Stat Comput (2017) 27:103–114
DOI 10.1007/s11222-015-9609-y

Multiple-population shrinkage estimation via sliced inverse regression

Tao Wang[1] · **Xuerong Meggie Wen**[2] · **Lixing Zhu**[3,4]

Received: 27 August 2014 / Accepted: 31 October 2015 / Published online: 20 November 2015
© Springer Science+Business Media New York 2015

Abstract The problem of dimension reduction in multiple regressions is investigated in this paper, in which data are from several populations that share the same variables. Assuming that the set of relevant predictors is the same across the regressions, a joint estimation and selection method is proposed, aiming to preserve the common structure, while allowing for population-specific characteristics. The new approach is based upon the relationship between sliced inverse regression and multiple linear regression, and is achieved through the lasso shrinkage penalty. A fast alternating algorithm is developed to solve the corresponding optimization problem. The performance of the proposed method is illustrated through simulated and real data examples.

Keywords Joint sparsity · Multiple regressions · Sliced inverse regression · Sufficient dimension reduction

1 Introduction

For a typical regression problem with a univariate response variable Y and a p-dimensional random vector $\boldsymbol{X}$ of predictors, Li (1991) and Cook (1998) proposed sufficient dimension reduction that aims at reducing the dimension of $\boldsymbol{X}$ while preserving the regression relationship between Y and $\boldsymbol{X}$. Specifically, the scope of sufficient dimension reduction is to seek a set of linear combinations of $\boldsymbol{X}$, say $\boldsymbol{\beta}^\top \boldsymbol{X}$, where $\boldsymbol{\beta}$ is a $p \times d$ matrix with $d \leq p$, such that

$$Y \perp\!\!\!\perp \boldsymbol{X} | \boldsymbol{\beta}^\top \boldsymbol{X}, \tag{1.1}$$

where the notation $\perp\!\!\!\perp$ indicates independence. The column space of $\boldsymbol{\beta}$ is called a dimension-reduction subspace. The smallest dimension-reduction subspace, denoted by $\mathcal{S}_{Y|\boldsymbol{X}}$, is called the central subspace for the regression of Y on $\boldsymbol{X}$; it is the intersection of all dimension-reduction subspaces. The goal of sufficient dimension reduction is to make inferences about the central subspace and its dimension, written as $d_{Y|\boldsymbol{X}}$ and called the structural dimension of the regression. Subsequent modeling and prediction can be built upon these new constructed predictors.

Sufficient dimension reduction has received considerable interest in recent years due to the ubiquity of complex and high-dimensional data sets. Many methods have been developed in the literature, including sliced inverse regression (Li 1991), sliced average variance estimation (Cook and Weisberg 1991), minimum average variance estimation (Xia et al. 2002), partial dimension reduction (Chiaromonte et al. 2002), directional regression (Li and Wang 2007), likelihood acquired directions (Cook and Forzani 2009), dimension reduction for a special structured $\boldsymbol{X}$ (Li et al. 2010), discretization-expectation estimation (Zhu et al. 2010), cumulative slicing estimation (Zhu et al. 2010), nonlinear sufficient dimension reduction (Lee et al. 2013), and many others. Generally, these estimation methods can be classified into two categories: eigen-decomposition-based

✉ Lixing Zhu
lzhu@hkbu.edu.hk

[1] Department of Biostatistics, Yale University, New Haven, CT 06520, USA

[2] Department of Mathematics and Statistics, Missouri University of Science and Technology, Rolla, MO 65409, USA

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

[4] School of Statistics, Beijing Normal University, Beijing, China

methods and optimization-based methods, with a majority of them belonging to the first category.

The focus so far in the literature has been on a single population. In many applications, however, data are acquired from multiple populations or sources, with different populations or sources sharing the same variables but differing in the dependence structure among variables. As a result, single-population approaches would mask the underlying heterogeneity, and thus they cannot help us achieve the goal of dimension reduction within each population. Toward this end, a natural way to deal with such heterogeneous data is to use a conditional analysis by applying single-population approaches to each population separately. This strategy, however, fails to reveal or share the common structure across populations.

The study described herein is motivated by the problem of estimating the coefficients of several multiple linear regressions. In this setting, borrowing strength across different regressions by jointly estimating these regression equations could discover a common structure and improve estimation performance, especially when the sample sizes are relatively small. Depending on how the information is shared among the regressions, different algorithms have been devised. For example, it is commonly assumed that only a small subset of the predictors are important for all or most of the regressions. Under this joint sparsity assumption on the regression coefficients, regularization methods have been proposed to recover the shared sparsity structure (e.g., Lounici et al. 2009), and it has been empirically and theoretically shown that, when applied appropriately, a joint analysis has advantages over a conditional analysis.

In the context of sufficient dimension reduction, when there are multiple populations, the situation becomes much more complicated. To be specific, we consider the following conditional independence setting

$$Y^{(k)} \perp\!\!\!\perp \boldsymbol{X}^{(k)} | {\boldsymbol{\beta}^{(k)}}^{\top} \boldsymbol{X}^{(k)}, \quad k = 1, \ldots, K, \tag{1.2}$$

where $Y^{(k)}$ is a univariate response variable, $\boldsymbol{X}^{(k)}$ is a p-dimensional random vector of predictors, and $\boldsymbol{\beta}^{(k)}$ is a $p \times d_k$ matrix with $d_k \leq p$ for the k-th population, for $k = 1, \ldots, K$. Throughout this paper, the population labels are assumed to be known. We note that Chavent et al. (2011) considered a related problem in which they required the same central subspace across different populations.

Two observations are noteworthy. First, there are two aspects of common structure. One of them is related to the aforementioned joint sparsity assumption. The other, which is more abstract, is based on the fact that central subspaces from two different populations can share a common subspace. Second, it is a nontrivial task to exploit the common structure, if any, in a model-free manner. As our first attempt to the multiple-population reduction problem, we in this paper concentrate on the common structure implied by the joint sparsity assumption. In this regard, variable selection becomes an essential tool. Since it seems impossible to extend eigen-decomposition-based methods to develop joint estimation and selection procedures, in the following we are concerned only with optimization-based methods.

For a single population, many variable selection approaches have been developed within the framework of sufficient dimension reduction. There are primarily two types of approaches: those that are test-based and those that are regularization-based. See, for example, Bernard-Michel et al. (2008, 2009), Li and Yin (2008), and Scrucca (2007). However, the first class of approaches are computationally intensive and unsatisfactory in terms of stability. Within the second class, and in terms of optimization-based dimension reduction, only a few variable selection procedures are available. For example, Ni et al. (2005) introduced a shrinkage version of sliced inverse regression by invoking the least squares formulation of sliced inverse regression originated by Cook (2004). More generally, Bondell and Li (2009) extended the idea of shrinkage to a family of inverse regression estimators, which are obtained by minimizing a quadratic objective function (Cook and Ni 2005), and derived the variable-selection consistency. However, these shrinkage estimators require a consistent initial estimator. As a result, their performance depends critically on that of the initial estimator, which may perform poorly when the sample size is small. Indeed, this is in some sense contrary to the spirit of a multiple-population analysis. More recently, Wu and Li (2011) and Wang and Zhu (2013) proposed penalized dimension-reduction estimators by using a general formulation of dimension reduction via multiple linear regression of a set of transformations of the response variable on the predictors. Nevertheless, the former is a two-step procedure, and one is only allowed to exploit the common structure in the first step because an eigen-decomposition problem is solved in the second step. To accomplish this joint estimation, in this paper we describe how the methodology of Wang and Zhu (2013) can be extended to dimension reduction in regression involving multiple populations.

The materials are organized in the following way. The methodology will be developed in Sects. 2 and 3 will contain simulation studies, and a real data example will be put in Sect. 4.

2 Methodology

Suppose we have a data set from (1.2). Specifically, for each $k = 1, \ldots, K$, the data from the k-th population contain n_k samples $(\mathbf{X}^{(k)}, \boldsymbol{y}^{(k)})$, where $\mathbf{X}^{(k)} = (\mathbf{X}_1^{(k)}, \ldots, \mathbf{X}_p^{(k)}) \in \mathbb{R}^{n_k \times p}$ is the matrix of predictor values and $\boldsymbol{y}^{(k)} = (y_1^{(k)}, \ldots, y_{n_k}^{(k)})^{\top} \in \mathbb{R}^{n_k}$ is the vector of response values. Through-

out, we assume that the data from different populations are independent and that the data within each population are independent and identically distributed. Without loss of generality, assume that the columns of $\mathbf{X}^{(k)}$ are centered.

As mentioned before, the most natural way to deal with heterogeneity is to carry out sufficient dimension reduction in a population-wise manner. In this paper, we concentrate on sliced inverse regression since it is a simple and useful first method for dimension reduction in regression. The use of sliced inverse regression, however, depends on its relationship to multiple linear regression, which enables us to propose our joint estimation method to improve the estimation accuracy under the joint sparsity assumption.

2.1 Sliced inverse regression via multiple linear regression

Consider the k-th population in the conditional independence setting (1.2). It is known that sliced inverse regression can be formulated as a generalized eigenvalue problem of the form

$$\begin{aligned} &\mathrm{cov}\left[E\left\{\boldsymbol{X}^{(k)} - E(\boldsymbol{X}^{(k)})|Y^{(k)}\right\}\right]\boldsymbol{v}_i^{(k)} \\ &= \lambda_i^{(k)}\mathrm{cov}(\boldsymbol{X}^{(k)})\boldsymbol{v}_i^{(k)}, \quad \text{for} \quad i = 1, \ldots, p, \end{aligned} \tag{2.1}$$

where the vectors $\boldsymbol{v}_1^{(k)}, \ldots, \boldsymbol{v}_p^{(k)}$ are the eigenvectors such that $\boldsymbol{v}_i^{(k)\top}\mathrm{cov}(\boldsymbol{X}^{(k)})\boldsymbol{v}_j^{(k)} = 1$ if $i = j$, and 0 if $i \neq j$, and $\lambda_1^{(k)} \geq \cdots \geq \lambda_p^{(k)} \geq 0$ are the corresponding eigenvalues. Under the linearity condition (Li 1991), the first $d^{(k)}$ eigenvectors $\{\boldsymbol{v}_1^{(k)}, \ldots, \boldsymbol{v}_{d^{(k)}}^{(k)}\}$, which correspond to the nonzero eigenvalues $\lambda_1^{(k)} \geq \cdots \geq \lambda_{d^{(k)}}^{(k)} > 0$, are contained in the central subspace $\mathcal{S}_{Y^{(k)}|\boldsymbol{X}^{(k)}}$. For simplicity, we assume that they form a basis for $\mathcal{S}_{Y^{(k)}|\boldsymbol{X}^{(k)}}$ with $d^{(k)} = d_{Y^{(k)}|\boldsymbol{X}^{(k)}}$.

Somewhat less known is that there is an intrinsic connection between sliced inverse regression and multiple linear regression: sliced inverse regression is a transformation-based projection pursuit that finds linear combinations of the predictors that maximize the correlation with the optimally transformed response variable (Chen and Li 1998). More precisely, the criterion of sliced inverse regression via multiple linear regression has the form

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

where $T_i^{(k)} = T_i^{(k)}(Y^{(k)})$, $i = 1, \ldots, p$. It has been shown that the i-th optimal transformation $T_i^{(k)}$ is identical, up to a scalar multiplication, to $E(\boldsymbol{v}_i^{(k)\top}\boldsymbol{X}^{(k)}|Y^{(k)})$, and that $\boldsymbol{b}_i^{(k)}$ is proportional to $\boldsymbol{v}_i^{(k)}$, where $\boldsymbol{v}_i^{(k)}$ is the i-th sliced inverse regression direction given in (2.1). See Chen and Li (1998) for details.

Following Wang and Zhu (2013), we use linear combinations of basis functions to represent these response transformations. To be specific, let $\{\phi_1^{(k)}(y), \ldots, \phi_{H_k}^{(k)}(y)\}$ be a known set of basis functions with $H_k \geq d^{(k)} + 1$. We then linearize $T_i^{(k)}(Y^{(k)})$ by $\boldsymbol{\theta}_i^{(k)\top}\boldsymbol{\phi}^{(k)}(Y^{(k)})$, where $\boldsymbol{\phi}^{(k)} = (\phi_1^{(k)}, \ldots, \phi_{H_k}^{(k)})^\top$, and $\boldsymbol{\theta}_i^{(k)} = (\theta_{i1}^{(k)}, \ldots, \theta_{iH_k}^{(k)})^\top$ is an H_k-dimensional coefficient vector. Let $\boldsymbol{\Phi}^{(k)} = \{\boldsymbol{\phi}^{(k)}(y_1^{(k)}), \ldots, \boldsymbol{\phi}^{(k)}(y_{n_k}^{(k)})\}^\top \in \mathbb{R}^{n_k \times H_k}$ be the matrix of basis function values. In the sample, sliced inverse regression via multiple linear regression solves

$$\begin{aligned} \underset{\boldsymbol{\theta}_i^{(k)} \in \mathbb{R}^{H_k}, \boldsymbol{\beta}_i^{(k)} \in \mathbb{R}^p}{\text{minimize}} \quad & \|\boldsymbol{\Phi}^{(k)}\boldsymbol{\theta}_i^{(k)} - \mathbf{X}^{(k)}\boldsymbol{\beta}_i^{(k)}\|_2^2 \\ \text{subject to} \quad & \boldsymbol{\theta}_i^{(k)\top}\boldsymbol{\Phi}^{(k)\top}\boldsymbol{\Phi}^{(k)}\boldsymbol{\theta}_i^{(k)} \\ & = n_k, \boldsymbol{\theta}_i^{(k)\top}\boldsymbol{\Phi}^{(k)\top}\boldsymbol{\Phi}^{(k)}\boldsymbol{\theta}_j^{(k)} = 0, \\ & j = 1, \ldots, i-1, \end{aligned} \tag{2.3}$$

where $\|\cdot\|_2$ stands for the vector l_2 norm. This procedure is known as optimal scoring in the machine learning literature (Hastie et al. 2009).

There are various choices for the basis functions. The original sliced inverse regression algorithm uses indicator functions for slices, with H_k being the number of slices. Alternatively, we can use B-spline basis functions (including an intercept). Since the columns of $\mathbf{X}^{(k)}$ are centered to have mean zero, one can see that for these two cases the constant coefficient vector $\boldsymbol{1}_{H_k} = (1, \ldots, 1)^\top$ of length H_k is trivial, and hence there are at most $H_k - 1$ nontrivial solutions to (2.3). We let $\{(\tilde{\boldsymbol{\theta}}_i^{(k)}, \tilde{\boldsymbol{\beta}}_i^{(k)})\}_{i=1}^{d^{(k)}}$ denote the first $d^{(k)}$ solutions. Then the estimator of the central subspace $\mathcal{S}_{Y^{(k)}|\boldsymbol{X}^{(k)}}$ is given by $\mathrm{span}\{\tilde{\mathbf{B}}^{(k)}\}$, where $\tilde{\mathbf{B}}^{(k)} = (\tilde{\boldsymbol{\beta}}_1^{(k)}, \ldots, \tilde{\boldsymbol{\beta}}_{d^{(k)}}^{(k)})$ with $\tilde{\boldsymbol{\beta}}_i^{(k)}$ as its i-th column, $i = 1, \ldots, d^{(k)}$.

2.2 Conditional shrinkage sliced inverse regression

When a subset of predictors is irrelevant or redundant, it is desirable to have the corresponding row estimates of $\tilde{\mathbf{B}}^{(k)}$ equal to zero, and consequently to achieve predictor selection. Following Ni et al. (2005) and Bondell and Li (2009), we can compute a shrinkage estimator by solving

$$\begin{aligned} \underset{\boldsymbol{\alpha}^{(k)}}{\text{minimize}} \quad & \sum_{i=1}^{d^{(k)}} \|\boldsymbol{\Phi}^{(k)}\tilde{\boldsymbol{\theta}}_i^{(k)} - \mathbf{X}^{(k)}\mathrm{diag}(\boldsymbol{\alpha}^{(k)})\tilde{\boldsymbol{\beta}}_i^{(k)}\|_2^2 \\ \text{subject to} \quad & \sum_{j=1}^{p} |\alpha_j^{(k)}| \leq \tau_k, \end{aligned} \tag{2.4}$$

where $\boldsymbol{\alpha}^{(k)} = (\alpha_1^{(k)}, \ldots, \alpha_p^{(k)})^\top \in \mathbb{R}^p$, and $\tau_k \geq 0$ is a shrinkage parameter. Denote the solution by $\check{\boldsymbol{\alpha}}^{(k)}(\tau_k)$. Let $\check{\mathbf{B}}^{(k)}(\tau_k) = \text{diag}\{\check{\boldsymbol{\alpha}}^{(k)}(\tau_k)\}\tilde{\mathbf{B}}^{(k)}$. Then $\text{span}\{\check{\mathbf{B}}^{(k)}(\tau_k)\}$ is a shrinkage estimator of $\mathcal{S}_{Y^{(k)}|\boldsymbol{X}^{(k)}}$.

The performance of $\check{\mathbf{B}}^{(k)}$, however, depends critically on that of the initial estimator $\tilde{\mathbf{B}}^{(k)}$. This is clearly undesirable, because $\tilde{\mathbf{B}}^{(k)}$ may perform poorly when the sample size n_k is small; otherwise, there is no need to take into account the shared information across populations. Thus, in some sense this shrinkage estimator, which is "two-step" in nature, is contrary to the spirit of a joint or multiple-population analysis. In this regard, a "multi-step" or "fully iterative" estimator is preferred. To this end, we propose an improved version of $\check{\mathbf{B}}^{(k)}$ by solving the following minimization problem:

$$\begin{aligned} \underset{\{(\boldsymbol{\theta}_i^{(k)}, \boldsymbol{\beta}_i^{(k)})\}_{i=1}^{d^{(k)}}, \boldsymbol{\alpha}^{(k)}}{\text{minimize}} \quad & \sum_{i=1}^{d^{(k)}} \|\boldsymbol{\Phi}^{(k)}\boldsymbol{\theta}_i^{(k)} - \mathbf{X}^{(k)}\text{diag}(\boldsymbol{\alpha}^{(k)})\boldsymbol{\beta}_i^{(k)}\|_2^2 \\ \text{subject to} \quad & {\boldsymbol{\theta}_i^{(k)}}^\top {\boldsymbol{\Phi}^{(k)}}^\top \boldsymbol{\Phi}^{(k)} \boldsymbol{\theta}_i^{(k)} \\ & = n_k, {\boldsymbol{\theta}_i^{(k)}}^\top {\boldsymbol{\Phi}^{(k)}}^\top \boldsymbol{\Phi}^{(k)} \boldsymbol{\theta}_j^{(k)} = 0, \\ & j = 1, \ldots, i-1, i = 1, \ldots, d^{(k)}, \\ & \sum_{j=1}^{p} |\alpha_j^{(k)}| \leq \tau_k, \quad \tau_k \geq 0. \end{aligned} \tag{2.5}$$

This can be minimized by an alternating optimization procedure given in Sect. 2.4 below. Denote the solution by $\check{\boldsymbol{\theta}}_i^{(k)}(\tau_k)$, $\check{\boldsymbol{\beta}}_i^{(k)}(\tau_k)$ and $\check{\boldsymbol{\alpha}}^{(k)}(\tau_k)$. The improved shrinkage estimator of $\mathcal{S}_{Y^{(k)}|\boldsymbol{X}^{(k)}}$ is given by $\text{span}\{\check{\mathbf{B}}^{(k)}(\tau_k)\}$, where $\check{\mathbf{B}}^{(k)}(\tau_k) = \text{diag}\{\check{\boldsymbol{\alpha}}^{(k)}(\tau_k)\}\{\check{\boldsymbol{\beta}}_1^{(k)}(\tau_k), \ldots, \check{\boldsymbol{\beta}}_{d^{(k)}}^{(k)}(\tau_k)\}$.

2.3 Multiple-population sliced inverse regression

To improve estimation under the joint sparsity assumption, we propose a joint estimation method as follows. Let $\mathcal{K} \subseteq \mathcal{K}_0 = \{1, \ldots, K\}$. Although $\mathcal{K} = \mathcal{K}_0$ is the focus of this paper, we feel this notation is more convenient. For a given $\mathcal{K}$, we consider the following shrinkage criterion

$$\begin{aligned} \underset{\{\boldsymbol{\Theta}^{\mathcal{K}}, \boldsymbol{\alpha}\}}{\text{minimize}} \quad & \sum_{k\in\mathcal{K}} \sum_{i=1}^{d^{(k)}} \|\boldsymbol{\Phi}^{(k)}\boldsymbol{\theta}_i^{(k)} - \mathbf{X}^{(k)}\text{diag}(\boldsymbol{\alpha})\boldsymbol{\beta}_i^{(k)}\|_2^2 \\ \text{subject to} \quad & {\boldsymbol{\theta}_i^{(k)}}^\top {\boldsymbol{\Phi}^{(k)}}^\top \boldsymbol{\Phi}^{(k)} \boldsymbol{\theta}_i^{(k)} \\ & = n_k, {\boldsymbol{\theta}_i^{(k)}}^\top {\boldsymbol{\Phi}^{(k)}}^\top \boldsymbol{\Phi}^{(k)} \boldsymbol{\theta}_j^{(k)} = 0, \\ & j = 1, \ldots, i-1, i = 1, \ldots, d^{(k)}, k \in \mathcal{K}, \\ & \sum_{j=1}^{p} |\alpha_j| \leq \tau^{\mathcal{K}}, \quad \tau^{\mathcal{K}} \geq 0, \end{aligned} \tag{2.6}$$

where $\boldsymbol{\Theta}^{\mathcal{K}} = [\{(\boldsymbol{\theta}_i^{(k)}, \boldsymbol{\beta}_i^{(k)})\}_{i=1}^{d^{(k)}}]_{k\in\mathcal{K}}$ and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)^\top$. It is clear that (2.6) reduces to (2.5) when $\mathcal{K} = \{k\}$. Our multiple-population estimation method corresponds to $\mathcal{K} = \mathcal{K}_0$.

Let $\hat{\boldsymbol{\theta}}_i^{(k)}(\tau^{\mathcal{K}})$, $\hat{\boldsymbol{\beta}}_i^{(k)}(\tau^{\mathcal{K}})$ and $\hat{\boldsymbol{\alpha}}(\tau^{\mathcal{K}})$ denote the solution. The joint estimator of $\mathcal{S}_{Y^{(k)}|\boldsymbol{X}^{(k)}}$ is given by $\text{span}\{\hat{\mathbf{B}}^{(k)}(\tau^{\mathcal{K}})\}$, where $\hat{\mathbf{B}}^{(k)}(\tau^{\mathcal{K}}) = \text{diag}\{\hat{\boldsymbol{\alpha}}(\tau^{\mathcal{K}})\}\{\hat{\boldsymbol{\beta}}_1^{(k)}(\tau^{\mathcal{K}}), \ldots, \hat{\boldsymbol{\beta}}_{d^{(k)}}^{(k)}(\tau^{\mathcal{K}})\}$.

2.4 The optimization algorithm

To estimate $\boldsymbol{\Theta}^{\mathcal{K}}$ and $\boldsymbol{\alpha}$, an iterative algorithm is used. That is, first, fix $\boldsymbol{\Theta}^{\mathcal{K}}$ and estimate $\boldsymbol{\alpha}$; second, fix $\boldsymbol{\alpha}$ and estimate $\boldsymbol{\Theta}^{\mathcal{K}}$; then iterate between these two steps until the solution converges. Since at each step, the value of the objective function in (2.6) decreases, the solution is guaranteed to converge. In general, this algorithm converges to a local minimizer, because the optimazation problem is non-convex. For identifiability of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}_i^{(k)}$, we assume in the sequel that $\alpha_j \geq 0$ and $\|\boldsymbol{\beta}_j^{\mathcal{K}}\|_2 = 1$, for $j = 1, \ldots, p$, where $\boldsymbol{\beta}_j^{\mathcal{K}} = [\{\beta_{ij}^{(k)}\}_{i=1}^{d^{(k)}}]_{k\in\mathcal{K}}$ denotes the set of coefficients corresponding to the j-th coordinate.

Solving (2.6) for $\boldsymbol{\alpha}$ with $[\{(\boldsymbol{\theta}_i^{(k)}, \boldsymbol{\beta}_i^{(k)})\}_{i=1}^{d^{(k)}}]_{k\in\mathcal{K}}$ fixed yields the optimization problem

$$\begin{aligned} \underset{\boldsymbol{\alpha}}{\text{minimize}} \quad & \sum_{k\in\mathcal{K}} \sum_{i=1}^{d^{(k)}} \|\boldsymbol{\Phi}^{(k)}\boldsymbol{\theta}_i^{(k)} - \mathbf{X}^{(k)}\text{diag}(\boldsymbol{\beta}_i^{(k)})\boldsymbol{\alpha}\|_2^2 \\ \text{subject to} \quad & \sum_{j=1}^{p} \alpha_j \leq \tau^{\mathcal{K}}, \quad \tau^{\mathcal{K}} \geq 0. \end{aligned} \tag{2.7}$$

Before continuing we introduce some notation. Let $\boldsymbol{y}^{*(k)}$ denote the vector formed by stacking the vectors $\boldsymbol{\Phi}^{(k)}\boldsymbol{\theta}_i^{(k)}$, $i = 1, \ldots, d^{(k)}$. Likewise $\mathbf{X}^{*(k)}$ denotes the matrix obtained by stacking the matrices $\mathbf{X}^{(k)}\text{diag}(\boldsymbol{\beta}_i^{(k)})$, $i = 1, \ldots, d^{(k)}$. Let $\boldsymbol{y}^{\mathcal{K}} = (\boldsymbol{y}^{*(k)}, k \in \mathcal{K})$, and let $\mathbf{X}^{\mathcal{K}} = \text{diag}(\mathbf{X}^{*(k)}, k \in \mathcal{K})$ denote the block diagonal matrix with submatrices $\mathbf{X}^{*(k)}$, $k \in \mathcal{K}$ along the diagonal, and zeros elsewhere. Now rewrite (2.7) as

$$\begin{aligned} \underset{\boldsymbol{\alpha}}{\text{minimize}} \quad & \|\boldsymbol{y}^{\mathcal{K}} - \mathbf{X}^{\mathcal{K}}\boldsymbol{\alpha}\|_2^2 \\ \text{subject to} \quad & \sum_{j=1}^{p} \alpha_j \leq \tau^{\mathcal{K}}, \quad \tau^{\mathcal{K}} \geq 0. \end{aligned} \tag{2.8}$$

This is a lasso-type problem (Tibshirani 1996), and the shrinkage factors α_j can be computed using either an efficient solution path algorithm or a quadratic programming package.

Solving (2.6) for $[\{(\boldsymbol{\theta}_i^{(k)}, \boldsymbol{\beta}_i^{(k)})\}_{i=1}^{d^{(k)}}]_{k\in\mathcal{K}}$ with $\boldsymbol{\alpha}$ fixed yields $|\mathcal{K}|$ individual optimization problems

$$\begin{aligned}
\underset{\{(\boldsymbol{\theta}_i^{(k)},\boldsymbol{\beta}_i^{(k)})\}_{i=1}^{d^{(k)}}}{\text{minimize}} \quad & \sum_{i=1}^{d^{(k)}} \|\boldsymbol{\Phi}^{(k)}\boldsymbol{\theta}_i^{(k)} - \mathbf{X}^{(k)}\mathrm{diag}(\boldsymbol{\alpha})\boldsymbol{\beta}_i^{(k)}\|_2^2 \\
\text{subject to} \quad & \boldsymbol{\theta}_i^{(k)\top}\boldsymbol{\Phi}^{(k)\top}\boldsymbol{\Phi}^{(k)}\boldsymbol{\theta}_i^{(k)} \\
& = n_k, \boldsymbol{\theta}_i^{(k)\top}\boldsymbol{\Phi}^{(k)\top}\boldsymbol{\Phi}^{(k)}\boldsymbol{\theta}_j^{(k)} = 0, \\
& j = 1, \ldots, i-1, i = 1, \ldots, d^{(k)}.
\end{aligned} \tag{2.9}$$

Let $\mathcal{A} = \{j : \alpha_j \neq 0\}$. Write $\boldsymbol{\beta}_i^{(k)} = (\beta_{i1}^{(k)}, \ldots, \beta_{ip}^{(k)})^\top$. We can simply set $\beta_{ij}^{(k)} = (d_k|\mathcal{K}|)^{-1/2}$ for $j \notin \mathcal{A}$. Let $\mathbf{X}_{\mathcal{A}}^{(k)} = (\mathbf{X}_j^{(k)}, j \in \mathcal{A})$, $\boldsymbol{\alpha}_{\mathcal{A}} = (\alpha_j, j \in \mathcal{A})$ and $\boldsymbol{\beta}_{i\mathcal{A}}^{(k)} = (\beta_{ij}^{(k)}, j \in \mathcal{A})$. Then the above criteria become

$$\begin{aligned}
\underset{\{(\boldsymbol{\theta}_i^{(k)},\boldsymbol{\beta}_{i\mathcal{A}}^{(k)})\}_{i=1}^{d^{(k)}}}{\text{minimize}} \quad & \sum_{i=1}^{d^{(k)}} \|\boldsymbol{\Phi}^{(k)}\boldsymbol{\theta}_i^{(k)} - \mathbf{X}_{\mathcal{A}}^{(k)}\mathrm{diag}(\boldsymbol{\alpha}_{\mathcal{A}})\boldsymbol{\beta}_{i\mathcal{A}}^{(k)}\|_2^2 \\
\text{subject to} \quad & \boldsymbol{\theta}_i^{(k)\top}\boldsymbol{\Phi}^{(k)\top}\boldsymbol{\Phi}^{(k)}\boldsymbol{\theta}_i^{(k)} \\
& = n_k, \boldsymbol{\theta}_i^{(k)\top}\boldsymbol{\Phi}^{(k)\top}\boldsymbol{\Phi}^{(k)}\boldsymbol{\theta}_j^{(k)} = 0, \\
& j = 1, \ldots, i-1, i = 1, \ldots, d^{(k)}.
\end{aligned} \tag{2.10}$$

We note that, for each $k \in \mathcal{K}$, this is the criterion for sliced inverse regression via multiple linear regression. In the literature, the standard way of solving (2.9) is by a suitable singular value decomposition. However, we propose to update $\{\boldsymbol{\theta}_i^{(k)}\}_{i=1}^{d^{(k)}}$ and $\{\boldsymbol{\beta}_{i\mathcal{A}}^{(k)}\}_{i=1}^{d^{(k)}}$ separately as follows. For fixed $\{\boldsymbol{\beta}_{i\mathcal{A}}^{(k)}\}_{i=1}^{d^{(k)}}$, the coefficient vectors $\{\boldsymbol{\theta}_i^{(k)}\}_{i=1}^{d^{(k)}}$ sequentially solve

$$\begin{aligned}
\underset{\boldsymbol{\theta}_i^{(k)}}{\text{minimize}} \quad & \|\boldsymbol{\Phi}^{(k)}\boldsymbol{\theta}_i^{(k)} - \mathbf{X}_{\mathcal{A}}^{(k)}\mathrm{diag}(\boldsymbol{\alpha}_{\mathcal{A}})\boldsymbol{\beta}_{i\mathcal{A}}^{(k)}\|_2^2 \\
\text{subject to} \quad & \boldsymbol{\theta}_i^{(k)\top}\boldsymbol{\Phi}^{(k)\top}\boldsymbol{\Phi}^{(k)}\boldsymbol{\theta}_i^{(k)} \\
& = n_k, \boldsymbol{\theta}_i^{(k)\top}\boldsymbol{\Phi}^{(k)\top}\boldsymbol{\Phi}^{(k)}\boldsymbol{\theta}_j^{(k)} = 0, \\
& j = 1, \ldots, i-1.
\end{aligned} \tag{2.11}$$

Let $\mathbf{Q}_i^{(k)} = (\boldsymbol{I}_{H_k}, \boldsymbol{\theta}_1^{(k)}, \ldots, \boldsymbol{\theta}_{i-1}^{(k)})$ denote the $H_k \times i$ matrix consisting of the previous $i-1$ solutions, as well as the H_k-dimensional vector of all ones. Let $\mathbf{D}^{(k)} = n_k^{-1}\boldsymbol{\Phi}^{(k)\top}\boldsymbol{\Phi}^{(k)}$. One can show that the i-th solution is given by

$$\begin{aligned}
\boldsymbol{\theta}_i^{(k)} = c_i^{(k)} & \left(\mathbf{I}_{H_k} - \mathbf{Q}_i^{(k)}\mathbf{Q}_i^{(k)\top}\mathbf{D}^{(k)}\right) \\
& \times \left(\mathbf{D}^{(k)}\right)^{-1} \boldsymbol{\Phi}^{(k)\top}\mathbf{X}_{\mathcal{A}}^{(k)}\mathrm{diag}\left(\boldsymbol{\alpha}_{\mathcal{A}}\right)\boldsymbol{\beta}_{i\mathcal{A}}^{(k)},
\end{aligned}$$

where $\mathbf{I}_m$ denotes the $m \times m$ identity matrix and $c_i^{(k)}$ is a constant such that $\boldsymbol{\theta}_i^{(k)\top}\mathbf{D}^{(k)}\boldsymbol{\theta}_i^{(k)} = 1$. For fixed $\{\boldsymbol{\theta}_i^{(k)}\}_{i=1}^{d^{(k)}}$, we obtain $d^{(k)}$ linear least squares problems

$$\underset{\boldsymbol{\beta}_{i\mathcal{A}}^{(k)}}{\text{minimize}} \quad \|\boldsymbol{\Phi}^{(k)}\boldsymbol{\theta}_i^{(k)} - \mathbf{X}_{\mathcal{A}}^{(k)}\mathrm{diag}(\boldsymbol{\alpha}_{\mathcal{A}})\boldsymbol{\beta}_{i\mathcal{A}}^{(k)}\|_2^2. \tag{2.12}$$

The i-solution is

$$\begin{aligned}
\boldsymbol{\beta}_{i\mathcal{A}}^{(k)} = & \left[\left\{\mathbf{X}_{\mathcal{A}}^{(k)}\mathrm{diag}\left(\boldsymbol{\alpha}_{\mathcal{A}}\right)\right\}^\top \mathbf{X}_{\mathcal{A}}^{(k)}\mathrm{diag}\left(\boldsymbol{\alpha}_{\mathcal{A}}\right)\right]^{-1} \\
& \times \left\{\mathbf{X}_{\mathcal{A}}^{(k)}\mathrm{diag}\left(\boldsymbol{\alpha}_{\mathcal{A}}\right)\right\}^\top \boldsymbol{\Phi}^{(k)}\boldsymbol{\theta}_i^{(k)}.
\end{aligned}$$

For identifiability of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}_{i\mathcal{A}}^{(k)}$, we normalize $\boldsymbol{\beta}_{i\mathcal{A}}^{(k)}$ so that $\sum_{k\in\mathcal{K}}\sum_{i=1}^{d^{(k)}}(\beta_{ij}^{(k)})^2 = 1$ for $j \in \mathcal{A}$.

In summary, the algorithm proceeds as follows:

Step 0. Initialization. Initialize $\{(\boldsymbol{\theta}_i^{(k)}, \boldsymbol{\beta}_i^{(k)})\}_{i=1}^{d^{(k)}}$ with some plausible values. For example, we can use the solutions $\{(\tilde{\boldsymbol{\theta}}_i^{(k)}, \tilde{\boldsymbol{\beta}}_i^{(k)})\}_{i=1}^{d^{(k)}}$ to (2.3). Let $\mathcal{A} = \{1, \ldots, p\}$.

Step 1. Update $\boldsymbol{\alpha}$. Set $\boldsymbol{y}_i^{*(k)} = \boldsymbol{\Phi}^{(k)}\boldsymbol{\theta}_i^{(k)}$ and $\mathbf{X}_i^{*(k)} = \mathbf{X}_{\mathcal{A}}^{(k)}\mathrm{diag}(\boldsymbol{\beta}_{i\mathcal{A}}^{(k)})$. Write

$$\boldsymbol{y}^{*(k)} = \left(\boldsymbol{y}_1^{*(k)\top}, \ldots, \boldsymbol{y}_{d^{(k)}}^{*\,(k)\top}\right)^\top \quad \text{and}$$
$$\mathbf{X}^{*(k)} = \left(\mathbf{X}_1^{*(k)\top}, \ldots, \mathbf{X}_{d^{(k)}}^{*\,(k)\top}\right)^\top.$$

Write $\boldsymbol{y}^{\mathcal{K}} = (\boldsymbol{y}^{*(k)}, k \in \mathcal{K})$ and $\mathbf{X}^{\mathcal{K}} = \mathrm{diag}(\mathbf{X}^{*(k)}, k \in \mathcal{K})$. Let $\boldsymbol{\alpha}_{\mathcal{A}}$ be the solution to the lasso problem

$$\begin{aligned}
\underset{\boldsymbol{\alpha}_{\mathcal{A}}}{\text{minimize}} \quad & \|\boldsymbol{y}^{\mathcal{K}} - \mathbf{X}^{\mathcal{K}}\boldsymbol{\alpha}_{\mathcal{A}}\|_2^2 \\
\text{subject to} \quad & \sum_{j\in\mathcal{A}} \alpha_j \leq \tau^{\mathcal{K}}, \quad \tau^{\mathcal{K}} \geq 0.
\end{aligned}$$

Let $\mathcal{A} = \{j : \alpha_j \neq 0\}$.

Step 2. Update $\left[\{\boldsymbol{\theta}_i^{(k)}\}_{i=1}^{d^{(k)}}\right]_{k\in\mathcal{K}}$. For each $k \in \mathcal{K}$ and $i = 1, \ldots, d^{(k)}$, let

$$\mathbf{Q}_i^{(k)} = \left(\boldsymbol{I}_{H_k}, \boldsymbol{\theta}_1^{(k)}, \ldots, \boldsymbol{\theta}_{i-1}^{(k)}\right)$$

and

$$\begin{aligned}
\boldsymbol{\theta}_i^{(k)} = & \left(\mathbf{I}_{H_k} - \mathbf{Q}_i^{(k)}\mathbf{Q}_i^{(k)\top}\mathbf{D}^{(k)}\right) \\
& \times \left(\mathbf{D}^{(k)}\right)^{-1} \boldsymbol{\Phi}^{(k)\top}\mathbf{X}_{\mathcal{A}}^{(k)}\mathrm{diag}\left(\boldsymbol{\alpha}_{\mathcal{A}}\right)\boldsymbol{\beta}_{i\mathcal{A}}^{(k)},
\end{aligned}$$

then normalize $\boldsymbol{\theta}_i^{(k)}$ so that $\boldsymbol{\theta}_i^{(k)\top}\mathbf{D}^{(k)}\boldsymbol{\theta}_i^{(k)} = 1$.

Step 3. Update $[\{\boldsymbol{\beta}_i^{(k)}\}_{i=1}^{d^{(k)}}]_{k\in\mathcal{K}}$. Let $\beta_{ij}^{(k)} = (d_k|\mathcal{K}|)^{-1/2}$ for $j \notin \mathcal{A}$. For each $k \in \mathcal{K}$ and $i = 1, \ldots, d^{(k)}$, let

$$\boldsymbol{\beta}_{i\mathcal{A}}^{(k)} = \left[\left\{\mathbf{X}_{\mathcal{A}}^{(k)}\text{diag}\,(\boldsymbol{\alpha}_{\mathcal{A}})\right\}^{\top}\mathbf{X}_{\mathcal{A}}^{(k)}\text{diag}\,(\boldsymbol{\alpha}_{\mathcal{A}})\right]^{-1} \times \left\{\mathbf{X}_{\mathcal{A}}^{(k)}\text{diag}\,(\boldsymbol{\alpha}_{\mathcal{A}})\right\}^{\top}\boldsymbol{\Phi}^{(k)}\boldsymbol{\theta}_i^{(k)}.$$

Normalize $\boldsymbol{\beta}_{i\mathcal{A}}^{(k)}$ so that $\sum_{k\in\mathcal{K}}\sum_{i=1}^{d^{(k)}}(\beta_{ij}^{(k)})^2 = 1$ for $j \in \mathcal{A}$.

Step 4. Iterate Steps 1 through 3 until convergence or until a maximum number of iterations is reached.

One way to measure the convergence is to use the maximum absolute distance between two consecutive solutions of $\boldsymbol{\alpha}$, that is, $\max_{j=1,\ldots,p}|\alpha_j(t+1) - \alpha_j(t)|$, where t is the index of iteration. In the numerical studies throughout this paper, we take the maximum number of iterations to be 30, and the tolerance level for the above convergence criterion to be 10^{-5}. Based on our limited experience, we find that the proposed algorithm usually takes less than 15 iterations to converge.

2.5 Tuning

For a given $\tau^{\mathcal{K}}$, denote the estimates of $\boldsymbol{\theta}_i^{(k)}$, $\boldsymbol{\beta}_i^{(k)}$ and $\boldsymbol{\alpha}$ by $\hat{\boldsymbol{\theta}}_i^{(k)}(\tau^{\mathcal{K}})$, $\hat{\boldsymbol{\beta}}_i^{(k)}(\tau^{\mathcal{K}})$ and $\hat{\boldsymbol{\alpha}}(\tau^{\mathcal{K}})$, respectively. Write $\hat{\boldsymbol{\alpha}}(\tau^{\mathcal{K}}) = \{\hat{\alpha}_1(\tau^{\mathcal{K}}), \ldots, \hat{\alpha}_p(\tau^{\mathcal{K}})\}^{\top}$ and let $\hat{\mathcal{A}}(\tau^{\mathcal{K}}) = \{j : \hat{\alpha}_j(\tau^{\mathcal{K}}) \neq 0\}$. In practice, the choice of the regularization parameter $\tau^{\mathcal{K}}$ is of great importance. Because the estimates are obtained using a penalized regression method, criteria that have been developed for selecting the tuning parameter for penalized regression can be applied. For computational easiness, in this section we propose to use a BIC-type criterion for choosing $\tau^{\mathcal{K}}$. Following Wang and Leng (2007) and Bondell and Li (2009), we define

$$\text{BIC}^{\mathcal{K}}\left(\tau^{\mathcal{K}}\right) = \log\left\{\text{RSS}^{\mathcal{K}}\left(\tau^{\mathcal{K}}\right)\right\} + p_e^{\mathcal{K}}\left(\tau^{\mathcal{K}}\right)\frac{\log n^{\mathcal{K}}}{n^{\mathcal{K}}}, \tag{2.13}$$

where

$$\text{RSS}^{\mathcal{K}}\left(\tau^{\mathcal{K}}\right) = \sum_{k\in\mathcal{K}}\sum_{i=1}^{d^{(k)}}\|\boldsymbol{\Phi}^{(k)}\hat{\boldsymbol{\theta}}_i^{(k)}\left(\tau^{\mathcal{K}}\right) - \mathbf{X}^{(k)}\text{diag}\left\{\hat{\boldsymbol{\alpha}}\left(\tau^{\mathcal{K}}\right)\right\}\hat{\boldsymbol{\beta}}_i^{(k)}\left(\tau^{\mathcal{K}}\right)\|_2^2,$$

$p_e^{\mathcal{K}}(\tau^{\mathcal{K}})$ denotes the effective number of parameters in the estimates of dimension reduction subspaces, and $n^{\mathcal{K}} = \sum_{k\in\mathcal{K}} d^{(k)}n_k$ is the effective sample size. We estimate $p_e^{\mathcal{K}}(\tau^{\mathcal{K}})$ by $|\hat{\mathcal{A}}(\tau^{\mathcal{K}})| \times \sum_{k\in\mathcal{K}} d^{(k)}$. Finally, we select $\tau^{\mathcal{K}}$ by minimizing $\text{BIC}^{\mathcal{K}}(\tau^{\mathcal{K}})$.

For simplicity, the structural dimensions $d^{(k)}$ are assumed to be known in this paper. Under the penalized regression framework, criterion-based approaches can potentially be used for selecting $d^{(k)}$. The investigation of the behavior of these criteria, in both numerical and theoretical aspects, is an interesting topic for a future study.

2.6 Properties

In this section we discuss the motivation behind the new methodology. It is convenient to re-express (2.6) in the equivalent Lagrangian form

$$\begin{aligned}\underset{\{\boldsymbol{\Theta}^{\mathcal{K}},\boldsymbol{\alpha}\}}{\text{minimize}}\quad & \sum_{k\in\mathcal{K}}\sum_{i=1}^{d^{(k)}}\|\boldsymbol{\Phi}^{(k)}\boldsymbol{\theta}_i^{(k)} - \mathbf{X}^{(k)}\text{diag}(\boldsymbol{\alpha})\boldsymbol{\beta}_i^{(k)}\|_2^2 + \lambda^{\mathcal{K}}\sum_{j=1}^{p}\alpha_j \\ \text{subject to}\quad & \boldsymbol{\theta}_i^{(k)\top}\boldsymbol{\Phi}^{(k)\top}\boldsymbol{\Phi}^{(k)}\boldsymbol{\theta}_i^{(k)} = n_k, \boldsymbol{\theta}_i^{(k)\top}\boldsymbol{\Phi}^{(k)\top}\boldsymbol{\Phi}^{(k)}\boldsymbol{\theta}_j^{(k)} = 0, \\ & j = 1, \ldots, i-1, i = 1, \ldots, d^{(k)}, k \in \mathcal{K},\end{aligned} \tag{2.14}$$

for some non-negative regularization parameter $\lambda^{\mathcal{K}}$. Let $\boldsymbol{\Xi}^{\mathcal{K}} = [\{\boldsymbol{\beta}_i^{(k)}\}_{i=1}^{d^{(k)}}]_{k\in\mathcal{K}}$ and assume for the moment that $[\{\boldsymbol{\theta}_i^{(k)}\}_{i=1}^{d^{(k)}}]_{k\in\mathcal{K}}$ is given. Then the above criterion becomes

$$\underset{\{\boldsymbol{\Xi}^{\mathcal{K}},\boldsymbol{\alpha}\}}{\text{minimize}}\quad \sum_{k\in\mathcal{K}}\sum_{i=1}^{d^{(k)}}\|\boldsymbol{\Phi}^{(k)}\boldsymbol{\theta}_i^{(k)} - \mathbf{X}^{(k)}\text{diag}(\boldsymbol{\alpha})\boldsymbol{\beta}_i^{(k)}\|_2^2 + \lambda^{\mathcal{K}}\sum_{j=1}^{p}\alpha_j. \tag{2.15}$$

Instead of estimating $\boldsymbol{\Xi}^{\mathcal{K}}$ and $\boldsymbol{\alpha}$ jointly, a more direct way to proceed is to use the group lasso (Yuan and Lin 2006) by minimizing

$$\sum_{k\in\mathcal{K}}\sum_{i=1}^{d^{(k)}}\|\boldsymbol{\Phi}^{(k)}\boldsymbol{\theta}_i^{(k)} - \mathbf{X}^{(k)}\boldsymbol{\beta}_i^{(k)}\|_2^2 + \mu^{\mathcal{K}}\sum_{j=1}^{p}\|\boldsymbol{\beta}_j^{\mathcal{K}}\|_2 \tag{2.16}$$

with respect to $\boldsymbol{\Xi}^{\mathcal{K}}$. As before, $\boldsymbol{\beta}_j^{\mathcal{K}}$ represents the set of coefficients corresponding to the j-th coordinate. From Lemma 2 of Lin and Zhang (2006), we know that an equivalent form is

$$\begin{aligned} \underset{\{\Xi^{\mathcal{K}},\xi\}}{\text{minimize}} \quad & \sum_{k\in\mathcal{K}}\sum_{i=1}^{d^{(k)}} \|\Phi^{(k)}\theta_i^{(k)} - \mathbf{X}^{(k)}\beta_i^{(k)}\|_2^2 \\ & + \sum_{j=1}^{p} \xi_j^{-1}\|\beta_j^{\mathcal{K}}\|_2^2 + \nu^{\mathcal{K}}\sum_{j=1}^{p}\xi_j \\ \text{subject to} \quad & \xi_j \geq 0,\ j = 1,\ldots,p, \end{aligned} \tag{2.17}$$

where $\xi = (\xi_1,\ldots,\xi_p)^\top$. Define $\xi^{-1} = (\xi_1^{-1},\ldots,\xi_p^{-1})^\top$ and $\beta_i^{*(k)} = \text{diag}(\xi^{-1})\beta_i^{(k)}$. Let $\Xi^{*\mathcal{K}} = [\{\beta_i^{*(k)}\}_{i=1}^{d^{(k)}}]_{k\in\mathcal{K}}$. Then we arrive at

$$\begin{aligned} \underset{\{\Xi^{*(k)},\xi\}}{\text{minimize}} \quad & \sum_{k\in\mathcal{K}}\sum_{i=1}^{d^{(k)}} \|\Phi^{(k)}\theta_i^{(k)} - \mathbf{X}^{(k)}\text{diag}(\xi)\beta_i^{*(k)}\|_2^2 \\ & + \sum_{j=1}^{p} \|\beta_j^{*\mathcal{K}}\|_2^2 + \nu^{\mathcal{K}}\sum_{j=1}^{p}\xi_j \\ \text{subject to} \quad & \xi_j \geq 0,\ j = 1,\ldots,p. \end{aligned} \tag{2.18}$$

Comparing (2.18) with (2.15), we see that our estimation procedure amounts to the group lasso. This is because, under the identifiability constraint, $\sum_{j=1}^{p}\|\beta_j^{*\mathcal{K}}\|_2^2 = p$. However, the performance of our method is empirically observed to be superior to the standard implementation of the group lasso, which is sometimes instable and computationally intensive within the context of joint estimation.

3 Simulation studies

In this section, we use simulation examples to evaluate the performance of the multiple-population shrinkage sliced inverse regression we proposed in terms of estimation accuracy and predictor selection, and compare it with the conditional shrinkage sliced inverse regression. For the latter, we consider both the naive conditional shrinkage estimator in (2.4) and its improved version in (2.5), which is a special case of the multiple-population estimator when $K = 1$.

Throughout B-spline basis functions are used as the transformation functions, because empirically it has been found that they are superior to the slice indicator functions (Wang and Zhu 2013). In particular, we use a cubic spline with $H_k - 4$ inner knots and fix H_k at 10. The entire R code is available from the authors upon request.

To evaluate the accuracy of each method, we use the distance measure suggested by Li et al. (2005). Specifically, let $\mathcal{S}_1$ and $\mathcal{S}_2$ be two subspaces of $\mathbb{R}^p$. Then we adopt the criterion

$$\text{DIST}(\mathcal{S}_1,\mathcal{S}_2) = \|\mathbf{P}_{\mathcal{S}_1} - \mathbf{P}_{\mathcal{S}_2}\|_F,$$

where $\mathbf{P}$ is the orthogonal projection operator and $\|\cdot\|_F$ denotes the Frobenius norm, that is, the maximum singular value of a matrix. This measure is similar to the one used in Xia et al. (2002). For each simulation configuration, we run 200 simulation samples and take the average of the aforementioned criterion. We also employ the average model size: the average number of identified predictors; the true positive rate: the ratio of the number of correctly identified predictors to the number of relevant predictors; and the false positive rate: the ratio of the number of falsely identified predictors to the number of irrelevant predictors, for assessing the performance of a method for selecting predictors.

We let $\mathbf{0}_p$ denote the p-dimensional vector of zeros, and $\mathbf{e}_i$ the p-dimensional vector whose i-th element is 1 and other elements are all 0, $i = 1,\ldots,p$. Let $\Sigma^{(k)} = (\Sigma_{ij}^{(k)}) = \text{cov}(\mathbf{X}^{(k)})$.

Example 1

$$Y^{(k)} = \frac{\mathbf{e}_1^\top\mathbf{X}^{(k)}}{0.5 + \left(\mathbf{e}_2^\top\mathbf{X}^{(k)} + 1.5\right)^2} + \epsilon^{(k)},$$

where $\mathbf{X}^{(k)} \sim N(\mathbf{0}_p, \Sigma^{(k)})$ with $\Sigma_{ij}^{(k)} = 0.5^{|i-j|}$, $1 \leq i, j \leq p = 8$, $\epsilon^{(k)} \sim N(0,\sigma_k^2)$, $k = 1, 2$. Two cases are explored: $(\sigma_1,\sigma_2) = (0.5, 0.5)$ and $(\sigma_1,\sigma_2) = (0.5, 0.8)$. In this example, we have $\beta^{(1)} = \beta^{(2)} = (\mathbf{e}_1,\mathbf{e}_2)$. We take $n_1 = n_2 = 80$.

Example 2

$$\begin{aligned} Y^{(1)} &= \frac{\mathbf{e}_1^\top\mathbf{X}^{(1)}}{0.5 + \left(\mathbf{e}_2^\top\mathbf{X}^{(1)} + 1.5\right)^2} + \epsilon^{(1)}, \\ Y^{(2)} &= \frac{\mathbf{e}_2^\top\mathbf{X}^{(2)}}{0.5 + \left(\mathbf{e}_1^\top\mathbf{X}^{(2)} + 1.5\right)^2} + \epsilon^{(2)}. \end{aligned}$$

The setup is the same as in Example 1, except here $\sigma_1 = \sigma_2 = 0.8$. In this example, we have $\beta^{(1)} = \beta^{(2)} = (\mathbf{e}_1,\mathbf{e}_2)$. We take $n_1 = n_2 = 60$ and $n_1 = n_2 = 80$.

Example 3

$$\begin{aligned} Y^{(1)} &= \frac{\mathbf{e}_1^\top\mathbf{X}^{(1)}}{0.5 + \left(\mathbf{e}_2^\top\mathbf{X}^{(1)} + 1.5\right)^2} + \epsilon^{(1)}, \\ Y^{(2)} &= \mathbf{e}_1^\top\mathbf{X}^{(2)} \times \left(\mathbf{e}_1^\top\mathbf{X}^{(2)} + \mathbf{e}_2^\top\mathbf{X}^{(2)} + 1\right) + \epsilon^{(2)}. \end{aligned}$$

The setup is the same as in Example 1, except that $\sigma_1 = \sigma_2 = 0.5$. In this example, we have $\beta^{(1)} = \beta^{(2)} = (\mathbf{e}_1,\mathbf{e}_2)$. We take $n_1 = n_2 = 80$.

Table 1 Summary of Examples 1 and 2

	$\beta^{(1)}$			$\beta^{(2)}$		
	CS-SIR-N	CS-SIR-I	MP-SIR	CS-SIR-N	CS-SIR-I	MP-SIR
Example 1: $(\sigma_1, \sigma_2) = (0.5, 0.5)$						
DIST	0.7561 (0.5113)	0.5904 (0.5865)	0.2621 (0.4668)	0.7465 (0.5146)	0.5597 (0.5872)	0.2873 (0.4986)
MS	3.1500	2.6000	2.3100	3.0750	2.6050	2.3100
TPR	0.8900	0.9025	0.9650	0.8750	0.9100	0.9650
FPR	0.2283	0.1325	0.0633	0.2208	0.1308	0.0633
Example 1: $(\sigma_1, \sigma_2) = (0.5, 0.8)$						
DIST	0.7561 (0.5113)	0.5904 (0.5865)	0.3596 (0.5440)	0.9901 (0.5076)	0.8949 (0.6102)	0.4150 (0.5965)
MS	3.1500	2.6000	2.3200	2.5050	2.6600	2.3200
TPR	0.8900	0.9025	0.9350	0.6850	0.8075	0.9350
FPR	0.2283	0.1325	0.0750	0.1891	0.1741	0.0750
Example 2: $n_1 = n_2 = 60$						
DIST	1.2387 (0.3629)	1.1528 (0.5201)	0.7219 (0.6632)	1.1516 (0.4499)	1.0126 (0.6119)	0.7042 (0.6618)
MS	2.6100	2.6500	2.4350	2.3600	2.6050	2.4350
TPR	0.5750	0.6850	0.8625	0.5775	0.7425	0.8625
FPR	0.2433	0.2133	0.1183	0.2008	0.1866	0.1183
Example 2: $n_1 = n_2 = 80$						
DIST	1.0364 (0.4712)	0.9264 (0.6304)	0.4276 (0.5877)	0.9735 (0.5213)	0.8940 (0.6228)	0.4301 (0.5788)
MS	2.4550	2.5000	2.3650	2.4800	2.5850	2.3650
TPR	0.6550	0.7725	0.9525	0.7000	0.7950	0.9525
FPR	0.1908	0.1591	0.0766	0.1800	0.1658	0.0766

The average distance measure (DIST) with standard error in parentheses, the average model size (MS), the true positive rate (TPR), and the false positive rate (FPR), based on 200 data replications, are reported
CS-SIR-N the naive conditional shrinkage sliced inverse regression, *CS-SIR-I* the improved conditional shrinkage sliced inverse regression, *MP-SIR* multiple-population sliced inverse regression

Table 2 Summary of Examples 3 and 4

	$\beta^{(1)}$			$\beta^{(2)}$		
	CS-SIR-N	CS-SIR-I	MP-SIR	CS-SIR-N	CS-SIR-I	MP-SIR
Example 3						
DIST	0.7561 (0.5113)	0.5904 (0.5865)	0.4585 (0.6017)	1.1035 (0.4938)	0.9823 (0.6444)	0.5184 (0.6362)
MS	3.1500	2.6000	2.3050	2.1100	2.5450	2.3050
TPR	0.8900	0.9025	0.8925	0.5300	0.7225	0.8925
FPR	0.2283	0.1325	0.0866	0.1750	0.1833	0.0866
Example 4						
DIST	0.8414 (0.3455)	0.7486 (0.3603)	0.6227 (0.3119)	1.0251 (0.3486)	0.9738 (0.3440)	0.7720 (0.3372)
MS	4.4350	3.5550	3.0650	3.2800	2.9400	3.0650
TPR	0.9783	0.9516	0.9466	0.7550	0.7733	0.9466
FPR	0.3000	0.1400	0.0450	0.2030	0.1240	0.0450

The average distance measure (DIST) with standard error in parentheses, the average model size (MS), the true positive rate (TPR), and the false positive rate (FPR), based on 200 data replications, are reported
CS-SIR-N the naive conditional shrinkage sliced inverse regression, *CS-SIR-I* the improved conditional shrinkage sliced inverse regression, *MP-SIR* multiple-population sliced inverse regression

Table 3 Summary of Example 5

	$\boldsymbol{\beta}^{(1)}$			$\boldsymbol{\beta}^{(2)}$			$\boldsymbol{\beta}^{(3)}$		
	CS-SIR-N	CS-SIR-I	MP-SIR	CS-SIR-N	CS-SIR-I	MP-SIR	CS-SIR-N	CS-SIR-I	MP-SIR
Case 1: $p = 8$									
DIST	0.2515 (0.1397)	0.2390 (0.1408)	0.1826 (0.1281)	0.2382 (0.1383)	0.2211 (0.1361)	0.1753 (0.1296)	0.7309 (0.5253)	0.5795 (0.5678)	0.1396 (0.3424)
MS	3.0350	2.8050	2.1850	2.8900	2.6400	2.1850	2.9950	2.6250	2.1850
TPR	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.8800	0.9050	1.0000
FPR	0.1725	0.1341	0.0308	0.1483	0.1066	0.0308	0.2058	0.1358	0.0308
Case 1: $p = 16$									
DIST	0.2559 (0.1455)	0.2377 (0.1405)	0.1810 (0.1268)	0.2708 (0.1670)	0.2358 (0.1541)	0.1735 (0.1350)	0.8383 (0.5034)	0.8183 (0.6278)	0.1530 (0.3749)
MS	3.2850	2.9150	2.2000	3.5650	2.9500	2.2000	2.2500	2.5850	2.2000
TPR	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.7300	0.8175	1.0000
FPR	0.0917	0.0653	0.0142	0.1117	0.0678	0.0142	0.0564	0.0678	0.0142
Case 2: $p = 8$									
DIST	0.3322 (0.1325)	0.3235 (0.1310)	0.3060 (0.1320)	0.3326 (0.1358)	0.3301 (0.1384)	0.3054 (0.1368)	1.2264 (0.2517)	1.2390 (0.2347)	1.0754 (0.2627)
MS	4.9900	4.7900	4.3950	4.9350	4.7900	4.3950	4.2450	3.4700	4.3950
TPR	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.7737	0.7112	1.0000
FPR	0.2475	0.1975	0.0987	0.2337	0.1975	0.0987	0.2875	0.1562	0.0987
Case 2: $p = 16$									
DIST	0.3653 (0.1438)	0.3528 (0.1398)	0.3209 (0.1352)	0.3797 (0.1485)	0.3591 (0.1352)	0.3261 (0.1258)	1.3469 (0.1753)	1.3420 (0.2042)	1.0594 (0.2823)
MS	5.5550	5.2750	4.4050	6.0400	5.3100	4.4050	2.9300	3.4500	4.4050
TPR	1.0000	1.0000	0.9987	0.9987	1.0000	0.9987	0.5425	0.6262	0.9987
FPR	0.1295	0.1062	0.0341	0.1704	0.1091	0.0341	0.0633	0.0787	0.0341

The average distance measure (DIST) with standard error in parentheses, the average model size (MS), the true positive rate (TPR), and the false positive rate (FPR), based on 200 data replications, are reported

CS-SIR-N the naive conditional shrinkage sliced inverse regression, *CS-SIR-I* the improved conditional shrinkage sliced inverse regression, *MP-SIR* multiple-population sliced inverse regression

Example 4

$$Y^{(1)} = \frac{\boldsymbol{e}_1^\top \boldsymbol{X}^{(1)} + \boldsymbol{e}_2^\top \boldsymbol{X}^{(1)}}{0.5 + \left(\boldsymbol{e}_3^\top \boldsymbol{X}^{(1)} + 1.5\right)^2} + \epsilon^{(1)},$$
$$Y^{(2)} = \left(\boldsymbol{e}_1^\top \boldsymbol{X}^{(2)} + \boldsymbol{e}_2^\top \boldsymbol{X}^{(2)} + 1\right) \times \boldsymbol{e}_3^\top \boldsymbol{X}^{(2)} + \epsilon^{(2)}.$$

The setup is the same as in Example 1, except that $\sigma_1 = \sigma_2 = 0.5$. In this example, we have $\boldsymbol{\beta}^{(1)} = \boldsymbol{\beta}^{(2)} = (\boldsymbol{e}_1 + \boldsymbol{e}_2, \boldsymbol{e}_3)$. We take $(n_1, n_2) = (80, 100)$.

Note that $K = 2$ in the above examples. Furthermore, the two regression functions are the same in Example 1, have the same shape in Example 2, and are very different in Examples 3 and 4. The simulation results from these four examples are summarized in Tables 1 and 2. From the tables, we can see that the overall best performer is the proposed multiple-population estimator, which is especially true in Example 1 with $\sigma_1 = \sigma_2 = 0.5$ and Example 2 with $n_1 = n_2 = 80$. On the other hand, the inferiority of the conditional estimators are manifested in Example 1 with $(\sigma_1, \sigma_2) = (0.5, 0.8)$, and Examples 2–4, where they have poor performance in one or both populations. We can also see that the improved conditional shrinkage estimator generally outperforms the naive one. Finally, the simulation results in Example 2 also show that there is a great improvement in performance as the sample size increases. To further examine the performance, we consider one more example with $K = 3$.

Example 5

$$Y^{(1)} = \exp\left\{0.5 \times (\boldsymbol{b}_1 + \boldsymbol{b}_2)^\top \boldsymbol{X}^{(1)}\right\} + \epsilon^{(1)},$$
$$Y^{(2)} = 2 \times \sin\left\{0.25 \times (\boldsymbol{b}_1 + \boldsymbol{b}_2)^\top \boldsymbol{X}^{(2)}\right\} + \epsilon^{(2)},$$
$$Y^{(3)} = \frac{\boldsymbol{b}_1^\top \boldsymbol{X}^{(3)}}{0.5 + \left(\boldsymbol{b}_2^\top \boldsymbol{X}^{(3)} + 1.5\right)^2} + \epsilon^{(3)},$$

where $\boldsymbol{X}^{(k)} \sim N(\boldsymbol{0}_p, \boldsymbol{\Sigma}^{(k)})$ with $\Sigma_{ij}^{(k)} = 0.5^{|i-j|}$, $1 \le i, j \le p$, $\epsilon^{(k)} \sim N(0, 0.5^2)$, $k = 1, 2, 3$. In this example, we have $\boldsymbol{\beta}^{(1)} = \boldsymbol{\beta}^{(2)} = \boldsymbol{b}_1 + \boldsymbol{b}_2$ and $\boldsymbol{\beta}^{(3)} = (\boldsymbol{b}_1, \boldsymbol{b}_2) \neq \boldsymbol{\beta}^{(1)}$. We explore two cases with different sparsity levels: (1) $\boldsymbol{b}_1 = \boldsymbol{e}_1, \boldsymbol{b}_2 = \boldsymbol{e}_2$, and (2) $\boldsymbol{b}_1 = \boldsymbol{e}_1 + \boldsymbol{e}_2, \boldsymbol{b}_2 = \boldsymbol{e}_3 + \boldsymbol{e}_4$. To illustrate the sensitivity of the method to the dimension, we let $p \in \{8, 16\}$. Finally, we take $n_1 = n_2 = n_3 = 80$. The simulation results are summarized in Table 3. We observe qualitatively similar results to those reported in the previous examples. In addition, we can see that the performance of all the three competitors deteriorates when the dimension p increases from 8 up to 16, and when the sparsity level increases from case (1) to case (2).

Finally, in order to numerically compare the proposed method with the group lasso algorithm mentioned in Sect. 2.6, we consider Example 1 with $\sigma_1 = \sigma_2 = 0.5$, Example 4, and Example 5 with $p = 16$. Tables 4 and 5 show the simulation results as well as the average run time (CPU seconds) for the two algorithms. As we can see, the group lasso algorithm is computationally more intensive than the proposed method. Further, it is numerically very instable: it outperforms the proposed method in Example 1 ($\sigma_1 = \sigma_2 = 0.5$), but performs poorly and is inferior to the proposed method in Example 4 and Example 5 ($p = 16$).

Table 4 Comparisons between the propose method and the group lasso algorithm for Example 1 with $\sigma_1 = \sigma_2 = 0.5$ and Example 4

Timings	Example 1 ($\sigma_1 = \sigma_2 = 0.5$)			
	MP-SIR 9.9690		GLASSO 38.9115	
	$\boldsymbol{\beta}^{(1)}$	$\boldsymbol{\beta}^{(2)}$	$\boldsymbol{\beta}^{(1)}$	$\boldsymbol{\beta}^{(2)}$
DIST	0.2621	0.2873	0.0223	0.0092
	(0.4668)	(0.4986)	(0.1390)	(0.0799)
MS	2.3100	2.3100	2.0600	2.0150
TPR	0.9650	0.9650	1.0000	1.0000
FPR	0.0633	0.0633	0.0100	0.0025
Timings	Example 4			
	MP-SIR 11.6886		GLASSO 32.5398	
	$\boldsymbol{\beta}^{(1)}$	$\boldsymbol{\beta}^{(2)}$	$\boldsymbol{\beta}^{(1)}$	$\boldsymbol{\beta}^{(2)}$
DIST	0.6227	0.7720	1.1108	1.0593
	(0.3119)	(0.3372)	(0.4114)	(0.4175)
MS	3.0650	3.0650	2.8950	2.8600
TPR	0.9466	0.9466	0.7933	0.8133
FPR	0.0450	0.0450	0.1030	0.0840

The average distance measure (DIST) with standard error in parentheses, the average model size (MS), the true positive rate (TPR), and the false positive rate (FPR), based on 200 data replications, are reported *MP-SIR* multiple-population sliced inverse regression, *GLASSO* the group lasso algorithm

4 Australian athletes data

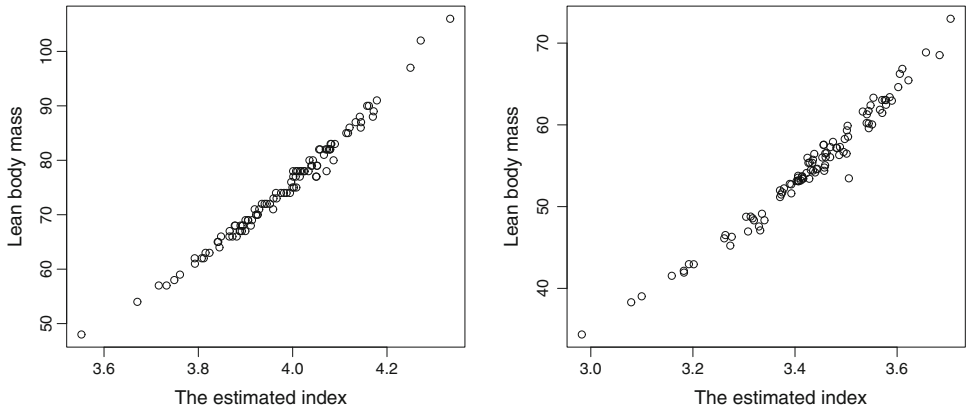
We next consider a data set on 102 male and 100 female athletes collected at the Australian Institute of Sport (Weisberg 2005, Sect. 6.4). We are interested in describing the conditional distribution of lean body mass, given eight predictors, height, weight, sum of skin folds, red cell count, white cell count, plasma ferritin concentration, hematocrit and hemoglobin, separately for each sex. It is natural to view this task as a two-population dimension-reduction problem, with the sex variable being the population indicator.

Chiaromonte et al. (2002) used this data set to illustrate dimension reduction in regression with categorical predic-

Table 5 Comparisons between the propose method and the group lasso algorithm for Example 5 with $p = 16$

Timings	Case (1)					
	MP-SIR 10.3779			GLASSO 43.3955		
	$\boldsymbol{\beta}^{(1)}$	$\boldsymbol{\beta}^{(2)}$	$\boldsymbol{\beta}^{(3)}$	$\boldsymbol{\beta}^{(1)}$	$\boldsymbol{\beta}^{(2)}$	$\boldsymbol{\beta}^{(3)}$
DIST	0.1810	0.1735	0.1530	0.5424	0.5407	1.0980
	(0.1268)	(0.1350)	(0.3749)	(0.3402)	(0.4208)	(0.5442)
MS	2.2000	2.2000	2.2000	5.1400	4.2150	4.5250
TPR	1.0000	1.0000	1.0000	0.9925	0.9025	0.8425
FPR	0.0142	0.0142	0.0142	0.2253	0.1721	0.2028
Timings	Case (2)					
	MP-SIR 11.3456			GLASSO 42.7013		
	$\boldsymbol{\beta}^{(1)}$	$\boldsymbol{\beta}^{(2)}$	$\boldsymbol{\beta}^{(3)}$	$\boldsymbol{\beta}^{(1)}$	$\boldsymbol{\beta}^{(2)}$	$\boldsymbol{\beta}^{(3)}$
DIST	0.3209	0.3261	1.0594	0.7520	0.7418	1.2706
	(0.1352)	(0.1258)	(0.2823)	(0.3030)	(0.3473)	(0.2705)
MS	4.4050	4.4050	4.4050	6.2700	5.5100	4.8650
TPR	0.9987	0.9987	0.9987	0.8737	0.8487	0.7862
FPR	0.0341	0.0341	0.0341	0.2312	0.1762	0.1433

The average distance measure (DIST) with standard error in parentheses, the average model size (MS), the true positive rate (TPR), and the false positive rate (FPR), based on 200 data replications, are reported *MP-SIR* multiple-population sliced inverse regression, *GLASSO* the group lasso algorithm



(a) Males **(b)** Females

Fig. 1 Plots of the estimated index from the joint estimation procedure versus the dependent variable for the male population (*left panel*) and the female population (*right panel*) **a** Males **b** Females

tors. They proposed partial dimension reduction in which the reduction of the vector of continuous predictors is done simultaneously for all distinct levels of categorical predictors. They applied partial sliced inverse regression to a restricted regression involving only five of the eight predictors mentioned above, and found that a single linear combination of the predictors is sufficient to describe both the male and female regressions. However, partial dimension reduction in general is not a good solution to the problem of multiple-population dimension reduction, because the same reduction applies to all levels of the categorical predictors, ignoring population-specific effects.

On the other hand, it is more appropriate to assume that there are common factors that are associated with lean body mass in the two regressions for males and females, that is, to assume the joint sparsity assumption. Before we continue, we log-transform each of the eight predictors in order to insure the linearity condition, following Chiaromonte et al. (2002) and Cook (2004). Using chi-squared tests (e.g., Li 1991), we infer that the central subspace for each population is one-dimensional, that is, $d^{(1)} = d^{(2)} = 1$. We then apply our multiple-population estimation method to the two regressions of lean body mass on the transformed predictors. The estimates for males and females are (0, 0.995, −0.100,

$0, -0.001, -0.003, 0, 0)^\top$ and $(0, 0.988, -0.157, 0, -0.009, -0.001, 0, 0)^\top$, respectively. We can see that the two predictors, weight and sum of skin folds, are highly relevant to lean body mass in both the regressions, and this is consistent with the conclusion drawn by Cook (2004) using testing procedures. We can also see that the two estimates are very close to each other, which indicates that for this particular data set, it might be reasonable to assume that the male and female regressions share a common dimension-reduction subspace, as was observed in Chiaromonte et al. (2002). Specifically, the squared cosine between the two direction estimates is 0.9965. Figure 1 shows the plot of the estimated index versus the dependent variable for each population. We can see that the two link functions are nearly linear. If we consider the union of males and females, the marginal direction estimate (without regularization) is

$$(0.0933, 0.9164, -0.1509, 0.1889, -0.004, 0.0056, -0.2747, 0.1315)^\top,$$

and the squared cosines between this estimate and the two previous ones are 0.8592 and 0.8626, respectively.

References

Bernard-Michel, C., Gardes, L., Girard, S.: A note on sliced inverse regression with regularizations. Biometrics **64**, 982–984 (2008)

Bernard-Michel, C., Gardes, L., Girard, S.: Gaussian regularized sliced inverse regression. Stat. Comput. **19**, 85–98 (2009)

Bondell, H.D., Li, L.: Shrinkage inverse regression estimation for model-free variable selection. J. R. Stat. Soc. Ser. B **71**, 287–299 (2009)

Chavent, M., Kuentz, V., Liquet, B., Saracco, J.: Sliced inverse regression for stratified population. Commun. Stat.-Theory Methods **40**, 1–22 (2011)

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

Chiaromonte, F., Cook, R.D., Li, B.: Sufficient dimension reduction in regressions with categorical predictors. Ann. Stat. **30**, 475–497 (2002)

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

Cook, R.D.: Testing predictor contributions in sufficient dimension reduction. Ann. Stat. **32**, 1061–1092 (2004)

Cook, R.D., Forzani, L.: Likelihood-based sufficient dimension reduction. J. Am. Stat. Assoc. **104**, 197–208 (2009)

Cook, R.D., Ni, L.: Sufficient dimension reduction via inverse regression: a minimum discrepancy approach. J. Am. Stat. Assoc. **100**, 410–428 (2005)

Cook, R.D., Weisberg, S.: Discussion of "Sliced inverse regression for dimension reduction" by Ker-Chau Li. J. Am. Stat. Assoc. **86**, 328–332 (1991)

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

Lee, K., Li, B., Chiaromonte, F.: A general theory for nonlinear sufficient dimension reduction: formulation and estimation. Ann Stat **41**, 221–249 (2013)

Li, B., Zha, H., Chiaromonte, F.: Contour regression: a general approach to dimension reduction. Ann. Stat. **33**, 1580–1616 (2005)

Li, B., Kim, M., Altman, N.: On dimension folding of matrix- or array-valued statistical objects. Ann. Stat. **38**, 1094–1121 (2010)

Li, B., Wang, S.: On directional regression for dimension reduction. J. Am. Stat. Assoc. **102**, 997–1008 (2007)

Li, K.C.: Sliced inverse regression for dimension reduction. J. Am. Stat. Assoc. **86**, 316–327 (1991)

Li, L., Yin, X.: Sliced inverse regression with regularizations. Biometrics **64**, 124–131 (2008)

Lin, Y., Zhang, H.H.: Component selection and smoothing in multivariate nonparametric regression. Ann. Stat. **34**, 2272–2297 (2006)

Lounici, K., Pontil, M., Tsybakov, A.B., Van De Geer, S.: Taking advantage of sparsity in multi-task learning. arXiv preprint arXiv:0903.1468 (2009)

Ni, L., Cook, R.D., Tsai, C.L.: A note on shrinkage sliced inverse regression. Biometrika **92**, 242–247 (2005)

Scrucca, L.: Class prediction and gene selection for DNA microarrays using regularized sliced inverse regression. Comput. Stat. Data Anal. **52**, 438–451 (2007)

Tibshirani, R.: Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B **58**, 267–288 (1996)

Wang, H., Leng, C.: Unified LASSO estimation by least squares approximation. J. Am. Stat. Assoc. **102**, 1039–1048 (2007)

Wang, T., Zhu, L.X.: Sparse sufficient dimension reduction using optimal scoring. Comput. Stat. Data Anal. **57**, 223–232 (2013)

Weisberg, S.: Applied Linear Regression. Wiley, New York (2005)

Wu, Y., Li, L.: Asymptotic properties of sufficient dimension reduction with a diverging number of predictors. Statistica Sinica **31**, 707–730 (2011)

Xia, Y., Tong, H., Li, W.K., Zhu, L.X.: An adaptive estimation of dimension reduction space. J. R. Stat. Soc. Ser. B **64**, 363–410 (2002)

Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. J. R. Stat. Soc. Ser. B **68**, 49–67 (2006)

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

Zhu, L.P., Zhu, L.X., Feng, Z.H.: Dimension reduction in regressions through cumulative slicing estimation. J. Am. Stat. Assoc. **105**, 1455–1466 (2010)