat{\mathbf{D}}\mathbf{F}(\mathbf{F}^\top\mathbf{F})^{-1/2} \\ &= (\mathbf{F}^\top\mathbf{F})^{-1/2}\mathbf{F}^\top\mathbf{X}\mathbf{X}^\top\mathbf{F}(\mathbf{F}^\top\mathbf{F})^{-1/2}.\end{aligned}$$

Write the eigen-decomposition of the term in the last line as

$$(\mathbf{F}^\top\mathbf{F})^{-1/2}\mathbf{F}^\top\mathbf{X}\mathbf{X}^\top\mathbf{F}(\mathbf{F}^\top\mathbf{F})^{-1/2} = \sum_{j=1}^{r}\rho_j\boldsymbol{\phi}_j\boldsymbol{\phi}_j^\top,$$

where $\rho_1 \geq \cdots \geq \rho_r \geq 0$ are the eigenvalues, and $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_r$ are the corresponding eigenvectors. A solution for $\boldsymbol{\beta}$ is then given by

$$\tilde{\boldsymbol{\beta}} = (\rho_1^{1/2}\boldsymbol{\phi}_1, \ldots, \rho_d^{1/2}\boldsymbol{\phi}_d)^\top (\mathbf{F}^\top \mathbf{F})^{-1/2}.$$

Furthermore,

$$\tilde{\boldsymbol{v}}_y = \tilde{\boldsymbol{\beta}} \boldsymbol{f}_y,$$

and the vector of coordinates $\tilde{\boldsymbol{v}}_{y^*}$ associated with a new observation $\boldsymbol{x}_{y^*}$ is, again, computed from (2.3).

4. Reduction when $\boldsymbol{\Delta}$ is unknown

4.1 The proposed method

In practice, $\boldsymbol{\Delta}$ is seldom known in advance, and has to be estimated from the data. Throughout the paper, we estimate $\boldsymbol{\Delta}$ by the residual sample covariance matrix from the multivariate linear regression of $\boldsymbol{X}_y$ on $\boldsymbol{f}_y$:

$$\hat{\boldsymbol{\Delta}} = \frac{1}{n}\mathbf{X}^\top(\mathbf{I}_n - \mathbf{P}_\mathbf{F})\mathbf{X},$$

where $\mathbf{P}_\mathbf{F} = \mathbf{F}(\mathbf{F}^\top\mathbf{F})^{-1}\mathbf{F}^\top$ is the hat matrix. Asymptotic properties of $\hat{\boldsymbol{\Delta}}$ can be found in Lemmas 1 and 2 in the Supplementary Materials. Theorem 3.1 of Cook and Forzani (2008) shows that this estimator and the maximum likelihood estimator under normality of errors are different but related.

We fix $\boldsymbol{\Delta}$ at $\hat{\boldsymbol{\Delta}}$, and base the analysis on the standardized data $\mathbf{X}\hat{\boldsymbol{\Delta}}^{-1/2}$. For simplicity, let us focus on the supervised inverse regression model (3.5). Replacing $\mathbf{X}$ by $\mathbf{X}\hat{\boldsymbol{\Delta}}^{-1/2}$, we compute

$$(\mathbf{F}^\top\mathbf{F})^{-1/2}\mathbf{F}^\top\mathbf{X}\hat{\boldsymbol{\Delta}}^{-1}\mathbf{X}^\top\mathbf{F}(\mathbf{F}^\top\mathbf{F})^{-1/2}$$

and its eigen-decomposition

$$\sum_{j=1}^{r}\hat{\rho}_j\hat{\boldsymbol{\phi}}_j\hat{\boldsymbol{\phi}}_j^\top.$$

We estimate $\boldsymbol{\beta}$ and $\boldsymbol{v}_y$ by

$$\hat{\boldsymbol{\beta}} = (\hat{\rho}_1^{1/2}\hat{\boldsymbol{\phi}}_1, \ldots, \hat{\rho}_d^{1/2}\hat{\boldsymbol{\phi}}_d)^\top(\mathbf{F}^\top\mathbf{F})^{-1/2}$$

and

$$\hat{\boldsymbol{v}}_y = \hat{\boldsymbol{\beta}}\boldsymbol{f}_y.$$

Define $\hat{s}_i = \|\hat{\boldsymbol{v}}_{y_i}\|_2^2 - \|\hat{\boldsymbol{\Delta}}^{-1/2}(\boldsymbol{x}_{y^*} - \boldsymbol{x}_{y_i})\|_2^2$. Let $\hat{\boldsymbol{s}} = (\hat{s}_1, \ldots, \hat{s}_n)^\top$ and $\hat{\mathbf{V}} = (\hat{\boldsymbol{v}}_{y_1}, \ldots, \hat{\boldsymbol{v}}_{y_n})^\top$. Then, the vector of coordinates of a new observation $\boldsymbol{x}_{y^*}$ is predicted by

$$\hat{\boldsymbol{v}}_{y^*} = \frac{1}{2}(\hat{\mathbf{V}}^\top\hat{\mathbf{V}})^{-1}\hat{\mathbf{V}}^\top\hat{\boldsymbol{s}}. \tag{4.6}$$

We call our method supervised reduction via inverse regression (SRIR).

As mentioned earlier, the advantage of working with coordinate vectors is that reduction can be done without the need to estimate $\boldsymbol{\Gamma}$ or $\mathcal{S}_{Y|\boldsymbol{X}}$.

Nevertheless, there are situations in which the inferential target is $\mathcal{S}_{Y|\boldsymbol{X}}$, as is the case in the traditional inquiry of sufficient dimension reduction. To conduct dimension reduction in the original predictor space, we have to determine both $\boldsymbol{\Delta}$ and $\boldsymbol{\Gamma}$, and it is generally infeasible to find a closed-form expression for these estimators; usually an alternating procedure is needed. Fortunately, the estimate $\hat{\boldsymbol{\Delta}}$ has nothing to do with $\boldsymbol{\Gamma}$, suggesting a one-step estimate for $\boldsymbol{\Gamma}$. Specifically, we estimate $\boldsymbol{\Gamma}$ by minimizing the residual sum-of-squares

$$\begin{aligned} \mathrm{RRS}(\boldsymbol{\Gamma}) &= \|\mathbf{P}_n\mathbf{X}\hat{\boldsymbol{\Delta}}^{-1/2} - \hat{\mathbf{V}}\boldsymbol{\Gamma}^\top\hat{\boldsymbol{\Delta}}^{-1/2}\|_F^2 \\ &= \mathrm{trace}\{(\mathbf{P}_n\mathbf{X} - \hat{\mathbf{V}}\boldsymbol{\Gamma}^\top)\hat{\boldsymbol{\Delta}}^{-1}(\mathbf{P}_n\mathbf{X} - \hat{\mathbf{V}}\boldsymbol{\Gamma}^\top)^\top\}. \end{aligned}$$

The solution is

$$\hat{\boldsymbol{\Gamma}} = \mathbf{X}^\top\hat{\mathbf{V}}(\hat{\mathbf{V}}^\top\hat{\mathbf{V}})^{-1} = \mathbf{X}^\top\mathbf{F}\hat{\boldsymbol{\beta}}^\top(\hat{\boldsymbol{\beta}}\mathbf{F}^\top\mathbf{F}\hat{\boldsymbol{\beta}}^\top)^{-1}. \tag{4.7}$$

Note that $\hat{\boldsymbol{\Gamma}}$ depends on $\hat{\boldsymbol{\Delta}}$ (and hence the observed responses y_i) through $\hat{\boldsymbol{\beta}}$. Finally, we estimate $\mathcal{S}_{Y|\boldsymbol{X}}$ by $\mathrm{span}(\hat{\boldsymbol{\Delta}}^{-1}\hat{\boldsymbol{\Gamma}})$.

4.2 Theoretical properties

The limiting behavior of $\hat{\boldsymbol{v}}_{y^*}$ is considered in the following theorem.

Theorem 1. *Assume that $\boldsymbol{v}_Y = \boldsymbol{\beta}\boldsymbol{f}_Y$ has finite sixth moments, and that $\boldsymbol{\epsilon}$*

has finite fourth moments. Then, for some $d \times d$ rotation matrix $\mathbf{R}$*,*

$$\hat{\boldsymbol{v}}_{y^*} = \mathbf{R}(\boldsymbol{v}_{y^*} + \boldsymbol{\Gamma}^\top \boldsymbol{\Delta}^{-1} \boldsymbol{\epsilon}_{y^*}) + O_P\left(\frac{1}{\sqrt{n}}\right).$$

For two d-dimensional random vectors $\boldsymbol{V}_1$ and $\boldsymbol{V}_2$, let $\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2$, and $\boldsymbol{\Sigma}_{12}$ denote the covariance matrix of $\boldsymbol{V}_1$, the covariance matrix of $\boldsymbol{V}_2$, and the covariance matrix between $\boldsymbol{V}_1$ and $\boldsymbol{V}_2$, respectively. To assess the prediction accuracy, we use the multiple correlation coefficient, which is defined as the positive square root of

$$\rho^2(\boldsymbol{V}_1, \boldsymbol{V}_2) = \frac{1}{d}\,\mathrm{trace}(\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_{12}^\top\boldsymbol{\Sigma}_1^{-1}).$$

This measure takes the maximum value of 1 if $\boldsymbol{V}_1$ and $\boldsymbol{V}_2$ are linearly related, and takes the minimum 0 if the components of the two vectors are uncorrelated; see Hall and Mathiason (1990) and Li and Dong (2009). We have the following corollary.

Corollary 1. *Assume the conditions of Theorem 1. Then,*

$$\rho^2(\hat{\boldsymbol{v}}_{Y^*}, \boldsymbol{v}_{Y^*}) = \frac{1}{d}\,\mathrm{trace}[\mathrm{Var}(\boldsymbol{v}_{Y^*})\{\mathrm{Var}(\boldsymbol{v}_{Y^*}) + \mathbf{I}_d\}^{-1}] + O_P\left(\frac{1}{\sqrt{n}}\right),$$

where the covariances in $\rho^2(\hat{\boldsymbol{v}}_{Y^*}, \boldsymbol{v}_{Y^*})$ *are taken with respect to the joint distribution of* Y^* *and* $\boldsymbol{\epsilon}_{Y^*}$*.*

We now consider the situation in which $\boldsymbol{f}_y$ is misspecified. Denote by

$$\{\mathrm{Var}(\boldsymbol{f}_Y)\}^{-1}\mathrm{Cov}(\boldsymbol{f}_Y, \boldsymbol{v}_Y) = \mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^\top$$

the singular value decomposition of $\{\mathrm{Var}(\boldsymbol{f}_Y)\}^{-1}\mathrm{Cov}(\boldsymbol{f}_Y, \boldsymbol{v}_Y)$; that is, $\mathbf{U} = (\mathbf{U}_1, \ldots, \mathbf{U}_d)$ is $r \times d$ with orthonormal columns, $\mathbf{V} = (\mathbf{V}_1, \ldots, \mathbf{V}_d)$ is $d \times d$ orthogonal, and $\mathbf{\Lambda}$ is a $d \times d$ diagonal matrix with diagonal entries $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \geq 0$. Let $\mathbf{\Phi} = (\lambda_1 \mathbf{U}_1, \ldots, \lambda_d \mathbf{U}_d)^\top$.

Theorem 2. *Assume that $\boldsymbol{f}_Y$ has finite sixth moments, and that $\boldsymbol{v}_Y$ and $\boldsymbol{\epsilon}$ both have finite fourth moments. If $\mathrm{Cov}(\boldsymbol{f}_Y, \boldsymbol{v}_Y)$ has full column rank, that is, $\lambda_d > 0$, then*

$$\hat{\boldsymbol{v}}_{y^*} = \mathbf{R}(\boldsymbol{c} + \mathbf{A}\boldsymbol{v}_{y^*} + \mathbf{A}\mathbf{\Gamma}^\top \mathbf{\Omega}^{-1}\boldsymbol{\epsilon}_{y^*}) + O_P\left(\frac{1}{\sqrt{n}}\right),$$

for some $d \times d$ rotation matrix $\mathbf{R}$, where

$$\boldsymbol{c} = \frac{1}{2}\{\mathbf{\Phi}\mathrm{Var}(\boldsymbol{f}_Y)\mathbf{\Phi}^\top\}^{-1}\{\mathrm{E}(\mathbf{\Phi}\boldsymbol{f}_Y\boldsymbol{f}_Y^\top\mathbf{\Phi}^\top\mathbf{\Phi}\boldsymbol{f}_Y) - \mathrm{E}(\mathbf{\Phi}\boldsymbol{f}_Y\boldsymbol{v}_Y^\top\boldsymbol{v}_Y)\},$$

$$\mathbf{A} = \{\mathbf{\Phi}\mathrm{Var}(\boldsymbol{f}_Y)\mathbf{\Phi}^\top\}^{-1}\mathbf{\Phi}\mathrm{Cov}(\boldsymbol{f}_Y, \boldsymbol{v}_Y),$$

and

$$\mathbf{\Omega} = \mathrm{Var}(\boldsymbol{X}) - \mathbf{\Gamma}\mathrm{Cov}(\boldsymbol{v}_Y, \boldsymbol{f}_Y)\{\mathrm{Var}(\boldsymbol{f}_Y)\}^{-1}\mathrm{Cov}(\boldsymbol{f}_Y, \boldsymbol{v}_Y)\mathbf{\Gamma}^\top.$$

This result indicates that, up to an affine transformation, that is, a linear transformation followed by a translation, the conclusion of Theorem 1 remains valid, given that $\boldsymbol{f}_Y$ and $\boldsymbol{v}_Y$ are sufficiently correlated. From a dimension reduction point of view, we can treat $\boldsymbol{v}_{y^*}$ and $\boldsymbol{c} + \mathbf{A}\boldsymbol{v}_{y^*}$ as the same reduction.

The following theorem gives us consistency of subspace estimation.

Theorem 3. *Assume the conditions of Theorem 1 or Theorem 2. Then,* $\text{span}(\hat{\boldsymbol{\Delta}}^{-1}\hat{\boldsymbol{\Gamma}})$ *is a* $\sqrt{n}$ *consistent estimate of* $\mathcal{S}_{Y|\boldsymbol{X}}$.

Let $\hat{\boldsymbol{\Sigma}} = \mathbf{X}^\top \mathbf{P}_{\mathbf{F}} \mathbf{X}$. Define $\mathcal{S}_d(\hat{\boldsymbol{\Delta}}, \hat{\boldsymbol{\Sigma}})$ to be the span of $\hat{\boldsymbol{\Delta}}^{-1/2}$ times the first d eigenvectors of $\hat{\boldsymbol{\Delta}}^{-1/2}\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{\Delta}}^{-1/2}$. One connection between our one-step subspace estimate and the maximum likelihood estimate is captured in the following theorem.

Theorem 4. *Assume that* $\boldsymbol{f}_y$ *is correctly specified. Then,* $\text{span}(\hat{\boldsymbol{\Delta}}^{-1}\hat{\boldsymbol{\Gamma}}) = \mathcal{S}_d(\hat{\boldsymbol{\Delta}}, \hat{\boldsymbol{\Sigma}})$. *Consequently, if* $\boldsymbol{\epsilon}$ *is normally distributed, then* $\text{span}(\hat{\boldsymbol{\Delta}}^{-1}\hat{\boldsymbol{\Gamma}})$ *is the maximum likelihood estimate of* $\mathcal{S}_{Y|\boldsymbol{X}}$.

4.3 Choice of d

In practice, the structural dimension d is unknown, and the choice of d is essential to the proposed method. In the literature, there are two useful techniques for determining d: one is based on a sequential test (Li, 1991), and the other is by using an information criterion (Zhu et al., 2006). Let d_0 denote the true dimension. To estimate d_0, we propose to use the Bayesian information criterion (Zhu et al., 2012). With

$$\text{BIC}_d = \frac{\sum_{j=1}^{d} \hat{\rho}_j^2}{\sum_{k=1}^{r} \hat{\rho}_k^2} - \frac{\log(n)}{n} \times \frac{d(d+1)}{2},$$

the estimated dimension is

$$\hat{d} = \arg\max_{1\le d\le r} \mathrm{BIC}_d. \tag{4.8}$$

Theorem 5. *Assume the conditions of Theorem 1 or Theorem 2. Then, $\hat{d}$ converges to d_0 in probability.*

5. Numerical studies

In this section, we first conduct Monte Carlo simulation studies to assess the finite sample performance of the proposed method. We then apply our method to the analysis of a real data set.

5.1 Simulations

Throughout the simulation study, we considered the structural dimension $d = 2$, the sample size $n = 200$, and the number of predictors $p \in \{10, 20\}$. We set $\mathbf{\Delta} = (\theta^{|i-j|})$, with θ taking 0 or 0.5. Let $\mathbf{\Gamma}_{01} = (1, 1, -1, -1, 0, \ldots, 0)^\top/2$, $\mathbf{\Gamma}_{02} = (1, 0, 1, 0, 1, 0, \ldots, 0)^\top/\sqrt{3}$, and $\mathbf{\Gamma}_0 = (\mathbf{\Gamma}_{01}, \mathbf{\Gamma}_{02})$. Set $\mathbf{\Gamma} = \mathbf{\Gamma}_0(\mathbf{\Gamma}_0^\top \mathbf{\Delta}^{-1}\mathbf{\Gamma}_0)^{-1/2}$.

We used the cubic polynomial basis (y, y^2, y^3) to fit the model, and then assessed the prediction accuracy on an independent test sample, $\{(\boldsymbol{x}_{y_i^*}, y_i^*)\}$, of size 100. Let $\hat{\boldsymbol{v}}_{y_i^*}$ be the predicted vector of coordinates of $\boldsymbol{x}_{y_i^*}$. To measure the closeness between $\hat{\boldsymbol{v}}_{y_i^*}$ and $\boldsymbol{v}_{y_i^*}$, we used the sample version

of the multiple correlation coefficient (MCC). For each configuration, the number of repetitions was 200.

Example 1. To gain first insights into the operating characteristics of the proposed method, consider the model

$$\boldsymbol{X}_y = \boldsymbol{\Gamma}\boldsymbol{v}_y + \boldsymbol{\epsilon},$$

where y is a draw from a normal distribution $N(0, \sigma^2)$, $\boldsymbol{v}_y = (y, y^2)^\top$, and $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \boldsymbol{\Delta})$. By Corollary 1,

$$\rho^2(\hat{\boldsymbol{v}}_{Y^*}, \boldsymbol{v}_{Y^*}) = g^2(\sigma) + O_P\left(\frac{1}{\sqrt{n}}\right).$$

Here, $g(\sigma) = \sqrt{\sigma^2/(2\sigma^2+2) + \sigma^4/(2\sigma^4+1)}$ is an increasing function of σ. Six values of σ were explored: 0.5, 0.8, 1, 1.5, 2, and 3. Figure 2 depicts $g(\sigma)$ and its sample estimate under different configurations. We see that there is an excellent agreement between theoretical prediction and empirical behavior.

Below we examined the behavior of our method in more detail. In addition to the prediction accuracy, we also assessed the performance in terms of subspace recovery. Specifically, we used the vector correlation coefficient (VCC) and the trace correlation coefficient (TCC) to measure the closeness between the true and estimated subspaces (Ye and Weiss, 2003). Let $\hat{\mathbf{B}}$ and $\mathbf{B}$ be basis matrices for the estimated and true subspaces,

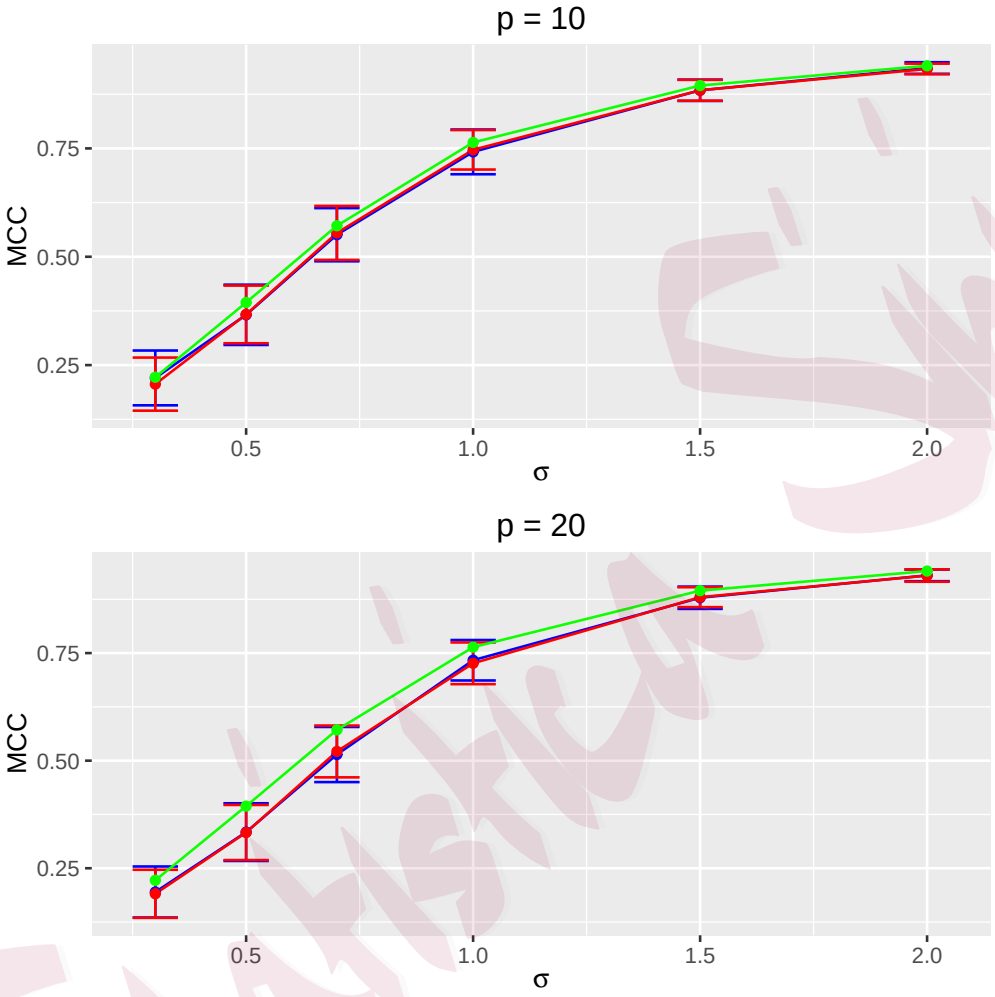

Figure 2: The estimated MCC curves, based on 200 data replications, for $\theta = 0$ (red) and $\theta = 0.5$ (blue). The error bars indicate one standard deviation. The theoretical MCC curve $g(\sigma)$ is shown in green.

respectively. Denote by $\rho_1 \geq \cdots \geq \rho_d$ the ordered eigenvalues of $\hat{\mathbf{B}}^\top \mathbf{B}\mathbf{B}^\top \hat{\mathbf{B}}$. VCC is defined as the square positive root of $q^2(\hat{\mathbf{B}}, \mathbf{B}) = \prod_{i=1}^d \rho_i$, and TCC is defined as the square positive root of $r^2(\hat{\mathbf{B}}, \mathbf{B}) = d^{-1} \sum_{i=1}^d \rho_i$.

Example 2. Consider the model

$$\boldsymbol{X}_y = \boldsymbol{\Gamma}\boldsymbol{\beta}\boldsymbol{f}_y + \boldsymbol{\epsilon},$$

where $\boldsymbol{f}_y = (y, |y|, y^2)^\top$, and $\boldsymbol{\epsilon}$ has mean vector $\mathbf{0}$ and covariance matrix $\boldsymbol{\Delta}$. Two types of non-Gaussian errors were explored, with covariance structures the same as in the previous example. In the former, $\boldsymbol{\epsilon}$ is from multivariate t-distribution with 5 degrees of freedom, and in the latter each component of $\boldsymbol{\epsilon}$ is uniformly distributed on $[-\sqrt{3}, \sqrt{3}]$. For the coefficient matrix $\boldsymbol{\beta}$, we set

$$\boldsymbol{\beta} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{or} \quad \begin{pmatrix} 1 & -0.5 & 0 \\ 0 & 0.5 & 1 \end{pmatrix}.$$

This corresponds to the setting where the cubic polynomial basis is correctly specified or misspecified. Finally, we generated Y from the standard normal distribution. The averaged values of MCC, VCC, and TCC, and their standard deviations, based on 200 data replications, are reported in Tables 1-4. From Tables 1 and 3 we see that the prediction accuracy for non-Gaussian errors is comparable to that under the Gaussian assumption (Figure 2, $\sigma = 1$). Furthermore, our method performs well in terms of subspace

estimation. Generally, the performance improves as the number of predictors decreases. From Tables 2 and 4, we see that our method is robust to misspecification of the basis functions, as expected from Theorems 2 and 3.

Table 1: Means and standard deviations (in parentheses) of MCC, VCC, and TCC, over 200 data applications. $\boldsymbol{\epsilon}$ is from multivariate t-distribution with 5 degrees of freedom and $\boldsymbol{f}_y$ is correctly specified.

		SRIR			PC		
		MCC	VCC	TCC	MCC	VCC	TCC
$p = 10$	$\theta = 0$	0.737 (0.056)	0.900 (0.036)	0.949 (0.018)	0.725 (0.066)	0.845 (0.170)	0.929 (0.061)
	$\theta = 0.5$	0.740 (0.048)	0.916 (0.037)	0.958 (0.018)	0.456 (0.085)	0.120 (0.083)	0.406 (0.096)
$p = 20$	$\theta = 0$	0.716 (0.055)	0.802 (0.047)	0.897 (0.025)	0.667 (0.081)	0.581 (0.279)	0.816 (0.103)
	$\theta = 0.5$	0.728 (0.055)	0.827 (0.048)	0.911 (0.025)	0.374 (0.085)	0.050 (0.049)	0.304 (0.077)

Table 2: Means and standard deviations (in parentheses) of MCC, VCC, and TCC, over 200 data applications. $\boldsymbol{\epsilon}$ is from multivariate t-distribution with 5 degrees of freedom and $\boldsymbol{f}_y$ is misspecified.

		SRIR			PC		
		MCC	VCC	TCC	MCC	VCC	TCC
$p = 10$	$\theta = 0$	0.768 (0.046)	0.899 (0.036)	0.949 (0.018)	0.758 (0.057)	0.843 (0.187)	0.930 (0.067)
	$\theta = 0.5$	0.767 (0.050)	0.917 (0.039)	0.958 (0.019)	0.502 (0.069)	0.130 (0.098)	0.431 (0.082)
$p = 20$	$\theta = 0$	0.752 (0.052)	0.815 (0.041)	0.903 (0.022)	0.711 (0.070)	0.607 (0.257)	0.834 (0.089)
	$\theta = 0.5$	0.752 (0.048)	0.826 (0.049)	0.911 (0.025)	0.457 (0.071)	0.058 (0.049)	0.353 (0.065)

Table 3: Means and standard deviations (in parentheses) of MCC, VCC, and TCC, over 200 data applications. Each component of $\boldsymbol{\epsilon}$ is uniformly distributed on $[-\sqrt{3}, \sqrt{3}]$ and $\boldsymbol{f}_y$ is correctly specified.

		SRIR			PC		
		MCC	VCC	TCC	MCC	VCC	TCC
$p = 10$	$\theta = 0$	0.745 (0.044)	0.931 (0.024)	0.965 (0.012)	0.747 (0.043)	0.946 (0.028)	0.973 (0.014)
	$\theta = 0.5$	0.742 (0.048)	0.942 (0.027)	0.971 (0.013)	0.436 (0.075)	0.097 (0.073)	0.377 (0.088)
$p = 20$	$\theta = 0$	0.727 (0.048)	0.852 (0.034)	0.923 (0.018)	0.731 (0.049)	0.867 (0.052)	0.933 (0.025)
	$\theta = 0.5$	0.730 (0.049)	0.859 (0.042)	0.928 (0.021)	0.380 (0.074)	0.040 (0.037)	0.290 (0.066)

Table 4: Means and standard deviations (in parentheses) of MCC, VCC, and TCC, over 200 data applications. Each component of $\boldsymbol{\epsilon}$ is uniformly distributed on $[-\sqrt{3}, \sqrt{3}]$ and $\boldsymbol{f}_y$ is misspecified.

		SRIR			PC		
		MCC	VCC	TCC	MCC	VCC	TCC
$p = 10$	$\theta = 0$	0.776 (0.033)	0.934 (0.022)	0.967 (0.011)	0.777 (0.033)	0.944 (0.028)	0.972 (0.014)
	$\theta = 0.5$	0.769 (0.036)	0.944 (0.026)	0.971 (0.012)	0.488 (0.060)	0.111 (0.078)	0.411 (0.066)
$p = 20$	$\theta = 0$	0.753 (0.037)	0.853 (0.034)	0.924 (0.018)	0.757 (0.039)	0.881 (0.049)	0.941 (0.024)
	$\theta = 0.5$	0.753 (0.045)	0.869 (0.037)	0.933 (0.019)	0.449 (0.065)	0.049 (0.045)	0.354 (0.058)

We also compared our method with principal components (PC) and principal fitted components (PFC) in Cook and Forzani (2008). The results of PC are shown in the last three columns of Tables 1-4. SRIR appears to

dominate PC in most cases, especially when $\theta = 0.5$. The results of PFC are the same as for SRIR and are omitted. For subspace estimation, this is expected from Theorem 4, but for prediction this comes as somewhat of a surprise. We provide a theoretical support in the Supplementary Materials.

So far the value of the structural dimension is assumed to be known. Using Example 2, we evaluated the performance of the BIC-type criterion (4.8). Tables 5 and 6 report the frequencies of decisions over 200 replications. We see that the proportion of correctly identifying the true dimension is larger than 80% in each configuration. We also see that misspecification can have some impact.

Table 5: Selection frequencies of BIC over 200 data replications. $\boldsymbol{\epsilon}$ is from multivariate t-distribution with 5 degrees of freedom.

		Correctly specified		Misspecified	
		$\hat{d} = 1$	$\hat{d} = 2$	$\hat{d} = 1$	$\hat{d} = 2$
$p = 10$	$\theta = 0$	2	198	39	161
	$\theta = 0.5$	2	198	31	169
$p = 20$	$\theta = 0$	5	195	33	167
	$\theta = 0.5$	4	196	30	170

Table 6: Selection frequencies of BIC over 200 data replications. Each component of $\boldsymbol{\epsilon}$ is uniformly distributed on $[-\sqrt{3}, \sqrt{3}]$.

		Correctly specified		Misspecified	
		$\hat{d}=1$	$\hat{d}=2$	$\hat{d}=1$	$\hat{d}=2$
$p=10$	$\theta=0$	0	200	29	171
	$\theta=0.5$	1	199	32	168
$p=20$	$\theta=0$	3	197	26	174
	$\theta=0.5$	1	199	30	170

5.2 Boston housing data

We applied SRIR to the Boston housing data (Harrison and Rubinfeld, 1978), which is available in the **MASS** library in **R**. This data set has 14 variables and 506 observations, with each observation representing a census tract in Boston Standard Metropolitan Statistical Areas. The variable of primary interest is the median value, in thousands of dollars, of owner occupied homes. The 13 explanatory variables include per capita crime rate by town, average number of rooms per house, percent of households with low socioeconomic status, and so on.

Fitting the supervised inverse regression model (3.5), with the cubic polynomial basis, resulted in BIC choosing $d = 2$, suggesting that two linear combinations of the 13 predictors are sufficient. The top panel of Figure 3 shows the 2-dimensional plot of the 506 observations, with coordinates computed using the formula (4.6). We see a horseshoe-like pattern in

the data cloud. We also see an association between the response and the coordinates, similar to the one in the toy example. For comparison, we also carried out classical multidimensional scaling. In the bottom panel, the ordination of the first two CMDS coordinates is shown. The unsupervised method failed to show any useful relationship.

Figure 4 gives plots of the response versus the SRIR coordinates. The upper panel shows a strong linear relation between the response and the first SRIR coordinate. In the lower panel, we see a nonlinear association between the response and the second SRIR coordinate.

6. Discussion

Linear reduction methods aim to construct a few linear combinations of the original predictors that are useful for subsequent analyses. Nearly all existing methods estimate a subspace in the primal predictor-based space, and then obtain the set of reduced predictors by projecting the original predictor vector onto this subspace. In this paper, we have proposed a principled reduction method in the dual sample-based space, on the basis of a supervised inverse regression model. Instead of estimating the subspace, our method directly estimate the projection coordinates of the predictor vector onto the subspace. The results extend the well-known

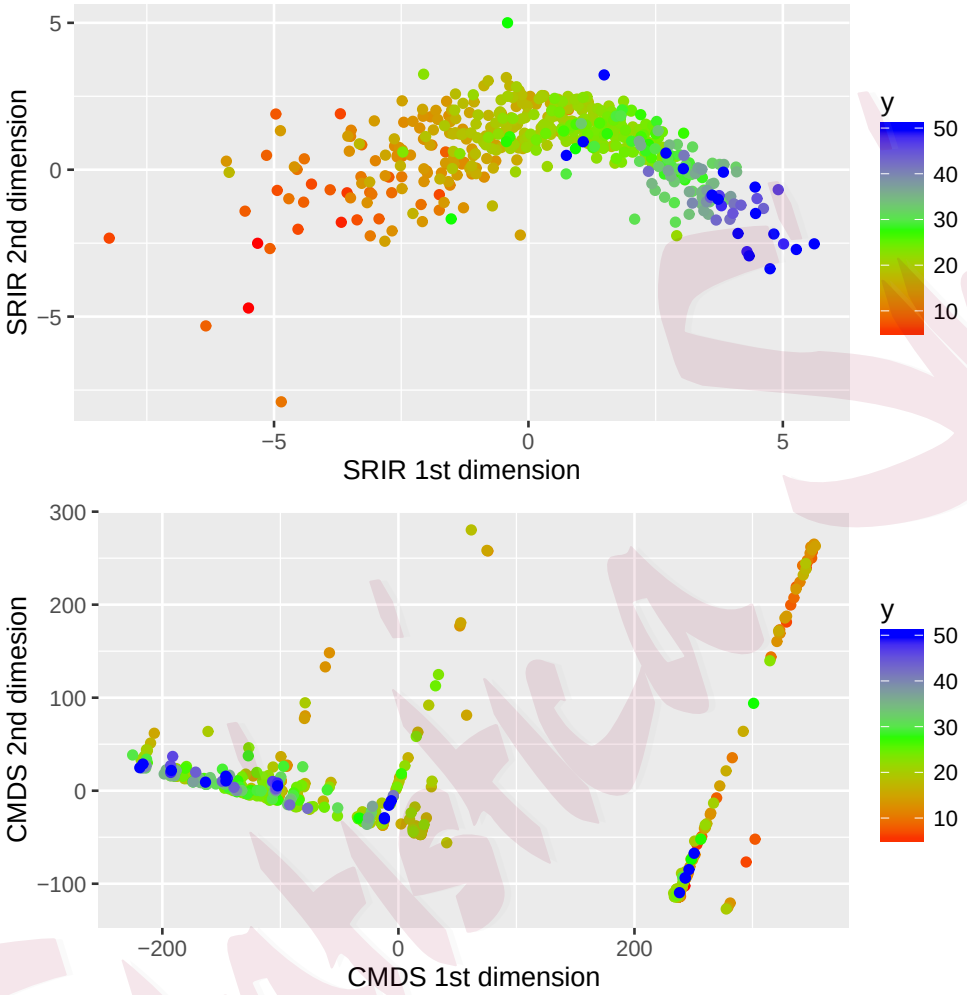

Figure 3: 2-dimensional plots for the Boston housing data. Top: The axes represent the first and second SRIR coordinates. Down: The axes represent the first and second CMDS coordinates.

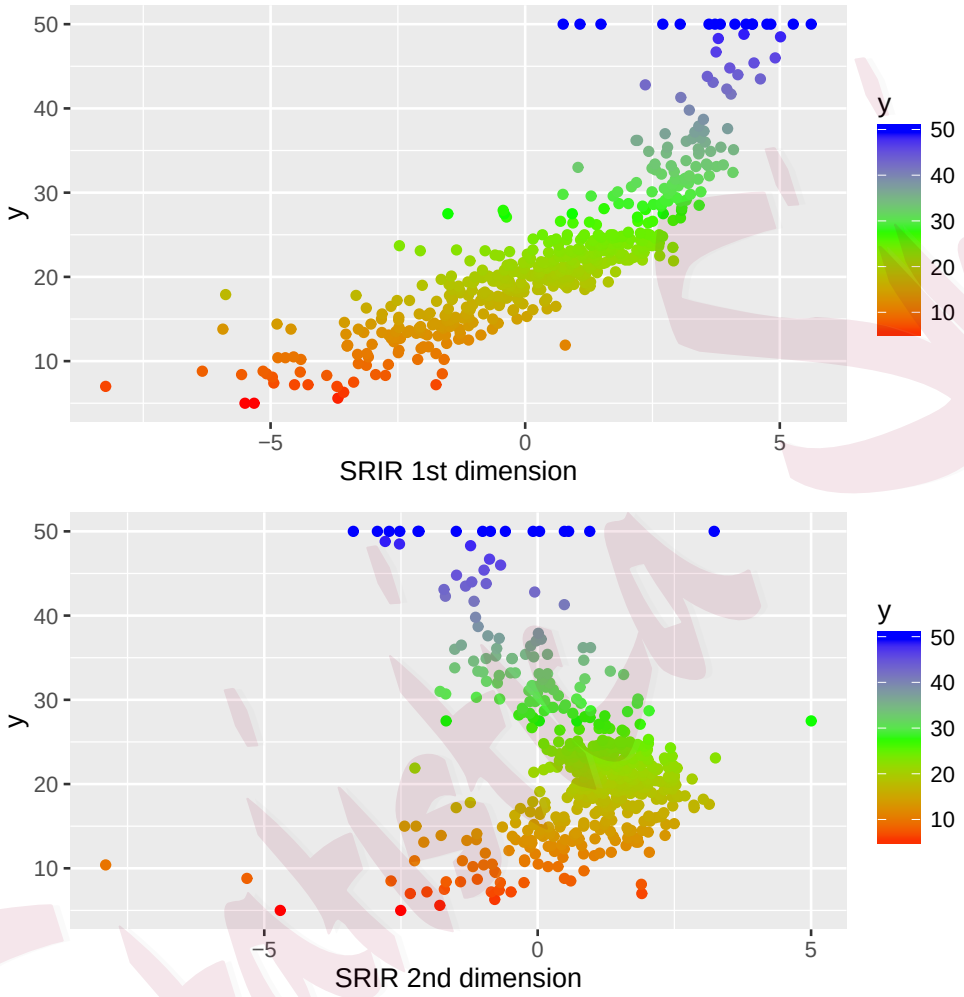

Figure 4: Boston housing data. Top: Response versus the first SRIR coordinate. Down: Response versus the second SRIR coordinate.

duality between principal component analysis and classical multidimensional scaling.

Computation of SRIR has the same order of computation as the maximum likelihood estimation (Cook and Forzani, 2008). However, in terms of generating the reduction for the observed data, our method has a smaller computational cost than the method of maximum likelihood. Specifically, the computational complexity of the former is $O(d \times n \times r^2)$, while that of the latter is $O(d \times n \times p^2)$.

We have studied the theoretical properties of our method, and used simulation results to support them. As with most reduction methods, we have adopted the traditional asymptotic reasoning by letting the sample size $n \to \infty$, with the number of predictors p fixed. Our method requires the inverse of the residual sample covariance matrix, and hence is problematic in situations where p is comparable to or even larger than n. Regularized versions in the dual space have a strong practical appeal and are currently under investigation.

Our method is related to a nonparametric multivariate analysis procedure in ecological studies (Mcardle and Anderson, 2001). This procedure, known as permutation multivariate analysis of variance, partitions the variability in multivariate ecological data according to factors in an experimental design.

The underlying intuition is the duality between $\mathbf{X}^\top\mathbf{X}$, an inner product matrix in the primal space, and $\mathbf{X}\mathbf{X}^\top$, an outer product matrix in the dual space, in the sense that $\text{trace}(\mathbf{X}^\top\mathbf{X}) = \text{trace}(\mathbf{X}\mathbf{X}^\top)$. This equivalence is important because an outer product matrix can be obtained from any symmetric distance matrix $\mathbf{D} = (d_{ij}) \in \mathbb{R}^{n\times n}$ (Gower, 1966). In particular, for a $p\times p$ positive definite matrix $\mathbf{B}$, if we let $d_{ij}(\mathbf{B}) = (\boldsymbol{x}_i - \boldsymbol{x}_j)^\top\mathbf{B}(\boldsymbol{x}_i - \boldsymbol{x}_j)$, then $\mathbf{X}\mathbf{B}\mathbf{X}^\top = -\mathbf{P}_n\mathbf{D}\mathbf{P}_n/2$, where $\mathbf{P}_n$ is the centering matrix. Similar to permutation multivariate analysis of variance, we can extend our supervised reduction method, based solely on measures of distance or dissimilarity between pairs of observations, even without assuming the inverse regression model. Alternatively, under a notion of nonlinear sufficient reduction (Zhang et al., 2008), it is possible to derive a kernel extension of the proposed method. Work along these lines is in progress.

Supplementary Materials The supplementary file contains the proofs.

Acknowledgements Tao Wang is the corresponding author. Peirong Xu was supported by the Natural Science Foundation of Shanghai (19ZR1437000). Tao Wang was supported in part by the National Natural Science Foundation of China (11601326), National Key R&D Program of China (2018YFC0910500),

Shanghai Municipal Science and Technology Major Project (2017SHZDZX01), and Neil Shen's SJTU Medical Research Fund.

References

Adragni, K. P. and R. D. Cook (2009). Sufficient dimension reduction and prediction in regression. *Philosophical Transactions of the Royal Society A 367*(1906), 4385–4405.

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

Cook, R. D. (2007). Fisher lecture: Dimension reduction in regression. *Statistical Science 22*(1), 1–26.

Cook, R. D. and L. Forzani (2008). Principal fitted components for dimension reduction in regression. *Statistical Science 23*(4), 485–501.

Cook, R. D., L. Forzani, and A. J. Rothman (2012). Estimating sufficient reductions of the predictors in abundant high-dimensional regressions. *The Annals of Statistics 40*(1), 353–384.

Cook, R. D. and S. Weisberg (1991). Comment. *Journal of the American Statistical Association 86*(414), 328–332.

Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika 53*, 325–338.

Gower, J. C. (1968). Adding a point to vector diagrams in multivariate analysis. *Biometrika 55*(3), 582–585.

Hall, W. J. and D. J. Mathiason (1990). On large-sample estimation and testing in parametric models. *International Statistical Review 58*(1), 77–97.

Harrison, D. and D. L. Rubinfeld (1978). Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management 5*(1), 81–102.

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

Li, B. and Y. Dong (2009). Dimension reduction for nonelliptically distributed predictors. *The Annals of Statistics 37*(3), 1272–1298.

Li, B. and S. Wang (2007). On directional regression for dimension reduction. *Journal of the American Statistical Association 102*, 997–1008.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association 86*(414), 316–327.

Mcardle, B. H. and M. J. Anderson (2001). Fitting multivariate models to community data: A comment on distance-based redundancy analysis. *Ecology 82*(1), 290–297.

Wang, T., M. Chen, H. Zhao, and L. Zhu (2018). Estimating a sparse reduction for general regression in high dimensions. *Statistics and Computing 28*(1), 33–46.

Wang, T. and L. Zhu (2013). Sparse sufficient dimension reduction using optimal scoring. *Computational Statistics & Data Analysis 57*(1), 223–232.

Ye, Z. and R. E. Weiss (2003). Using the bootstrap to select one of a new class of dimension reduction methods. *Journal of the American Statistical Association 98*(464), 968–979.

Zhang, Z., D. Yeung, J. T. Kwok, and E. Y. Chang (2008). Sliced coordinate analysis for effective dimension reduction and nonlinear extensions. *Journal of Computational and Graphical Statistics 17*(1), 225–242.

Zhu, L., B. Miao, and H. Peng (2006). On sliced inverse regression

with high-dimensional covariates. *Journal of the American Statistical Association 101*(474), 630–643.

Zhu, L., L. Zhu, and Z. Feng (2012). Dimension reduction in regressions through cumulative slicing estimation. *Journal of the American Statistical Association 105*(492), 1455–1466.

College of Mathematics and Sciences, Shanghai Normal University, Shanghai 200234, China

E-mail: prxu@shnu.edu.cn

Department of Bioinformatics and Biostatistics, Shanghai Jiao Tong University, Shanghai 200240, China

SJTU-Yale Joint Center for Biostatistics and Data Science, Shanghai Jiao Tong University, Shanghai 200240, China

MoE Key Lab of Artificial Intelligence, Shanghai Jiao Tong University, Shanghai 200240, China

E-mail: neowangtao@sjtu.edu.cn