Sparse Sliced Inverse Regression for High Dimensional Data

Qian Lin

Center of Mathematical Sciences and Applications
Harvard University
Cambridge, MA 02138
USA

E-mail: qianlin88@gmail.com

Zhigen Zhao

Department of Statistical Science
Fox School of Business
Temple University
Philadelphia, PA 19122

E-mail: zhaozhg@temple.edu

Jun S. Liu

Department of Statistics
Harvard University
Cambridge, MA 02138-2901
USA

E-mail: jliu@stat.harvard.edu

Summary. For multiple index models, it has recently been shown that the sliced inverse regression (SIR) is consistent for estimating the sufficient dimension reduction (SDR) subspace if and only if the dimension p and sample size n satisfies that $\rho = \lim \frac{p}{n} = 0$. Thus, when p is of the same or a higher order of n, additional assumptions such as sparsity have to be imposed in order to ensure consistency for SIR. By constructing artificial response variables made up from top eigenvectors of the estimated conditional covariance matrix, $\widehat{var}(\mathbb{E}[\boldsymbol{x}|y])$, we introduce a simple Lasso regression method to obtain an estimate of the SDR subspace. The resulting algorithm, Lasso-SIR, is shown to be consistent and achieve the optimal convergence rate under certain sparsity conditions when p is of order $o(n^2\lambda^2)$ where λ is the generalized signal noise ratio. We also demonstrate the superior performance of Lasso-SIR compared with existing approaches via extensive numerical studies and several real data examples.

1. Introduction

Dimension reduction and variable selection have become indispensable steps for most data analysts in this big data era, where thousands or even millions of features are easily obtained for only hundreds or thousands of samples (n). With these ultra high-dimensional data, an effective modeling strategy is to assume that only a few features and/or a few linear combinations of these features carry the information that researchers are interested in. Namely,

one can consider the following *multiple index model* [Li, 1991]:

$$y = f(\boldsymbol{\beta}_1^\tau \boldsymbol{x}, \boldsymbol{\beta}_2^\tau \boldsymbol{x}, ..., \boldsymbol{\beta}_d^\tau \boldsymbol{x}, \epsilon), \tag{1}$$

where $\boldsymbol{x}$ follows a p dimensional elliptical distribution with covariance matrix $\boldsymbol{\Sigma}$, the β_i's are unknown projection vectors, d is unknown but is assumed to be much smaller than p, and the error ϵ is independent of $\boldsymbol{x}$ and has mean 0. When p is very large, it is perhaps reasonable to further restrict each β_i to be a sparse vector.

Since the introduction of the sliced inverse regression (SIR) method by Li [1991], many sufficient dimension reduction (SDR, a term coined by Cook [1998]) methods have been proposed to estimate the space spanned by $(\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_d)$ with few assumptions on the link function $f(\cdot)$. With the multiple index model (1), the objective of all SDR methods is to find the minimal subspace $\mathcal{S} \subseteq \mathbb{R}^p$ such that $y \perp\!\!\!\perp \boldsymbol{x} \mid P_{\mathcal{S}}\boldsymbol{x}$, where $P_{\mathcal{S}}$ stands for the projection operator to the subspace $\mathcal{S}$. When the dimension of $\boldsymbol{x}$ is moderate, all SDR methods including SIR, which is most popular due to its simplicity and computational efficiency, have shown successes in various application fields [Xia et al., 2002, Ni et al., 2005, Li and Nachtsheim, 2006, Li, 2007, Zhu et al., 2006]. However, these methods were previously known to work only when the sample size n grows much faster than the dimension p, an assumption that becomes inappropriate for many modern datasets, such as those from biomedical researches. It is thus of great interest to have a thorough investigation of "the behavior of these SDR estimators when n is not large relative to p", as raised by Cook et al. [2012].

Lin et al. [2015] made an attempt to address the aforementioned challenge for SIR. They showed that, under mild conditions, the SIR estimate of the central space is consistent if and only if $\rho_n = \lim \frac{p}{n}$ goes to zero as n grows. Additionally, they showed that the convergence rate of the SIR estimate of the central space (without any sparsity assumption) is ρ_n. When p is much greater than n, however, certain constraints must be imposed in order for SIR to be consistent. The sparsity assumption, i.e., the number of active variables (s) must be an order of magnitude smaller than n and p, appears to be a reasonable one. In a follow-up work, Neykov et al. [2016] studied the sign support recovery problem of the single index model ($d = 1$), suggesting that the correct optimal convergence rate for estimating the central space might be $\frac{s\log(p)}{n}$, a speculation that is partially confirmed in Lin et al. [2016]. It is shown that, for multiple index models with bounded dimension d and the identity covariance matrix, the optimal rate for estimating the central space is $\frac{ds+s\log(p/s)}{n\lambda}$, where λ is the smallest non-zero eigenvalue of $var(\mathbb{E}[\boldsymbol{x}|y])$. They further showed that the Diagonal-Thresholding algorithm proposed in Lin et al. [2015] achieves the optimal rate for the single index model with the identity covariance matrix.

The main idea. In this article, we introduce an efficient Lasso variant of SIR for the multiple index model (1) with a general covariance matrix $\boldsymbol{\Sigma}$. Consider the single index model: $y = f(\boldsymbol{\beta}^\tau \boldsymbol{x}, \epsilon)$. Let $\boldsymbol{\eta}$ be the eigenvector associated with the largest eigenvalue of $var(\mathbb{E}[\boldsymbol{x}|y])$. Since $\boldsymbol{\beta} \propto \boldsymbol{\Sigma}^{-1}\boldsymbol{\eta}$, there are two immediate ways to estimate the space spanned by $\boldsymbol{\beta}$. The first approach, as discussed in Lin et al. [2015], estimates the inverse of $\boldsymbol{\Sigma}$ and $\boldsymbol{\eta}$ separately (see Algorithm 1). The second one avoids estimating the inverse of the covariance matrix $\boldsymbol{\Sigma}$ directly by solving the following penalized least square problem $\|\frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\tau\boldsymbol{\beta} - \boldsymbol{\eta}\|_2^2 + \mu\|\boldsymbol{\beta}\|_1$, where $\boldsymbol{X}$ is the $p \times n$ covariate matrix formed by the n samples (see Algorithm 2). However,

similar to most L_1 penalized approaches for nonlinear models, theoretical underpinning of this approach has not been well understood. Since these two approaches provide good estimates compared with other existing approaches (e.g.,Li [1991], Li and Nachtsheim [2006], Li [2007]), we set them as benchmarks for further comparisons.

We note that an eigenvector $\widehat{\boldsymbol{\eta}}$ of $\widehat{var}(\mathbb{E}[\boldsymbol{x}|y])$, where $\widehat{var}(\mathbb{E}[\boldsymbol{x}|y])$ is an estimate of the conditional covariance matrix $var(\mathbb{E}[\boldsymbol{x}|y])$ using SIR [Li, 1991], must be a linear combination of the column vectors of $\boldsymbol{X}$. Thus, we can construct an artificial response vector $\widetilde{\boldsymbol{y}} \in \mathbb{R}^n$ such that $\widehat{\boldsymbol{\eta}} = \frac{1}{n}\boldsymbol{X}\widetilde{\boldsymbol{y}}$, and estimate $\boldsymbol{\beta}$ by solving another penalized least square problem: $\frac{1}{2n}\|\widetilde{\boldsymbol{y}} - \boldsymbol{X}^\tau\boldsymbol{\beta}\|_2^2 + \mu\|\boldsymbol{\beta}\|_1$ (see Algorithm 3). We call this algorithm "Lasso-SIR", which is computationally very efficient. In Section 3, we further show that the convergence rate of the estimator resulting from Lasso-SIR is $\frac{s\log(p)}{n\lambda}$, which is optimal if $s = O(p^{1-\delta})$ for some positive constant δ. Note that Lasso-SIR can be easily extended to other regularization and SDR methods, such as SCAD (Fan and Li [2001]), Group Lasso (Yuan and Lin [2006]), sparse Group Lasso (Simon et al. [2013]), SAVE (Cook [2000]), etc.

Connection to Other work. Estimating the central space is widely considered as a generalized eigenvector problem in the literature [Li, 1991, Li and Nachtsheim, 2006, Li, 2007, Chen and Li, 1998]. Lin et al. [2016] explicitly described the similarities and differences between SIR and PCA (as first studied by Jung and Marron [2009]) under the "high dimension, low sample size (HDLSS)" scenario. However, after comparing their results with those for Lasso regression, Lin et al. [2016] advocated that a more appropriate prototype of SIR (at least for the single index model) should be the linear regression. In the past three decades, tremendous efforts have been put into the study of linear regression models $y = \boldsymbol{x}^\tau\boldsymbol{\beta} + \epsilon$ for HDLSS data. By imposing the L_1 penalty on the regression coefficients, the Lasso approach [Tibshirani, 1996] is rate optimal in estimating $\boldsymbol{\beta}$ in the sense that it achieves the maximum convergence rate [Raskutti et al., 2011]. Because of apparent limitations of linear models, there are many attempts to build flexible and computationally friendly semi-parametric models, such as the projection pursuit regression [Friedman and Stuetzle, 1981, Chen, 1991], sliced inverse regression [Li, 1991], and MAVE [Xia et al., 2002]. However, none of these methods work under the HDLSS setting. Existing theoretical results for HDLSS data mainly focus on linear regressions [Raskutti et al., 2011] and submatrix detections [Butucea et al., 2013], and are not applicable to index models. In this paper, we provide a new framework for the theoretical investigation of regularized SDR methods for HDLSS data.

The rest of the paper is organized as follows. After briefly reviewing SIR, we present the Lasso-SIR algorithm in Section 2. The consistency of the Lasso-SIR estimate and its connection to the Lasso regression are presented in Section 3. Numerical simulations and real data applications are reported in Sections 4 and 5. Some potential extensions are briefly discussed in Section 6. To improve the readability, we defer all the proofs and brief reviews of some existing results to the appendix.

2. Sparse SIR for High Dimensional Data

2.1. Notation

We adopt the following notations throughout this paper. For a matrix $\boldsymbol{V}$, we call the space generated by its column vectors the column space and denote it by $col(\boldsymbol{V})$. The i-th row and j-th column of the matrix are denoted by $\boldsymbol{V}_{i,*}$ and $\boldsymbol{V}_{*,j}$, respectively. For (column) vectors $\boldsymbol{x}$ and $\boldsymbol{\beta} \in \mathbb{R}^p$, we denote their inner product $\langle \boldsymbol{x}, \boldsymbol{\beta} \rangle$ by $\boldsymbol{x}(\boldsymbol{\beta})$, and the k-th entry of $\boldsymbol{x}$ by $\boldsymbol{x}(k)$. For two positive numbers a, b, we use $a \vee b$ and $a \wedge b$ to denote $\max\{a, b\}$ and $\min\{a, b\}$ respectively; We use C, C', C_1 and C_2 to denote generic absolute constants, though the actual value may vary from case to case. For two sequences $\{a_n\}$ and $\{b_n\}$, we denote $a_n \succ b_n$ and $a_n \prec b_n$ if there exist positive constants C and C' such that $a_n \geq C b_n$ and $a_n \leq C' b_n$, respectively. We denote $a_n \asymp b_n$ if both $a_n \succ b_n$ and $a_n \prec b_n$ hold. The $(1, \infty)$ norm and (∞, ∞) norm of matrix A are defined by $\|A\|_{1,\infty} = \max_{1 \leq j \leq p} \sum_{i=1}^{p} |A_{i,j}|$ and $\max_{1 \leq i,j \leq n} \|A_{i,j}\|$ respectively. To simplify discussions, we assume that $\frac{s \log(p)}{n\lambda}$ is sufficiently small. We emphasize again that our covariate data X is a $p \times n$ instead of the traditional $n \times p$ matrix.

2.2. A brief review of Sliced Inverse Regression (SIR)

In the multiple index model (1), the matrix $\boldsymbol{B}$ formed by the vectors $\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_d$ is not identifiable. However, $col(\boldsymbol{B})$, the space spanned by the columns of $\boldsymbol{B}$ is uniquely defined. Given n *i.i.d.* samples $(y_i, \boldsymbol{x}_i)$, $i = 1, \cdots, n$, as proposed in Li [1991], SIR first divides the data into H equal-sized slices according to the order statistics $y_{(i)}$, $i = 1, \ldots, n$. To ease notations and arguments, we assume that $n = cH$ and $\mathbb{E}[\boldsymbol{x}] = 0$, and re-express the data as $y_{h,j}$ and $\boldsymbol{x}_{h,j}$, where h refers to the slice number and j refers to the order number of a sample in the h-th slice, i.e., $y_{h,j} = y_{(c(h-1)+j)}$, $\boldsymbol{x}_{h,j} = \boldsymbol{x}_{(c(h-1)+j)}$. Here $\boldsymbol{x}_{(k)}$ is the concomitant of $y_{(k)}$. Let the sample mean in the h-th slice be denoted by $\overline{\boldsymbol{x}}_{h,\cdot}$, then $\boldsymbol{\Lambda} \triangleq var(\mathbb{E}[\boldsymbol{x}|y])$ can be estimated by:

$$\widehat{\boldsymbol{\Lambda}}_H = \frac{1}{H} \sum_{h=1}^{H} \bar{\boldsymbol{x}}_{h,\cdot} \bar{\boldsymbol{x}}_{h,\cdot}^{\tau} = \frac{1}{H} \boldsymbol{X}_H^{\tau} \boldsymbol{X}_H \tag{2}$$

where $\boldsymbol{X}_H$ is a $p \times H$ matrix formed by the H sample means, i.e., $\boldsymbol{X}_H = (\boldsymbol{x}_{1,\cdot}, \ldots, \boldsymbol{x}_{H,\cdot})$. Thus, $col(\boldsymbol{\Lambda})$ is estimated by $col(\widehat{\boldsymbol{V}}_H)$, where $\widehat{\boldsymbol{V}}_H$ is the matrix formed by the top d eigenvectors of $\widehat{\boldsymbol{\Lambda}}_H$. The $col(\widehat{\boldsymbol{V}}_H)$ was shown to be a consistent estimator of $col(\boldsymbol{\Lambda})$ under a few technical conditions when p is fixed [Duan and Li, 1991, Hsing and Carroll, 1992, Zhu et al., 2006, Li, 1991, Lin et al., 2015], which are summarized in the online supplementary file. Recently, Lin et al. [2015, 2016] showed that $col(\widehat{\boldsymbol{V}}_H)$ is consistent for $col(\boldsymbol{\Lambda})$ if and only if $\rho_n = \frac{p}{n} \to 0$ as $n \to \infty$, and that the number of slices H can be chosen as a fixed integer independent of n and p when the dimension d of the central space is bounded. When $\boldsymbol{x}$'s distribution is elliptically symmetric, Li [1991] showed that

$$\boldsymbol{\Sigma} col(\boldsymbol{B}) = col(\boldsymbol{\Lambda}), \tag{3}$$

and thus our goal is to recover $col(\boldsymbol{B})$ by solving the above equation, which can be easily and consistently achieved by $\widehat{col(\boldsymbol{B})} = \widehat{\Sigma}^{-1} col(\widehat{\boldsymbol{V}}_H)$ when $\rho_n \to 0$ [Lin et al., 2015], where

$\widehat{\Sigma} = \frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\tau$, the sample covariance matrix of $\boldsymbol{x}$. However, this simple approach breaks down when $\rho_n \not\to 0$, especially when $p \gg n$. Although stepwise methods [Zhong et al., 2012, Jiang and Liu, 2014] can work under HDLSS settings, the sparse SDR algorithms proposed in Li [2007] and Li and Nachtsheim [2006] appeared to be ineffective. Below we describe two intuitive non-stepwise methods for HDLSS scenarios, which will be used as benchmarks in our simulation studies to measure the performance of newly proposed SDR algorithms.

Diagonal Thresholding-SIR. When $p \gg n$, the Diagonal Thresholding (DT) screening method [Lin et al., 2015] proceeds by marginally screening all the variables via the diagonal elements of $\widehat{\boldsymbol{\Lambda}}_H$ and then applying SIR to those retained variables to obtain an estimate of $col(\boldsymbol{B})$. The procedure is shown to be consistent if the number of nonzero entries in each row of $\boldsymbol{\Sigma}$ is bounded even when $p \gg n$.

Algorithm 1 DT-SIR

1: Use the magnitudes of the diagonal elements of $\widehat{\boldsymbol{\Lambda}}_H$ to select the set of important predictors $\mathcal{I}$, with $|\mathcal{I}| = o(n)$

2: Apply SIR to the data $(\boldsymbol{y}, \boldsymbol{x}_\mathcal{I})$ to estimate a subspace $\widehat{\mathcal{S}}_\mathcal{I}$.

3: Extend $\widehat{\mathcal{S}}_\mathcal{I}$ to a subspace in $\mathbb{R}^p$ by filling in $0's$ for unimportant predictors.

Matrix Lasso. We can bypass the estimation and inversion of Σ by solving an L_1 penalization problem. Since (3) holds at the population level, we speculate that a reasonable estimate of $col(\boldsymbol{B})$ can be obtained by solving a sample-version of the equation with an appropriate regularization term to cope with the high dimensionality. Let $\widehat{\boldsymbol{\eta}}_1, \cdots, \widehat{\boldsymbol{\eta}}_d$ be the eigenvectors associated with the largest d eigenvalues of $\widehat{\boldsymbol{\Lambda}}_H$. Replacing $\boldsymbol{\Sigma}$ by its sample version $\frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\tau$ and imposing an L_1 penalty, we obtain a penalized sample version of (3):

$$\|\frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\tau\boldsymbol{\beta} - \widehat{\boldsymbol{\eta}}_i\|_2^2 + \mu_i\|\boldsymbol{\beta}\|_1 \tag{4}$$

for an appropriate choice of parameter μ_i.

Algorithm 2 Matrix Lasso

1: Let $\widehat{\boldsymbol{\eta}}_1, ..., \widehat{\boldsymbol{\eta}}_d$ be the eienvectors associated with the largest d eigenvalues of $\widehat{\boldsymbol{\Lambda}}_H$;

2: For $1 \le i \le d$, let $\widehat{\boldsymbol{\beta}}_i$ be the minimizer of equation (4) for an appropriate choice of μ_i;

3: Estimate the central space $col(\boldsymbol{B})$ by $col(\widehat{\boldsymbol{\beta}}_1, ..., \widehat{\boldsymbol{\beta}}_d)$.

This simple procedure can be easily implemented to produce sparse estimates of $\boldsymbol{\beta}_i$'s. Empirically it works reasonably well, so we set it as another benchmark to compare with. Since we later observed that its numerical performance was consistently worse than that of our main algorithm, Lasso-SIR, we did not further investigate its theoretical properties.

2.3. Lasso-SIR

Let us first consider the single index model

$$y = f(\boldsymbol{x}^\tau \boldsymbol{\beta}_0, \boldsymbol{\epsilon}). \tag{5}$$

Without loss of generality, we assume that $(\boldsymbol{x}_i, y_i),\ i = 1, \ldots, n$, are sorted $i.i.d.$ samples such that $y_1 \leq y_2 \leq \cdots \leq y_n$. Construct an $n \times H$ matrix $\boldsymbol{M} = \boldsymbol{I}_H \otimes \mathbf{1}_c$, where $\mathbf{1}_c$ is the $c \times 1$ vector with all entries being 1. Then, according to the definition of $\boldsymbol{X}_H$, we can write $\boldsymbol{X}_H = \boldsymbol{X}\boldsymbol{M}/c$. Let $\widehat{\lambda}$ be the largest eigenvalue of $\widehat{\boldsymbol{\Lambda}}_H = \frac{1}{H}\boldsymbol{X}_H\boldsymbol{X}_H^\tau$ and let $\widehat{\boldsymbol{\eta}}$ be the corresponding eigenvector of length 1. That is,

$$\hat{\lambda}\widehat{\boldsymbol{\eta}} = \frac{1}{H}\boldsymbol{X}_H^\tau \boldsymbol{X}_H \widehat{\boldsymbol{\eta}} = \frac{1}{nc}\boldsymbol{X}^\tau \boldsymbol{M}\boldsymbol{M}^\tau \boldsymbol{X}\widehat{\boldsymbol{\eta}}.$$

Thus, by defining

$$\widetilde{\boldsymbol{y}} = \frac{1}{c\widehat{\lambda}}\boldsymbol{M}\boldsymbol{M}^\tau \boldsymbol{X}\widehat{\boldsymbol{\eta}} \tag{6}$$

we have $\hat{\boldsymbol{\eta}} = \frac{1}{n}\boldsymbol{X}\widetilde{\boldsymbol{y}}$. Note that a key in estimating the central space $col(\boldsymbol{\beta})$ of SIR is the equation $\boldsymbol{\eta} \propto \boldsymbol{\Sigma}\boldsymbol{\beta}$. If approximating $\boldsymbol{\eta}$ and $\boldsymbol{\Sigma}$ by $\widehat{\boldsymbol{\eta}}$ and $\frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\tau$ respectively, this equation can be written as $\frac{1}{n}\boldsymbol{X}\widetilde{\boldsymbol{y}} \propto \frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\tau\boldsymbol{\beta}$. To recover a sparse vector $\widehat{\boldsymbol{\beta}} \propto \boldsymbol{\beta}$, one can consider the following optimization problem

$$\min ||\boldsymbol{\beta}||_1, \quad \text{subject to} \quad ||\boldsymbol{X}(\widetilde{\boldsymbol{y}} - \boldsymbol{X}^\tau\boldsymbol{\beta})||_\infty \leq \mu,$$

which is known as the Dantzig selector [Candes and Tao, 2007]. A related formulation is the Lasso regression, where $\boldsymbol{\beta}$ is estimated by the minimizer of

$$\mathcal{L}_{\boldsymbol{\beta}} = \frac{1}{2n}\|\widetilde{\boldsymbol{y}} - \boldsymbol{X}^\tau\boldsymbol{\beta}\|_2^2 + \mu\|\boldsymbol{\beta}\|_1. \tag{7}$$

As shown by Bickel et al. [2009], the Dantzig Selector is asymptotically equivalent to the Lasso for linear regressions. We thus propose and study the Lasso-SIR algorithm in this paper and leave the discussion of the Dantzig selector in a future work.

Algorithm 3 Lasso-SIR (for single index models)

1: Let $\widehat{\lambda}$ and $\widehat{\boldsymbol{\eta}}$ be the first eigenvalue and eigenvector of $\widehat{\boldsymbol{\Lambda}}_H$, respectively;
2: Let $\widetilde{\boldsymbol{y}} = \frac{1}{c\widehat{\lambda}}\boldsymbol{M}\boldsymbol{M}^\tau\boldsymbol{X}\widehat{\boldsymbol{\eta}}$ and solve the Lasso optimization problem

$$\widehat{\boldsymbol{\beta}}(\mu) = \arg\min \mathcal{L}_{\boldsymbol{\beta}}, \text{ where } \mathcal{L}_{\boldsymbol{\beta}} = \frac{1}{2n}\|\widetilde{\boldsymbol{y}} - \boldsymbol{X}^\tau\boldsymbol{\beta}\|_2^2 + \mu\|\boldsymbol{\beta}\|_1;$$

3: Estimate $P_{\boldsymbol{\beta}}$ by $P_{\widehat{\boldsymbol{\beta}}(\mu)}$.

Clearly, we do not need to estimate the inverse of $\boldsymbol{\Sigma}$ in Lasso-SIR. Moreover, since the optimization problem (7) is well studied for linear regression models [Tibshirani, 1996,

Efron et al., 2004, Friedman et al., 2010], we may formally "transplant" their results to the index models. Specifically, we use the R package *glmnet* to solve the optimization problem where the tuning parameter μ is chosen based on cross-validation.

Last but not least, Lasso-SIR can be easily generalized to the multiple index model (1). Let $\widehat{\lambda}_i, 1 \le i \le d$, be the d-top eigenvalues of $\widehat{\mathbf{\Lambda}}_H$ and $\widehat{\boldsymbol{\eta}} = (\widehat{\boldsymbol{\eta}}_1, \cdots, \widehat{\boldsymbol{\eta}}_d)$ be the corresponding eigenvectors. Similar to the definition of the *"pseudo response variable"* for the single index model, we define a multivariate pseudo response $\widetilde{\boldsymbol{Y}}$ as

$$\widetilde{\boldsymbol{Y}} = \frac{1}{c}\boldsymbol{M}\boldsymbol{M}^{\tau}\boldsymbol{X}\widehat{\boldsymbol{\eta}}\ diag(\frac{1}{\widehat{\lambda}_1}, \cdots, \frac{1}{\widehat{\lambda}_d}). \tag{8}$$

We then apply the Lasso on each column of the pseudo response matrix to produce the corresponding $\widehat{\boldsymbol{\beta}}_i's$.

Algorithm 4 Lasso-SIR (for multiple index model)

1: Let $\widehat{\lambda}_i$ and $\widehat{\boldsymbol{\eta}}_i$, $i = 1, \cdots, d$ be the top d eigenvalues and eigenvectors of $\widehat{\mathbf{\Lambda}}_H$ respectively.

2: Let $\widetilde{\boldsymbol{Y}} = \frac{1}{c}\boldsymbol{M}\boldsymbol{M}^{\tau}\boldsymbol{X}\widehat{\boldsymbol{\eta}}\ diag(\frac{1}{\widehat{\lambda}_1}, \cdots, \frac{1}{\widehat{\lambda}_d})$ and for each $1 \le i \le d$ solve the Lasso optimization problem

$$\widehat{\boldsymbol{\beta}}_i = \arg\min \mathcal{L}_{\boldsymbol{\beta},i} \text{ where } \mathcal{L}_{\boldsymbol{\beta},i} = \frac{1}{2n}\|\widetilde{\boldsymbol{Y}}_{*,i} - \boldsymbol{X}^{\tau}\boldsymbol{\beta}\|_2^2 + \mu\|\boldsymbol{\beta}\|_1.$$

3: Let $\widehat{\boldsymbol{B}}$ be the matrix formed by $\widehat{\boldsymbol{\beta}}_1, \cdots, \widehat{\boldsymbol{\beta}}_d$. The estimate of $P_{\boldsymbol{B}}$ is given by $P_{\widehat{\boldsymbol{B}}}$.

The number of directions d plays an important role when implementing Algorithm 4. A common practice is to locate the maximum gap among the ordered eigenvalues of the matrix $\widehat{\mathbf{\Lambda}}_H$, which does not work well under HDLSS settings. In Section 3, we show that there exists a gap among the adjusted eigenvalues $\hat{\lambda}_i^a = \hat{\lambda}_i||\hat{\beta}_i||_2$ where $\hat{\beta}_i$ is the i-th output of Algorithm 4. Motivated by this, we estimate d according to the following algorithm:

Algorithm 5 Estimation of the number of directions d

1: Apply Algorithm 4 by setting $d = H$;
2: For each i, calculate $\hat{\lambda}_i^a = \hat{\lambda}_i||\hat{\beta}_i||_2$;
3: Apply the k-means method on $\hat{\lambda}_i^a$ with k being 2 and the total number of points in the cluster with larger $\hat{\lambda}^a$ is the estimated value of d.

3. Consistency of Lasso-SIR

For simplicity, we assume that $\boldsymbol{x} \sim N(0, \mathbf{\Sigma})$. The normality assumption can be relaxed to elliptically symmetric distributions with sub-Gaussian tail; however, this will make technical arguments unnecessarily tedious and is not the main focus of this paper. From now on, we assume that d, the dimension of the central space, is bounded; thus we can assume that H,

the number of slices, is also finite [Lin et al., 2016, 2015]. In order to prove the consistency, we need the following technical conditions:

A1) There exist two positive constants $C_{\min} < C_{\max}$, such that $C_{\min} < \lambda_{\min}(\mathbf{\Sigma}) \leq \lambda_{\max}(\mathbf{\Sigma}) < C_{\max}$;

A2) There exists a constant $\kappa \geq 1$, such that $0 < \lambda = \lambda_d(var(\mathbb{E}[\boldsymbol{x}|y]) \leq ... \leq \lambda_1(var(\mathbb{E}[\boldsymbol{x}|y]) \leq \kappa\lambda \leq \lambda_{max}(\mathbf{\Sigma})$;

A3) The central curve $\boldsymbol{m}(y) = \mathbb{E}[\boldsymbol{x}|y]$ satisfies the sliced stability condition.

Condition A1 is quite mild and standard. Condition A2 is merely a refinement of the coverage condition, $rank(var(\mathbb{E}[\boldsymbol{x}|y])) = d$, a common assumption in the SIR literature. A3 is a condition on the central curve, or equivalently, a regularity condition on the link function $f(\cdot)$ and the noise $\boldsymbol{\epsilon}$. It guarantees that $\gamma^\tau \hat{\mathbf{\Lambda}}_H \gamma$ converges to $var(\gamma^\tau \mathbb{E}[\boldsymbol{x}|y])$ for any $\gamma \in col(var(\mathbb{E}[\boldsymbol{x}|y]))$. These conditions play indispensable roles in establishing the phase transition phenomenon of SIR [Lin et al., 2015] and the minimax rate of SIR over a large class of models [Lin et al., 2016]. As discussed in Lin et al. [2016] and Neykov et al. [2016], Condition A3 can be derived from a slightly stronger version of the commonly used condition introduced by Hsing and Carroll [1992]. To improve the readability and avoid duplicating existing results, we leave further discussion of Condition A3 in the online supplementary file. For single index model, there is a more intuitive explanation of the condition A2. Since $rank(var(\mathbb{E}[\boldsymbol{x}|y])) = 1$, condition A2 is simplified to $0 < \lambda = \lambda_1 \leq \lambda_{max}(\mathbf{\Sigma})$ which is a direct corollary of the total variance decomposition identity (i.e., $var(\boldsymbol{x}) = var(\mathbb{E}[\boldsymbol{x}|y]) + \mathbb{E}[var(\boldsymbol{x}|y)]$). We may treat λ as a generalized SNR and A2 simply requires that the generalized SNR is non-zero.

Remark 1 (Generalized SNR and eigenvalue bound). Recall that the the signal-to-noise ratio (SNR) for the linear model $y = \boldsymbol{\beta}^\tau \boldsymbol{x} + \epsilon$, where $\boldsymbol{x} \sim N(0, \mathbf{\Sigma})$ and $\epsilon \sim N(0, 1)$, is defined as

$$SNR = \frac{E[(\boldsymbol{\beta}^\tau \boldsymbol{x})^2]}{\mathbb{E}[y^2]} = \frac{\|\boldsymbol{\beta}\|_2^2 \boldsymbol{\beta}_0^\tau \mathbf{\Sigma} \boldsymbol{\beta}_0}{1 + \|\boldsymbol{\beta}\|_2^2 \boldsymbol{\beta}_0^\tau \mathbf{\Sigma} \boldsymbol{\beta}_0}.$$

where $\boldsymbol{\beta}_0 = \boldsymbol{\beta}/\|\boldsymbol{\beta}\|_2$. A simple calculation shows that

$$var(\mathbb{E}[\boldsymbol{x}|y]) = \frac{\mathbf{\Sigma}\boldsymbol{\beta}\boldsymbol{\beta}^\tau\mathbf{\Sigma}}{\boldsymbol{\beta}_0^\tau \mathbf{\Sigma} \boldsymbol{\beta}_0 \|\boldsymbol{\beta}\|_2^2 + 1}, \quad \text{and} \quad \lambda(var(\mathbb{E}[\boldsymbol{x}|y])) = \frac{\boldsymbol{\beta}_0^\tau \mathbf{\Sigma}\mathbf{\Sigma}\boldsymbol{\beta}_0 \|\boldsymbol{\beta}\|_2^2}{\boldsymbol{\beta}_0^\tau \mathbf{\Sigma} \boldsymbol{\beta}_0 \|\boldsymbol{\beta}\|_2^2 + 1},$$

where $\lambda(var(\mathbb{E}[\boldsymbol{x}|y]))$ is the unique non-zero eigenvalue of $var(\mathbb{E}[\boldsymbol{x}|y])$. This leads to the identity

$$\lambda(var(\mathbb{E}[\boldsymbol{x}|y])) = \frac{\boldsymbol{\beta}_0^\tau \mathbf{\Sigma}\mathbf{\Sigma}\boldsymbol{\beta}_0}{\boldsymbol{\beta}_0^\tau \mathbf{\Sigma} \boldsymbol{\beta}_0} SNR.$$

Thus, we call λ, the smallest non-zero eigenvalue of $var(\mathbb{E}[\boldsymbol{x}|y])$, the generalized SNR for multiple index models.

THEOREM 1 (CONSISTENCY OF LASSO-SIR FOR SINGLE INDEX MODELS). *Let $\widehat{\boldsymbol{\beta}}$ be the output of Algorithm 3. Assume that $n\lambda = p^\alpha$ for some $\alpha > 1/2$ and that conditions A1-A3 hold for the single index model, $y = f(\boldsymbol{\beta}_0^\tau \boldsymbol{x}, \epsilon)$, where $\boldsymbol{\beta}_0$ is a unit vector, then*

$$\|P_{\widehat{\boldsymbol{\beta}}} - P_{\boldsymbol{\beta}_0}\|_F \leq C_1 \sqrt{\frac{s\log(p)}{n\lambda}}$$

holds with probability at least $1 - C_2\exp(-C_3\log(p))$ for some constants C_2 and C_3.

The technical assumption $n\lambda = p^\alpha$ for some $\alpha > 1/2$ could be replaced by some sparsity conditions on the covariance matrix $\boldsymbol{\Sigma}$. Since our focus here is to introduce the Lasso-SIR algorithm, we leave this extension in our future study. Next, we state the theoretical result regarding the multiple index model (1).

THEOREM 2 (CONSISTENCY OF LASSO-SIR). *Let $\widehat{B}$ be the output of Algorithm 4 where d, the dimension of central subspace, is known. Assume that that $n\lambda = p^\alpha$ for some $\alpha > 1/2$ (where λ is the smallest nonzero eigenvalue of $var(\mathbb{E}[\boldsymbol{x}|y])$), and that conditions A1-A3 hold for the multiple index model* (1)*, then*

$$\|P_{\widehat{\boldsymbol{B}}} - P_{\boldsymbol{B}}\|_F \leq C_1 \sqrt{\frac{s\log(p)}{n\lambda}}.$$

holds with probability at least $1 - C_2\exp(-C_3\log(p))$ for some constants C_2 and C_3.

Lin et al. [2016] have speculated that the lower bound of the risk $\mathbb{E}\|P_{\widehat{\boldsymbol{B}}} - P_{\boldsymbol{B}}\|_F^2$ is $\frac{s\log(p/s)}{n\lambda}$ when the dimension of the central space is bounded. This implies that if $s = O(p^{1-\delta})$ for some positive constant δ, the Lasso-SIR algorithm achieves the optimal rate, i.e., we have the following corollary.

COROLLARY 1. *Assume that conditions A1-A3 hold for the multiple index model* (1) *and d, the dimension of central subspace, is bounded. If $n\lambda = p^\alpha$ for some $\alpha > 1/2$ and $s = O(p^{1-\delta})$, the Lasso-SIR estimate $P_{\widehat{\boldsymbol{B}}}$ achieves the minimax rate.*

REMARK 2 (COMPARISON WITH LASSO REGRESSION). Consider the linear regression $y = \boldsymbol{\beta}^\tau \boldsymbol{x} + \epsilon$, where $\boldsymbol{x} \sim N(0, \boldsymbol{\Sigma})$ and $\epsilon \sim N(0, 1)$. It is shown in Raskutti et al. [2011] that the lower bound of the minimax rate of the l_2 distance between any estimator and the true $\boldsymbol{\beta}$ is $\frac{s\log(p/s)}{n}$ and the convergence rate of Lasso estimator $\widehat{\boldsymbol{\beta}}_{Lasso}$ is $\frac{s\log(p)}{n}$. Namely, the Lasso estimator is rate optimal for linear regression when $s = O(p^{1-\delta})$ for some positive constant δ. A simple calculation shows that $\lambda(var(\mathbb{E}[\boldsymbol{x}|y])) \sim \|\boldsymbol{\beta}\|_2^2$, if $\|\boldsymbol{\beta}\|$ is bounded away from ∞. Consequently,

$$\|P_{\widehat{\boldsymbol{\beta}}_{Lasso}} - P_{\boldsymbol{\beta}}\|_F \leq 4\frac{\|\widehat{\boldsymbol{\beta}}_{Lasso} - \boldsymbol{\beta}\|_2}{\|\boldsymbol{\beta}\|_2} \leq C\sqrt{\frac{s\log(p)}{n\lambda(var(\mathbb{E}[\boldsymbol{x}|y]))}} \tag{9}$$

holds with high probability. In other words, the projection matrix $P_{\widehat{\boldsymbol{\beta}}_{Lasso}}$ based on Lasso is rate optimal. From this view point, the Lasso-SIR has extended the Lasso to the non-linear

multiple index models. This justifies a suggestion in Chen and Li [1998], stating that SIR should be viewed as an alternative or generalization of the multiple linear regression. The connection also justifies a speculation in Lin et al. [2016] that a more appropriate prototype of the high dimensional SIR problem should be the sparse linear regression rather than the sparse PCA and the generalized eigenvector problem.

Determining the dimension d of the central space is a challenge problem for SDR, especially for HDLSS cases. If we want to discern signals (i.e., the true directions) from noises (i.e., the other directions) simply via the eigenvalues $\widehat{\lambda}_i$ of $\widehat{\mathbf{\Lambda}}_H$, $i = 1, ..., H$, we face the problem that all these $\widehat{\lambda}_i$'s are of order p/n, but the gap between the signals and noises is of order λ ($\leq C_{\max}$). With the Lasso-SIR, we can bypass this difficulty by using the adjusted eigenvalues $\hat{\lambda}_i^a = \widehat{\lambda}_i \|\widehat{\boldsymbol{\beta}}_i\|_2$, $i = 1, ..., H$. To this end, we have the following theorem.

THEOREM 3. *Let $\widehat{\boldsymbol{\beta}}_i$ be the output of Algorithm 4 for $i = 1, ..., H$. Assume that $n\lambda = p^\alpha$ for some $\alpha > 1/2$, $s\log(p) = o(n\lambda)$, and $H > d$, then, for some constants C_1, C_2 C_3 and C_4,*

$$\hat{\lambda}_i^a \geq C_1\sqrt{\lambda} - C_2\sqrt{\frac{s\log(p)}{n}}, \text{ for } 1 \leq i \leq d, \text{ and}$$

$$\hat{\lambda}_i^a \leq C_3\frac{\sqrt{p\log(p)}}{n\lambda}\sqrt{\lambda} + C_4\sqrt{\frac{s\log(p)}{n}}, \text{ for } d+1 \leq i \leq H,$$

hold with probability at least $1 - C_5\exp(-C_6\log(p))$ for some constants C_5 and C_6.

In the sparse linear regression literature, the region $s^2 = o(p)$ is often referred to as the "highly sparse" region [Ingster et al., 2010]. Theorem 3 shows that there is a clear gap between the signal (i.e., the first d adjusted eigenvalues) and noise (i.e., the other adjusted eigenvalues) in the highly sparse region if $p^{1/2} = o(n\lambda)$.

4. Simulation Stuides

4.1. Single Index Models

Let $\boldsymbol{\beta}$ be the vector of coefficients and let $\mathcal{S}$ be the active set; namely, $\beta_i = 0, \forall i \in \mathcal{S}^c$. Furthermore, for each $i \in \mathcal{S}$, we simulated independently $\beta_i \sim N(0, 1)$. Let $\boldsymbol{x}$ be the design matrix with each row following $N(0, \mathbf{\Sigma})$. We consider two types of covariance matrices: (i) $\mathbf{\Sigma} = (\sigma_{ij})$ where $\sigma_{ii} = 1$ and $\sigma_{ij} = \rho^{|i-j|}$; and (ii) $\sigma_{ii} = 1$, $\sigma_{i,j} = \rho$ when $i, j \in \mathcal{S}$ or $i, j \in \mathcal{S}^c$, and $\sigma_{i,j} = 0.1$ when $i \in \mathcal{S}, j \in \mathcal{S}^c$ or vice versa. The first one represents a covariance matrix which is essentially sparse and we choose ρ among 0, 0.3, 0.5, and 0.8. The second one represents a dense covariance matrix with ρ chosen as 0.2. In all the simulations, we set $n = 1,000$ and let p vary among 100, 1,000, 2,000, and 4,000. For all the settings, the random error $\boldsymbol{\epsilon}$ follows $N(0, \boldsymbol{I}_n)$. For single index models, we consider the following model settings:

I. $\boldsymbol{y} = \boldsymbol{x\beta} + \boldsymbol{\epsilon}$ where $\mathcal{S} = \{1, 2, \cdots, 10\}$;

II. $\boldsymbol{y} = (\boldsymbol{x\beta})^3/2 + \epsilon$ where $\mathcal{S} = \{1, 2, \cdots, 20\}$;

III. $\boldsymbol{y} = \sin(\boldsymbol{x\beta}) * exp(\boldsymbol{x\beta}) + \epsilon$ where $\mathcal{S} = \{1, 2, \cdots, 10\}$;

Table 1. Estimation error for the first type covariance matrix with $\rho = 0.5$.

	p	Lasso-SIR	DT-SIR	Lasso	M-Lasso	Lasso-SIR(Known d)	$\hat{d}$
I	100	0.12	0.47	0.11	0.19	0.12	1
	1000	0.18	0.65	0.15	0.26	0.18	1
	2000	0.20	0.74	0.16	0.30	0.20	1
	4000	0.23	0.90	0.18	0.39	0.23	1
II	100	0.07	0.60	0.23	0.27	0.07	1
	1000	0.12	0.78	0.31	0.17	0.12	1
	2000	0.15	0.86	0.34	0.20	0.15	1
	4000	0.20	0.99	0.37	0.28	0.19	1
III	100	0.21	0.55	1.25	0.26	0.21	1
	1000	0.28	0.74	1.32	0.51	0.27	1
	2000	0.35	0.87	1.34	0.66	0.31	1.1
	4000	0.46	1.00	1.33	0.83	0.39	1.1
IV	100	0.46	0.92	0.78	0.58	0.45	1
	1000	0.62	1.07	0.87	0.78	0.58	1.1
	2000	0.71	1.22	0.89	0.94	0.59	1.3
	4000	0.71	1.30	0.91	1.00	0.63	1.2
V	100	0.12	0.37	0.42	0.15	0.12	1
	1000	0.20	0.55	0.55	0.41	0.20	1
	2000	0.38	0.80	0.60	0.67	0.29	1.2
	4000	0.78	1.22	0.77	1.06	0.48	1.5

IV. $\boldsymbol{y} = \exp(\boldsymbol{x}\boldsymbol{\beta}/10) + \epsilon$ where $\mathcal{S} = \{1, 2, \cdots, 50\}$;

V. $\boldsymbol{y} = \exp(\boldsymbol{x}\boldsymbol{\beta} + \epsilon)$ where $\mathcal{S} = \{1, 2, \cdots, 7\}$.

The goal is to estimate $col(\boldsymbol{\beta})$, the space spanned by $\boldsymbol{\beta}$. As in Lin et al. [2015], the estimation error is defined as $\mathcal{D}(\widehat{col(\boldsymbol{\beta})}, col(\boldsymbol{\beta}))$, where $\mathcal{D}(M, N)$, the distance between two subspaces $M, N \subset \mathcal{R}^p$, is defined as the Frobenius norm of $P_M - P_N$ where P_M and P_N are the projection matrices associated with these two spaces. The methods we compared with are DT-SIR, matrix Lasso (M-Lasso), and Lasso. The number of slices H is chosen as 20 in all simulation studies. The number of directions d is chosen according to Algorithm 5. Note that both benchmarks (i.e., DT-SIR and M-Lasso) require the knowledge of d as well. To be fair, we use the d estimated based on Algorithm 5 for both benchmarks. For comparison, we have also included the estimation error of Lasso-SIR assuming d is known. For each p, n, and ρ, we replicate the above steps 100 times to calculate the average estimation error for each setting. We tabulated the results for the first type of covariance matrix with $\rho = 0.5$ in Table 1 and put the results for other settings in Tables C1-C4 in the online supplementary file. The average of estimated directions $\hat{d}$ is reported in the last column of these tables.

The simulations show that Lasso-SIR outperformed both DT-SIR and M-Lasso under all settings. The performance of DT-SIR has become worse when the dependence is stronger and denser. The reason is that this method is based on the diagonal threshold and is only supposed to work well for the diagonal covariance matrix. Overall, Algorithm 5 provided a reasonable good estimate of d especially for moderate covariance matrix. When assuming d is known, the performances of both DT-SIR and M-Lasso are inferior to Lasso-SIR, and are

Table 2. Estimation error for the first type covariance matrix with $\rho = 0.5$.

	p	Lasso-SIR	DT-SIR	M-Lasso	Lasso-SIR(Known d)	$\hat{d}$
VI	100	0.26	0.57	0.31	0.26	2
	1000	0.33	0.74	0.62	0.33	2
	2000	0.36	0.92	0.73	0.38	2
	4000	0.44	1.12	0.87	0.42	2
VII	100	0.32	0.67	0.42	0.32	2
	1000	0.60	0.93	1.02	0.66	2.1
	2000	0.95	1.18	1.35	0.83	2.3
	4000	1.17	1.43	1.47	1.08	2.1
VIII	100	0.29	0.61	0.34	0.25	2
	1000	0.37	0.82	0.69	0.35	2
	2000	0.54	1.00	0.92	0.47	2.2
	4000	0.88	1.37	1.27	0.71	2.5
IX	100	0.43	0.74	0.48	0.43	2
	1000	0.47	0.91	0.91	0.48	2
	2000	0.58	1.11	1.12	0.50	2.1
	4000	0.57	1.25	1.23	0.56	2

thus not reported.

Under Setting I when the true model is linear, Lasso performed the best among all the methods, as expected. However, the difference between Lasso and Lasso-SIR is not significant, implying that Lasso-SIR does not sacrifice much efficiency without the knowledge of the underlying linearity. On the other hand, when the models are not linear (Case II-VI), Lasso-SIR worked much better than Lasso. It is seen that Lasso works better than Lasso-SIR for Setting V when ρ is large or the covariance matrix is dense. One explanation is that Lasso-SIR tends to overestimate d under these conditions while Lasso used the actual d. If assuming $d = 1$ in Lasso-SIR, its estimation error is smaller than that of Lasso.

4.2. Multiple Index Models

Let $\boldsymbol{\beta}$ be the $p \times 2$ matrix of coefficients and $\mathcal{S}$ be the active set. Let $\boldsymbol{x}$ be simulated similarly as in Section 4.1, and denote $\boldsymbol{z} = \boldsymbol{x}\boldsymbol{\beta}$. Consider the following settings:

VI. $y_i = |z_{i2}/4 + 2|^3 * sgn(z_{i1}) + \epsilon_i$ where $\mathcal{S} = \{1, 2, \cdots, 7\}$ and $\beta_{1:4,1} = 1, \beta_{5:7,2} = 1$, and $\beta_{i,j} = 0$ otherwise;

VII. $y_i = z_{i1} * exp(z_{i2}) + \epsilon_i$ where $\mathcal{S} = \{1, 2, \cdots, 12\}$ and $\beta_{1:7,1}, \beta_{8:12,2} \sim N(0, 1)$, and $\beta_{i,j} = 0$ otherwise;

VIII. $y_i = z_{i1} * exp(z_{i2} + \epsilon_i)$ where $\mathcal{S} = \{1, 2, \cdots, 12\}$ and $\beta_{1:7,1}, \beta_{8:12,2} \sim N(0, 1)$, and $\beta_{i,j} = 0$ otherwise;

IX. $y_i = z_{i1} * (2 + z_{i2}/3)^2 + \epsilon_i$ where $\mathcal{S} = \{1, 2, \cdots, 12\}$ and $\beta_{1:8,1} = 1, \beta_{9:12,2} = 1$ and $\beta_{i,j} = 0$ otherwise.

For the multiple index models, we compared both benchmarks (DT-SIR and M-Lasso) with Lasso-SIR. Lasso is not applicable for these cases and is thus not included. Similar to

Section 4.1, we tabulated the results for the first type covariance matrix with $\rho = 0.5$ in Table 2 and put the results for others in Tables C5-C8 in the online supplementary file.

For the identity covariance matrix ($\rho = 0$), there was little difference between performances of Lasso-SIR and DT-SIR. However, Lasso-SIR was substantially better than DT-SIR in other cases. Under all settings, Lasso-SIR worked much better than the matrix Lasso. For the dense covariance matrix $\mathbf{\Sigma}_2$, Algorithm 5 tended to underestimate d, which is worthy of further investigation.

4.3. Discrete Response

We consider the following simulation settings where for the response variable Y is discrete.

X. $\boldsymbol{y} = 1(\boldsymbol{x\beta} + \boldsymbol{\epsilon} > 0)$ where $\mathcal{S} = \{1, 2, \cdots, 10\}$;
XI. $\boldsymbol{y} = 1(exp(\boldsymbol{x\beta}) + \epsilon > 0)$ where $\mathcal{S} = \{1, 2, \cdots, 7\}$;
XII. $\boldsymbol{y} = 1((\boldsymbol{x\beta})^3/2 + \epsilon)$ where $\mathcal{S} = \{1, 2, \cdots, 20\}$;
XIII. Let $\boldsymbol{z} = \boldsymbol{x\beta}$ where $\mathcal{S} = \{1, 2, \cdots, 12\}$, $\boldsymbol{\beta}$ is a p by 2 matrix with $\beta_{1:7,1}, \beta_{8:12,2} \sim N(0, 1)$ and $\beta_{i,j} = 0$ otherwise. The response y_i is

$$y_i = \begin{cases} 1, & \text{if } z_{i1} + \epsilon_{i1} < 0, \\ 2, & \text{if } z_{i1} + \epsilon_{i1} > 0 \text{ and } z_{i2} + \epsilon_{i2} < 0, \\ 3, & \text{if } z_{i1} + \epsilon_{i1} > 0 \text{ and } z_{i2} + \epsilon_{i2} > 0, \end{cases}$$

where $\epsilon_{ij} \sim N(0, 1)$.

In settings X, XI, and XII, the response variable is dichotomous, and $\beta_i \sim N(0, 1)$ when $i \in \mathcal{S}$ and $\beta_i = 0$ otherwise. Thus the number of slices H can only be 2. For Setting XIII where the response variable is trichotomous, the number of slices H is chosen as 3. The number of direction d is chosen as $H - 1$ in all these simulations.

Similar to the previous two sections, we calculated the average estimation errors for Lasso-SIR (Algorithm 4), DT-SIR, M-Lasso, and generalized-Lasso based on 100 replications and reported the result in Table 3 for the first type covariance matrix with $\rho = 0.5$ and the results for other cases in Tables C9-C12 in online supplementary file. It is clearly seen that Lasso-SIR performed much better than DT-SIR and M-Lasso under all settings and the improvements were very significant. The generalized Lasso performed as good as Lasso-SIR for the dichotomous response; however, it performed substantially worse for Setting XIII.

5. Applications to Real Data

5.1. Arcene Data Set

We first apply the methods to a two-class classification problem, which aims to distinguish between cancer patients and normal subjects from using their mass-spectrometric measurements. The data were obtained by the National Cancer Institute (NCI) and the Eastern Virginia Medical School (EVMS) using the SELDI technique, including samples from 44 patients with ovarian and prostate cancers and 56 normal controls. The dataset was downloaded from the UCI machine learning repository (Lichman [2013]), where a detailed description can be found. It has also been used in the NIPS 2003 feature selection challenge (Guyon et al.

Table 3. Estimation error for the first type covariance matrix with $\rho = 0.5$.

	p	Lasso-SIR	DT-SIR	M-Lasso	Lasso
X	100	0.21	0.58	0.26	0.21
	1000	0.26	1.21	0.51	0.28
	2000	0.27	1.33	0.59	0.29
	4000	0.28	1.39	0.65	0.31
XI	100	0.36	0.85	0.57	0.34
	1000	0.44	1.32	1.07	0.45
	2000	0.46	1.38	1.16	0.48
	4000	0.52	1.41	1.20	0.52
XII	100	0.24	0.63	0.49	0.22
	1000	0.34	1.18	0.53	0.32
	2000	0.36	1.29	0.61	0.35
	4000	0.40	1.37	0.68	0.39
XIII	100	0.38	1.08	0.60	1.06
	1000	0.38	1.79	1.13	1.08
	2000	0.40	1.92	1.24	1.09
	4000	0.41	1.98	1.31	1.10

[2004]). For each subject, there are 10,000 features where 7,000 of them are real variables and 3,000 of them are random probes. There are 100 subjects in the validation set.

After standardizing $\boldsymbol{X}$, we estimated the number of directions d as 1 using Algorithm 5. We then applied Algorithm 3 and the sparse PCA to calculate the direction of $\boldsymbol{\beta}$ and the corresponding components, followed by a logistic regression model. We applied the fitted model to the validation set and calculated the probability of each subject being a cancer patient. We also fitted a Lasso logistic regression model to the training set and applied it to the validation set to calculate the corresponding probabilities.

In Figure 1, we plot the Receiver Operating Characteristic (ROC) curves for various methods. Lasso-SIR, represented by the red curve, was slightly better than Lasso (insignificant) and the sparse PCA, represented by the green and blue curves respectively. The areas under these three curves are 0.754, 0.742, and 0.671, respectively.

5.2. HapMap

In this section, we analyzed a data set with a continuous response. We consider the gene expression data from 45 Japanese and 45 Chinese from the international "HapMap" project (Thorisson et al. [2005], Thorgeirsson et al. [2010]). The total number of probes is 47,293. According to Thorgeirsson et al. [2010], the gene *CHRNA6* is the subject of many nicotine addiction studies. Similar to Fan et al. [2015], we treat the mRNA expression of *CHRNA6* as the response Y and expressions of other genes as the covariates. Consequently, the number of dimension p is 47,292, much greater than the number of subjects n=90.

We first applied Lasso-SIR to the data set with d being chosen as 1 according to Algorithm 5. The number of selected variables was 13. Based on the estimated coefficients $\boldsymbol{\beta}$ and $\boldsymbol{X}$, we calculated the first component and the scatter plot between the response Y and this component, showing a moderate linear relationship between them. We then fitted a linear

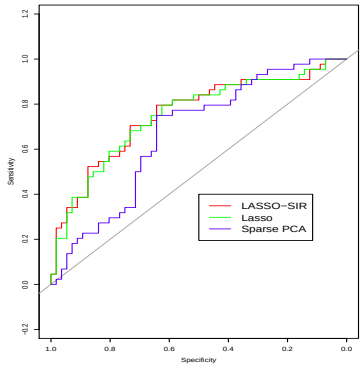

Fig. 1. ROC curve of various methods for Arcene Data set.

regression between them. The R-sq of this model is 0.5596 and the mean squared error of the fitted model 0.045.

We also applied Lasso to estimate the direction $\boldsymbol{\beta}$. The tuning parameter λ is chosen as 0.1215 such that the number of selected variables is also 13. When fitting a regression model between Y and the component based on the estimated $\boldsymbol{\beta}$, the R-sq is 0.5782 and the mean squared error is 0.044. There is no significant difference between these two approaches. This confirms the message that Lasso-SIR performs as good as Lasso when the linearity assumption is appropriate.

We have also calculated a direction and the corresponding components based on the sparse PCA (Zou et al. [2006]). We then fitted a regression model. The R-sq is only 0.1013 and the mean squared error is 0.093, significantly worse than the above two approaches.

5.3. Classify Wine Cultivars

We finally investigate the popular wine data set that people have used to compare various classification methods. This is a three-class classification problem. The data, available from the UCI machine learning repository (Lichman [2013]), consists of 178 wines grown in the same region in Italy under three different cultivars. For each wine, the chemical analysis was conducted and the quantities of 13 constituents are obtained, which are Alcohol, Malic acid, Ash, Alkalinity of ash, Magnesium, Total Phenols, Flavanoids, Nonflavanoid Phenols, Proanthocyanins, Color intensity, Hue, OD280/OD315 of diluted wines, and Proline. One of the goals is to use these 13 features to classify the cultivar.

The number of directions d is chosen as 2 according to Algorithm 5. We tested PCA, DT-SIR, M-Lasso, and Lasso-SIR, for obtaining these two directions. In Figure 2, we plotted the projection of the data onto the space spanned by two components. The colors of the points correspond to three different cultivars. It is clearly seen that Lasso-SIR provided the best separation of the three cultivars. When using one vertical and one horizontal line to classify three groups, only one subject would be wrongly classified.

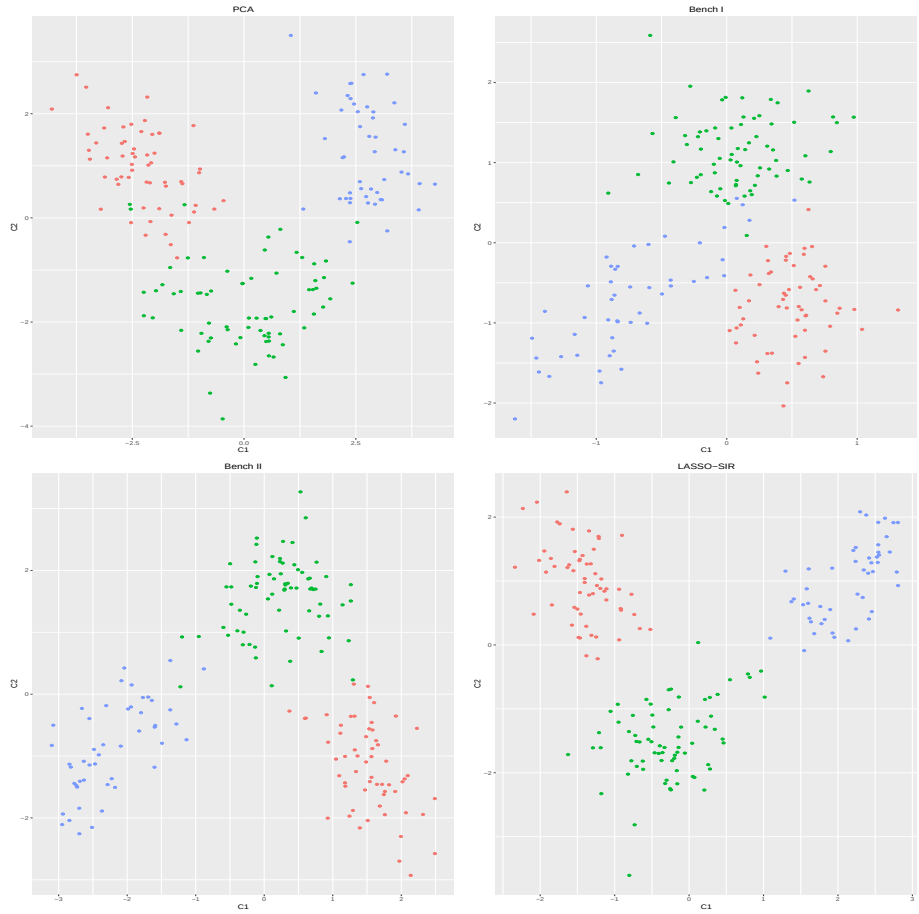

Fig. 2. We plotted the second component versus the first component for all the wines, which are labeled with different colors, representing different cultivars (1–red, 2–green, 3–blue). The four methods for calculating the directions are PCA, DT-SIR, M-Lasso, and Lasso-SIR from top-left to bottom-right. It is clearly seen that Lasso-SIR offered the best separation among these three groups.

6. Discussion

In this paper, we have proposed Lasso-SIR, an efficient high-dimensional variant of the sliced inverse regression [Li, 1991], to obtain a sparse solution to the estimation of the sufficient dimensional reduction space for multiple index models, and showed that it is rate optimal if $n\lambda = p^\alpha$ for some $\alpha > 1/2$, where λ is the smallest nonzero eigenvalue of $var(\mathbb{E}[\boldsymbol{x}|y])$. This technical assumption on n, λ and p is slightly disappointing from the ultra-high dimensional perspective. We believe that this technical assumption arises from an intrinsic limitation in estimating the central subspace, i.e., some further sparsity assumptions on either $\boldsymbol{\Sigma}$ or $var(\mathbb{E}[\boldsymbol{x}|y])$ or both are needed to show the consistency of any estimation method. We will address such extensions in our future researches.

Cautious reader may find that the concept of "pseudo-response variable" is not essential for developing the theory of the Lasso-SIR algorithm. However, by re-formulating the SIR method as a linear regression problem using the pseudo-response variable, we can formally consider the model selection consistency, regularization path and many others for the Lasso-SIR. In other words, the Lasso-SIR does not only provide an efficient high dimensional variant of SIR, but also extends the rich theory developed for Lasso linear regression in the past decades to the multiple index models.

Acknowledgments

This research was partly supported by NSF Grant IIS-1633283, NSF Grant DMS-1120368, and NIH Grant R01 GM113242-01.

Appendix: Sktech of Proofs of Theorems 1 and 2

We need two propositions to prove Theorems 1 and 2. The first one reveals some properties of the eigenvectors of $\widehat{\boldsymbol{\Lambda}}_H$.

PROPOSITION 1. *Assume that Conditions* **A1**), **A2**) *and* **A3**) *hold. Let* $\widehat{\boldsymbol{\Lambda}}_H = \frac{1}{H}\boldsymbol{X}_H\boldsymbol{X}_H^\tau$ *be the SIR estimate of* $\boldsymbol{\Lambda} = var(\mathbb{E}[\boldsymbol{x}|y])$ *using* (2). *Let* $\widehat{\boldsymbol{\eta}}_j$ *be the eigenvector of unit length associated with the* j*-th eigenvalue* $\widehat{\lambda}_j$ *of* $\widehat{\boldsymbol{\Lambda}}$, $j = 1, ..., H$. *If* $n\lambda = p^\alpha$ *for some* $\alpha > 1/2$, *there exist positive constants* C_1 *and* C_2 *such that*

$$\|P_{\boldsymbol{\Lambda}}\widehat{\boldsymbol{\eta}}_j\|_2 \geq C_1\sqrt{\frac{\lambda}{\widehat{\lambda}_j}}, 1 \leq j \leq d \text{ and } \|P_{\boldsymbol{\Lambda}}\widehat{\boldsymbol{\eta}}_j\|_2 \leq C_2\frac{\sqrt{p\log(p)}}{n\lambda}\sqrt{\frac{\lambda}{\widehat{\lambda}_j}}, d+1 \leq j \leq H \qquad (10)$$

hold with high probability. Here we say that an event Ω *happens with high probability if* $\mathbb{P}(\Omega) \geq 1 - C_3\exp(-C_4\log(p))$ *for some constants* C_3 *and* C_4.

Remark: This result might be of independent interest. In order to justify that the sparsity assumption for the high dimensional setting is necessary, Lin et al. [2015] have shown that $\mathbb{E}[\angle(\boldsymbol{\eta}_1, \widehat{\boldsymbol{\eta}}_1)] = 0$ if and only if $\lim \frac{p}{n\lambda} = 0$. Proposition 1 states that the projection of $\sqrt{\widehat{\lambda}_j}\widehat{\boldsymbol{\eta}}_j$, $j = 1, ..., d$, onto the true direction is non-zero if $n\lambda > p^{1/2}$.

The proof of Proposition 1 is technical. To improve the readability, we sketch the key steps here and put the detailed proof in the online supplementary file. Let $\boldsymbol{x} = \boldsymbol{z} + \boldsymbol{w}$ be the orthogonal decomposition with respect to $col(var(\mathbb{E}[\boldsymbol{x}|y]))$ and its orthogonal complement. We define two $p \times H$ matrices $\boldsymbol{Z}_H = (\boldsymbol{z}_{1,\cdot}, ..., \boldsymbol{z}_{H,\cdot})$ and $\boldsymbol{W}_H = (\boldsymbol{w}_{1,\cdot}, ..., \boldsymbol{w}_{H,\cdot})$ whose definition are similar to the definition of $\boldsymbol{X}_H$. We then have the following decomposition

$$\boldsymbol{X}_H = \boldsymbol{Z}_H + \boldsymbol{W}_H. \tag{11}$$

By definition, we know that $\boldsymbol{Z}_H \perp\!\!\!\perp \boldsymbol{W}_H$, $\boldsymbol{Z}_H^\tau \boldsymbol{W}_H = 0$, and $\boldsymbol{W}_H = \frac{1}{\sqrt{c}} \boldsymbol{\Sigma}_1^{1/2} \boldsymbol{E}_H$, where $\boldsymbol{\Sigma}_1$ is the covariance matrix of $\boldsymbol{w}$ and $\boldsymbol{E}_H$ is a $p \times H$ matrix with $i.i.d.$ standard normal entries. We have the following lemma:

LEMMA 1. *Assume that ν_1 is a large enough constant only depends on the constant κ. For sufficiently large constant $\mathfrak{a}$, the event $\Omega = \Omega_1 \cap \Omega_2$ holds with high probability where*

$$\Omega_1 = \Big\{ \omega \mid (1 - \frac{\kappa}{2\nu_1})\lambda \le \lambda_{\min}(\frac{1}{H}\boldsymbol{Z}_H^\tau \boldsymbol{Z}_H) \le \lambda_{\max}(\frac{1}{H}\boldsymbol{Z}_H^\tau \boldsymbol{Z}_H) \le (1 + \frac{1}{2\nu_1})\kappa\lambda \Big\} \tag{12}$$

$$\Omega_2 = \Big\{ \omega \mid \Big\| \frac{1}{H}\boldsymbol{W}_H^\tau \boldsymbol{W}_H - \frac{tr(\boldsymbol{\Sigma}_1)}{n} \boldsymbol{I}_H \Big\|_F \le \mathfrak{a} \frac{\sqrt{p \log(p)}}{n} \Big\}. \tag{13}$$

PROOF. *The proof is presented in the online supplementary file.* □

For any $\omega \in \Omega$, we can choose a $p \times p$ orthogonal matrix T and an $H \times H$ orthogonal matrix S such that

$$\frac{1}{H} T \boldsymbol{Z}_H(\omega) S = \begin{pmatrix} B_1 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} \text{ and } \frac{1}{H} T \boldsymbol{W}_H(\omega) S = \begin{pmatrix} 0 & 0 \\ B_2 & B_3 \\ 0 & B_4 \end{pmatrix} \tag{14}$$

where B_1 is a $d \times d$ matrix, B_2 is a $d \times d$ matrix, B_3 is a $d \times (H - d)$ matrix and B_4 is a $(p - 2d) \times (H - d)$ matrix. By definition of the event ω, we have

$$\begin{aligned} &(1 - \frac{\kappa}{2\nu_1})\lambda \le \lambda_{\min}(B_1^\tau B_1) \le \lambda_{\max}(B_1^\tau B_1) \le (1 + \frac{1}{2\nu_1})\kappa\lambda \\ &\Big\| \begin{pmatrix} B_2^\tau B_2 & B_2^\tau B_3 \\ B_3^\tau B_2 & B_3^\tau B_3 + B_4^\tau B_4 \end{pmatrix} - \frac{tr(\boldsymbol{\Sigma}_1)}{n} \boldsymbol{I}_H \Big\|_F \le \mathfrak{a}' \frac{\sqrt{p \log(p)}}{n} \end{aligned} \tag{15}$$

for some constant $\mathfrak{a}'$. Thus, we only need to prove the following lemma:

LEMMA 2. *Assume that $n\lambda = p^\alpha$ for some $\alpha > 1/2$. Let $\boldsymbol{M} = \begin{pmatrix} B_1 & 0 \\ B_2 & B_3 \\ 0 & B_4 \end{pmatrix}$ be a $p \times H$ matrix, where B_1 is a $d \times d$ matrix, B_2 is a $d \times d$ matrix, B_3 is a $d \times (H - d)$ matrix and B_4 is a $(p - 2d) \times (H - d)$ matrix satisfying (15). Let $\widehat{\boldsymbol{\eta}}_j$ be the eigenvector associated with the j-th eigenvalue $\widehat{\lambda}_j$ of $\boldsymbol{M}\boldsymbol{M}^\tau$, $j = 1, ..., H$. Then the length of the projection of $\widehat{\boldsymbol{\eta}}_j$ onto its first d-coordinates is at least $C\sqrt{\frac{\lambda}{\widehat{\lambda}_j}}$ for $j = 1, ..., d$ and is at most $C\frac{\sqrt{p \log(p)}}{n\lambda}\sqrt{\frac{\lambda}{\widehat{\lambda}_j}}$ for $j = d + 1, ..., H$.*

PROOF. *See the online supplementary file.*

Next, we briefly review the restricted eigenvalue (RE) property, which was first introduced in Raskutti et al. [2010], before stating the second proposition. Given a set $S \subset [p] = \{1, ..., p\}$, define the set $\mathcal{C}(S, \alpha)$ as

$$\mathcal{C}(S, \alpha) = \{\theta \in \mathbb{R}^p \mid \|\theta_{S^c}\|_1 \leq \alpha \|\theta_S\|_1\}.$$

We say that a sample matrix $\boldsymbol{X}\boldsymbol{X}^\tau/n$ satisfies the restricted eigenvalue condition over S with parameter $(\alpha, \kappa) \in [1, \infty) \times (0, \infty)$ if

$$\frac{1}{n}\theta^\tau \boldsymbol{X}\boldsymbol{X}^\tau \theta \geq \kappa^2 \|\theta\|_2^2, \quad \forall \theta \in \mathcal{C}(S, \alpha). \tag{16}$$

If (16) holds uniformly for all the subsets S with cardinality s, we say that $\boldsymbol{X}\boldsymbol{X}^\tau/n$ satisfies the restricted eigenvalue condition of order s with parameter (α, κ). Similarly, we say that the covariance matrix $\boldsymbol{\Sigma}$ satisfies the RE condition over S with parameter (α, κ) if $\|\boldsymbol{\Sigma}^{1/2}\theta\|_2 \geq \kappa\|\theta\|$ for all $\theta \in \mathcal{C}(S, \alpha)$. Additionally, if this condition holds uniformly for all the subsets S with cardinality s, we say that $\boldsymbol{\Sigma}$ satisfies the restricted eigenvalue condition of order s with parameter (α, κ). The following Corollary is borrowed from Raskutti et al. [2010].

COROLLARY 2. *Suppose that $\boldsymbol{\Sigma}$ satisfies the RE condition of order s with parameter (α, κ). Let $\boldsymbol{X}$ be the $p \times n$ matrix formed by n i.i.d samples from $N(0, \boldsymbol{\Sigma})$. For some universal positive constants $\mathfrak{a}_1, \mathfrak{a}_2$ and $\mathfrak{a}_3$, if the sample size satisfies*

$$n > \mathfrak{a}_3 \frac{(1+\alpha)^2 \max_{i \in [p]} \boldsymbol{\Sigma}_{ii}}{\kappa^2} s \log(p),$$

then the matrix $\frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\tau$ satisfies the RE condition of order s with parameter $(\alpha, \frac{\kappa}{8})$ with probability at least $1 - \mathfrak{a}_1 \exp(-\mathfrak{a}_2 n)$.

It is clear that $\lambda_{\min}(\boldsymbol{\Sigma}) \geq C_{\min}$ implies that $\boldsymbol{\Sigma}$ satisfies the RE condition of any order s with parameter $(3, \sqrt{C_{\min}})$. Thus, we have the following proposition.

PROPOSITION 2. *Assume that Condition* **A1)** *holds. For some universal constants $\mathfrak{a}_1$, $\mathfrak{a}_2$ and $\mathfrak{a}_3$, if the sample size satisfies that $n > \mathfrak{a}_1 s \log(p)$, then the matrix $\frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\tau$ satisfies the RE condition for any order s with parameter $(3, \sqrt{C_{\min}}/8)$ with probability at least $1 - \mathfrak{a}_2 \exp(-\mathfrak{a}_3 n)$.*

Next, we sketch key points and intermediate results of the proofs of the two theorems and leave details in the online supplementary files.

Sketch of the proof of Theorem 1. Recall that for single index model $y = f(\boldsymbol{\beta}_0^\tau \boldsymbol{x}, \epsilon)$ where $\boldsymbol{\beta}_0$ is a unit vector, we have denoted by $\widehat{\boldsymbol{\eta}}$ the eigenvector of $\widehat{\boldsymbol{\Lambda}}_H$ associated with the largest eigenvalue $\widehat{\lambda}$. Let $\widehat{\boldsymbol{\beta}}$ be the minimizer of

$$\mathcal{L}_{\boldsymbol{\beta}} = \frac{1}{2n}\|\widetilde{\boldsymbol{y}} - \boldsymbol{X}^\tau \boldsymbol{\beta}\| + \mu\|\boldsymbol{\beta}\|_1,$$

where $\widetilde{\boldsymbol{y}} \in \mathbb{R}^n$ such that $\widehat{\boldsymbol{\eta}} = \frac{1}{n}\boldsymbol{X}\widetilde{\boldsymbol{y}}$. Let $\boldsymbol{\eta}_0 = \boldsymbol{\Sigma}\boldsymbol{\beta}_0$, $\widetilde{\boldsymbol{\eta}} = P_{\eta_0}\widehat{\boldsymbol{\eta}}$ and $\widetilde{\boldsymbol{\beta}} = \boldsymbol{\Sigma}^{-1}\widetilde{\boldsymbol{\eta}} \propto \boldsymbol{\beta}_0$. Since we are interested in the distance between the directions of $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}_0$, we consider the difference $\boldsymbol{\delta} = \widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}$. A slight modification of the argument in Bickel et al. [2009] implies that, if we choose $\mu = C\sqrt{\frac{\log(p)}{n\widehat{\lambda}}}$ for sufficiently large constant C, we have $\|\boldsymbol{\delta}\|_2 \leq C_1\sqrt{\frac{s\log(p)}{n\widehat{\lambda}}}$ with high probability. The detailed arguments are put in the online supplementary file. The Proposition 1, Condition **A1**) and $\widetilde{\boldsymbol{\beta}} = \boldsymbol{\Sigma}^{-1}\widetilde{\boldsymbol{\eta}}$, imply that $C_1\sqrt{\frac{\widehat{\lambda}}{\widetilde{\lambda}}} \leq \|\widetilde{\boldsymbol{\beta}}\|_2 \leq C_2$ holds with high probability for some constants C_1 and C_2. Thus, we have

$$\|P_{\widehat{\boldsymbol{\beta}}} - P_{\boldsymbol{\beta}_0}\|_F = \|P_{\widehat{\boldsymbol{\beta}}} - P_{\widetilde{\boldsymbol{\beta}}}\|_F \leq 4\frac{\|\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}\|_2}{\|\widetilde{\boldsymbol{\beta}}\|_2} = 4\|\boldsymbol{\delta}\|_2/\|\widetilde{\boldsymbol{\beta}}\|_2 \leq C\sqrt{\frac{s\log(p)}{n\lambda}} \tag{17}$$

holds with high probability. □

Proof of Theorem 2. Let $\widehat{\boldsymbol{\eta}}_j$ be the (unit) eigenvectors associated with the j-th eigenvalues of $\widehat{\boldsymbol{\Lambda}}_H$, $j = 1, ..., d$. Let $\widetilde{\boldsymbol{\eta}}_j = P_{\boldsymbol{\Lambda}}\widehat{\boldsymbol{\eta}}_j$, $\widetilde{\boldsymbol{\beta}}_j = \boldsymbol{\Sigma}^{-1}\widetilde{\boldsymbol{\eta}}_j$ and $\boldsymbol{\gamma}_j = \widetilde{\boldsymbol{\beta}}_j/\|\widetilde{\boldsymbol{\beta}}_j\|_2$. Applying the argument in Theorem 1 to other eigenvectors gives us that

$$\|\widehat{\boldsymbol{\beta}}_j - \widetilde{\boldsymbol{\beta}}_j\|_2 \leq C\sqrt{\frac{s\log(p)}{n\widehat{\lambda}_j}} \text{ and } \|P_{\widehat{\boldsymbol{\beta}}_j} - P_{\widetilde{\boldsymbol{\beta}}_j}\|_F \leq C\sqrt{\frac{s\log(p)}{n\lambda}} \tag{18}$$

hold with high probability. Thus, if i) the lengths of $\widetilde{\boldsymbol{\beta}}_j, j = 1, ..., d$, are bounded below by $C\sqrt{\frac{\lambda}{\widehat{\lambda}_j}}$ with high probability and, ii) the angles between any two vectors of $\widetilde{\boldsymbol{\beta}}_j, j = 1, ..., d$, are bounded below by some constant with high probability, then the Gram-Schmit process implies that $\|P_{\widehat{\boldsymbol{B}}} - P_{\boldsymbol{B}}\|_F \leq C\sqrt{\frac{s\log(p)}{n\lambda}}$ holds with high probability from (18) and the assumption that d is bounded.

We only need to verify i) and ii). i) follows from the Proposition 1 and the Condition **A1**), since $\widetilde{\boldsymbol{\beta}}_j = \boldsymbol{\Sigma}^{-1}\widetilde{\boldsymbol{\eta}}_j, j = 1, ..., d$. ii) can be easily derived from the following two intuitively straightforward statements.

I. The angles between any two vectors in $\widetilde{\boldsymbol{\eta}}_j$'s , $j = 1, ..., d$ are nearly $\pi/2$. Since $n\lambda = p^\alpha$ for some $\alpha > 1/2$, we only need to prove that

$$\left|\cos\left(\angle(\widetilde{\boldsymbol{\eta}}_j, \widetilde{\boldsymbol{\eta}}_j)\right)\right| \leq C\frac{\sqrt{p\log(p)}}{n\lambda} \tag{19}$$

holds with high probability for any $i \neq j$. Recall that We have the following decomposition $\boldsymbol{X}_H = \boldsymbol{Z}_H + \boldsymbol{W}_H$. It is easy to see that $col(\boldsymbol{Z}_H) = col(var(\mathbb{E}[\boldsymbol{x}|y]))$ and $\sqrt{c}\, cov(\boldsymbol{w})^{-1/2}\boldsymbol{W}_H$ is identically distributed to a matrix, $\mathcal{E}_1$, with all the entries are $i.i.d.$ standard normal random variables. Let us choose an orthogonal matrix T such that $\frac{1}{\sqrt{H}}T\boldsymbol{Z}_H = \begin{pmatrix} \boldsymbol{A} \\ 0 \end{pmatrix}$ and $\frac{1}{\sqrt{H}}T\boldsymbol{W}_H = \begin{pmatrix} 0 \\ \boldsymbol{B} \end{pmatrix}$ where $\boldsymbol{A}$ is a $d \times H$ matrix and $\boldsymbol{B}$ is a $(p-d)\times$ H matrix. Thus, $T\widehat{\boldsymbol{\eta}}_j$

is the eigenvector of $\frac{1}{H}T\boldsymbol{X}_H\boldsymbol{X}_H^\tau T^\tau$ associated with the j-th eigenvalue $\widehat{\lambda}_j$, $j=1,...,d$. If we have *a)* $\lambda_{\min}(\boldsymbol{A}\boldsymbol{A}^\tau)\geq\lambda$, *b)* $\|P_{col(T\boldsymbol{Z}_H)}(T\widehat{\boldsymbol{\eta}}_j)\|_2\geq C\sqrt{\frac{\lambda}{\widehat{\lambda}_j}}$ and *c)* $\|\boldsymbol{B}^\tau\boldsymbol{B}-\mu\boldsymbol{I}_H\|_F\leq C\frac{\sqrt{p\log(p)}}{n}$ for some scalar $\mu>0$, then the statement **I** is reduced to the following linear algebra lemma.

LEMMA 3. *Let $\boldsymbol{A}$ be a $d\times H$ matrix $(d<H)$ with $\lambda_{\min}(\boldsymbol{A}\boldsymbol{A}^\tau)=\lambda$. Let $\boldsymbol{B}$ be a $(p-d)\times H$ matrix such that $\|\boldsymbol{B}^\tau\boldsymbol{B}-\mu\boldsymbol{I}_H\|_F^2\leq C\frac{\sqrt{p\log(p)}}{n}$. Let $\widehat{\boldsymbol{\xi}}_j$ be the j-th (unit) eigenvector of $\boldsymbol{C}\boldsymbol{C}^\tau$ associated with the j-th eigenvalue $\widehat{\lambda}_j$ where $\boldsymbol{C}^\tau=(\boldsymbol{A}^\tau,\boldsymbol{B}^\tau)$ and $\widetilde{\boldsymbol{\xi}}_j$ be the projection of $\widehat{\boldsymbol{\xi}}_j$ onto its first d-coordinates. If $\|\widetilde{\boldsymbol{\xi}}_j\|_2\geq C\sqrt{\frac{\lambda}{\widehat{\lambda}_j}}$, then for any $i\neq j$,*

$$\left|\cos\left(\angle(\widetilde{\boldsymbol{\xi}}_i,\widetilde{\boldsymbol{\xi}}_j)\right)\right|\leq C\frac{\sqrt{p\log(p)}}{n\lambda}. \tag{20}$$

Thus, $\widetilde{\boldsymbol{\xi}}_i{}'s$ are nearly orthogonal if $n\lambda=p^\alpha$ for some $\alpha>1/2$.

PROOF. *Let $\widehat{\boldsymbol{\alpha}}_j=\boldsymbol{C}^\tau\widehat{\boldsymbol{\xi}}_j$, then $\widehat{\boldsymbol{\xi}}_j=\frac{1}{\widehat{\lambda}_j}\boldsymbol{C}\boldsymbol{\alpha}_j$ and $\widetilde{\boldsymbol{\xi}}_j=\frac{1}{\widehat{\lambda}_j}\boldsymbol{A}\boldsymbol{\alpha}_j$. It is easy to see that $\|\widehat{\boldsymbol{\alpha}}_j\|_2=\sqrt{\widehat{\lambda}_j}$ and $\|\boldsymbol{C}\widehat{\boldsymbol{\alpha}}_j\|_2\geq\widehat{\lambda}_j$. Since $\widehat{\boldsymbol{\alpha}}_j/\sqrt{\widehat{\lambda}_j}$ is also the (unit) eigenvector of*

$$\boldsymbol{C}^\tau\boldsymbol{C}=\boldsymbol{A}^\tau\boldsymbol{A}+\mu\boldsymbol{I}+(\boldsymbol{B}^\tau\boldsymbol{B}-\mu\boldsymbol{I}),$$

for any $i\neq j$, we have

$$\begin{aligned}0&=\widehat{\boldsymbol{\alpha}}_j^\tau\boldsymbol{C}^\tau\boldsymbol{C}\widehat{\boldsymbol{\alpha}}_i=\widehat{\boldsymbol{\alpha}}_j^\tau\boldsymbol{A}^\tau\boldsymbol{A}\widehat{\boldsymbol{\alpha}}_j+\mu\widehat{\boldsymbol{\alpha}}_j^\tau\widehat{\boldsymbol{\alpha}}_i+\widehat{\boldsymbol{\alpha}}_j^\tau(\boldsymbol{B}^\tau\boldsymbol{B}-\mu\boldsymbol{I})\widehat{\boldsymbol{\alpha}}_i\\&=\widehat{\lambda}_j\widehat{\lambda}_i\widetilde{\boldsymbol{\xi}}_j^\tau\widetilde{\boldsymbol{\xi}}_i+\widehat{\boldsymbol{\alpha}}_j^\tau(\boldsymbol{B}^\tau\boldsymbol{B}-\mu\boldsymbol{I})\widehat{\boldsymbol{\alpha}}_i.\end{aligned}$$

Since $\|\boldsymbol{B}^\tau\boldsymbol{B}-tr(\boldsymbol{\Sigma})\boldsymbol{I}_H\|_F\leq C\frac{\sqrt{p\log(p)}}{n}$ and $\|\widehat{\boldsymbol{\xi}}_j\|_2\geq C\sqrt{\frac{\lambda}{\widehat{\lambda}_j}}$, $\forall i\neq j$, we have

$$\left|\frac{\boldsymbol{\xi}_j^\tau\boldsymbol{\xi}_i}{\|\boldsymbol{\xi}_i\|_2\|\boldsymbol{\xi}_j\|_2}\right|\leq C\left|\widetilde{\boldsymbol{\xi}}_j^\tau\widetilde{\boldsymbol{\xi}}_i\frac{\widehat{\lambda}_j^{1/2}\widehat{\lambda}_i^{1/2}}{\lambda}\right|=C\left|\frac{1}{\lambda}\frac{\widehat{\boldsymbol{\alpha}}_j^\tau}{\widehat{\lambda}_j^{1/2}}(\boldsymbol{B}^\tau\boldsymbol{B}-\mu\boldsymbol{I})\frac{\widehat{\boldsymbol{\alpha}}_i}{\widehat{\lambda}_i^{1/2}}\right|\leq C\frac{\sqrt{p\log(p)}}{n\lambda}.$$

□

Note that *a)* follows from the Lemma 3, *b)* follows from Proposition 1 and *c)* follows from the Lemma 8. Thus statement **I** holds.

II. The angles between any two vectors in $\widetilde{\boldsymbol{\beta}}_j{}'s$ are bounded away from 0. Since $\widetilde{\boldsymbol{\beta}}_j=\boldsymbol{\Sigma}^{-1}\widetilde{\boldsymbol{\eta}}_j$, we only need to prove that there exists a positive constant $\zeta<1$ such that

$$\left|\frac{\widetilde{\boldsymbol{\eta}}_i^\tau\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}^{-1}\widetilde{\boldsymbol{\eta}}_i}{\|\boldsymbol{\Sigma}^{-1}\widetilde{\boldsymbol{\eta}}_i\|_2\|\boldsymbol{\Sigma}^{-1}\widetilde{\boldsymbol{\eta}}_j\|_2}\right|\leq\zeta. \tag{21}$$

Let $(\widetilde{\boldsymbol{\eta}}_1/\|\widetilde{\boldsymbol{\eta}}_1\|_2,...,\widetilde{\boldsymbol{\eta}}_d/\|\widetilde{\boldsymbol{\eta}}_d\|_2)=\boldsymbol{T}\boldsymbol{M}$, where $\boldsymbol{T}$ is a $p\times d$ orthogonal matrix. Since $\widetilde{\boldsymbol{\eta}}_j/\|\widetilde{\boldsymbol{\eta}}_j\|_2'{}s$ are nearly mutually orthogonal, we know that $\boldsymbol{M}^\tau\boldsymbol{M}$ is nearly an identity matrix. Thus, by some continuity argument, the statement is reduced to the following linear algebra lemma.

LEMMA 4. *Let $\boldsymbol{A}$ be a $p\times p$ positive definite matrix such that $C_{\min} \leq \lambda_{\min}(\boldsymbol{A}) \leq \lambda_{\max}(\boldsymbol{A}) \leq C_{\max}$ for some positive constants $C_{\min}$ and $C_{\max}$. There exists constant $0 < \zeta < 1$ such that for any $p \times d$ orthogonal matrix $\boldsymbol{B}$, we have*

$$\left| \frac{\boldsymbol{B}_{*,i}^{\tau} \boldsymbol{A}^{\tau} \boldsymbol{A} \boldsymbol{B}_{*,j}}{\|\boldsymbol{A}\boldsymbol{B}_{*,i}\|_2 \|\boldsymbol{A}\boldsymbol{B}_{*,j}\|_2} \right| \leq \zeta \quad \forall i \neq j. \tag{22}$$

PROOF. *When d is finite, without loss of generality, we can assume that B is a $p \times 2$ matrix. Note that the expression on the left side is invariant under orthogonal transformation of B. We can simply assume that B is a matrix with the last $p-2$-rows consisting of all zeros. The result follows immediately based on basic calculation.* □

References

P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

Cristina Butucea, Yuri I Ingster, et al. Detection of a sparse submatrix of a high-dimensional noisy matrix. *Bernoulli*, 19(5B):2652–2688, 2013.

E. Candes and T. Tao. The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, 2007.

C. H. Chen and K. C. Li. Can SIR be as popular as multiple linear regression? *Statistica Sinica*, 8(2):289–316, 1998.

H. Chen. Estimation of a projection-pursuit type regression model. *The Annals of Statistics*, 19(1):142–157, 1991.

D. R. Cook. *Regression graphics*. Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons, Inc., New York, 1998.

D. R. Cook. SAVE: a method for dimension reduction and graphics in regression. *Communications in statistics-Theory and methods*, 29(9-10):2109–2121, 2000.

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

N. Duan and K. C. Li. Slicing regression: a link-free regression method. *The Annals of Statistics*, 19(2):505–530, 1991.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

J. Fan, Q. Shao, and W. Zhou. Are discoveries spurious? distributions of maximum spurious correlations and their applications. *arXiv preprint arXiv:1502.04237*, 2015.

J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.

J. H. Friedman and W. Stuetzle. Projection pursuit regression. *Journal of the American statistical Association*, 76(376):817–823, 1981.

I. Guyon, S. Gunn, A. Ben-Hur, and G. Dror. Result analysis of the NIPS 2003 feature selection challenge. In *Advances in neural information processing systems*, pages 545–552, 2004.

T. Hsing and R. J. Carroll. An asymptotic theory for sliced inverse regression. *The Annals of Statistics*, 20(2):1040–1061, 1992.

Yuri I Ingster, Alexandre B Tsybakov, Nicolas Verzelen, et al. Detection boundary in sparse regression. *Electronic Journal of Statistics*, 4:1476–1526, 2010.

B. Jiang and J. S. Liu. Variable selection for general index models via sliced inverse regression. *The Annals of Statistics*, 42(5):1751–1786, 2014.

S. Jung and J. S. Marron. PCA consistency in high dimension, low sample size context. *The Annals of Statistics*, 37(6B):4104–4130, 2009.

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

L. Li. Sparse sufficient dimension reduction. *Biometrika*, 94(3):603–613, 2007.

L. Li and C. J. Nachtsheim. Sparse sliced inverse regression. *Technometrics*, 48(4):503–510, 2006.

M. Lichman. UCI machine learning repository, 2013. URL `http://archive.ics.uci.edu/ml`.

Q. Lin, Z. Zhao, and J. S. Liu. On consistency and sparsity for sliced inverse regression in high dimensions. *arXiv preprint arXiv:1507.03895*, 2015.

Q. Lin, X. Li, H. Dong, and J. S. Liu. On optimality of sliced inverse regression in high dimensions. 2016.

M. Neykov, Q. Lin, and J. S. Liu. Signed support recovery for single index models in high-dimensions. *Annals of Mathematical Sciences and Applications*, 1(2):379–426, 2016.

L. Ni, D. R. Cook, and C. L. Tsai. A note on shrinkage sliced inverse regression. *Biometrika*, 92(1):242–247, 2005.

G. Raskutti, M. J. Wainwright, and B. Yu. Restricted eigenvalue properties for correlated gaussian designs. *The Journal of Machine Learning Research*, 11:2241–2259, 2010.

G. Raskutti, M. J. Wainwright, and B. Yu. Minimax rates of estimation for high-dimensional linear regression over-balls. *Information Theory, IEEE Transactions on*, 57(10):6976–6994, 2011.

N. Simon, J. Friedman, T. Hastie, and R. Tibshirani. A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245, 2013.

T. E. Thorgeirsson, D. F. Gudbjartsson, and many others. Sequence variants at CHRNB3-CHRNA6 and CYP2A6 affect smoking behavior. *Nature genetics*, 42(5):448–453, 2010.

G. A. Thorisson, A. V. Smith, L. Krishnan, and L. D. Stein. The international HapMap project web site. *Genome research*, 15(11):1592–1593, 2005.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B*, pages 267–288, 1996.

Y. Xia, H. Tong, W. K. Li, and L. X. Zhu. An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3): 363–410, 2002.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

W. Zhong, T. Zhang, Y. Zhu, and J. S. Liu. Correlation pursuit: forward stepwise variable selection for index models. *Journal of the Royal Statistical Society: Series B*, 74(5):849–870, 2012.

L. Zhu, B. Miao, and H. Peng. On sliced inverse regression with high-dimensional covariates. *Journal of the American Statistical Association*, 101(474):640–643, 2006.

Hui Zou, Trevor Hastie, and Robert Tibshirani. Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286, 2006.

Web-based supporting materials for

"Sparse Sliced Inverse Regression for High Dimensional Data"

by "Lin, Q., Zhao, Z., and Liu, S. J."

A. Proofs of Theorems

A.1. Proof of Theorem 1

Let $T = supp(\boldsymbol{\beta}_0)$, then $|T| \leq s$. For a vector $\boldsymbol{\gamma} \in \mathbb{R}^p$, let $\boldsymbol{\gamma}_T$ and $\boldsymbol{\gamma}_{T^c}$ be the sub-vector consists of the components of $\boldsymbol{\gamma}$ in T and T^c respectively. Let us introduce the event $\mathsf{E} = \mathsf{E}_1 \cap \mathsf{E}_2 \cap \mathsf{E}_3 \cap \mathsf{E}_4$ where

$$\mathsf{E}_1 = \Big\{ \omega \mid \|\boldsymbol{\delta}_{T^c}\|_1 \leq 3\|\boldsymbol{\delta}_T\|_1 \Big\},$$

$$\mathsf{E}_2 = \Big\{ \omega \mid \frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\tau \text{ satisfies the RE condition of order } s \text{ with parameter}(3, \frac{\sqrt{C_{\min}}}{8}) \Big\}$$

$$\mathsf{E}_3 = \Big\{ \omega \mid \|\frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\tau\boldsymbol{\delta}\|_\infty \leq \mu \text{ where } \mu = \mathfrak{b}_1\sqrt{\frac{\log(p)}{n\widehat{\lambda}}} \Big\}$$

$$\mathsf{E}_4 = \Big\{ \omega \mid \|P_{\boldsymbol{\Lambda}}\widehat{\boldsymbol{\eta}}\|_2 \geq \mathfrak{b}_2\sqrt{\frac{\widetilde{\lambda}}{\widehat{\lambda}}} \Big\}.$$

Here $\mathfrak{b}_1$ and $\mathfrak{b}_2$ are sufficiently large constants to be specified later. Because Proposition 1 implies that E_4 happens with high probability(i.e., $\mathbb{P}(\mathsf{E}_4) \geq 1 - C_1\exp(-C_2\log(p))$ for some constants C_1 and C_2), Proposition 2 implies that E_2 happens with high probability, Lemma 5 (see below) implies that E_3 happens with high probability, and Lemma 7 (see below) implies that E_1 happens with high probability, we conclude that E happens with high probability. Conditioning on E, we have

$$\frac{1}{64}C_{\min}\|\boldsymbol{\delta}\|_2^2 \leq \frac{1}{n}\|\boldsymbol{X}^\tau\boldsymbol{\delta}\|^2 \leq \frac{1}{n}\|\boldsymbol{X}\boldsymbol{X}^\tau\boldsymbol{\delta}\|_\infty\|\boldsymbol{\delta}\|_1 \leq 4\mu\|\boldsymbol{\delta}_T\|_1 \leq 4\sqrt{s}\mu\|\boldsymbol{\delta}_T\|_2.$$

Thus, conditioning on E, we have $\|\boldsymbol{\delta}\|_2 \leq \frac{256}{C_{\min}}\sqrt{s}\mu = C\sqrt{\frac{s\log(p)}{n\widehat{\lambda}}}$. Since we have $C_1\sqrt{\frac{\widetilde{\lambda}}{\widehat{\lambda}}} \leq \|\widetilde{\boldsymbol{\beta}}\|_2 \leq C_2$ conditioning on E, we know that

$$\|P_{\widehat{\boldsymbol{\beta}}} - P_{\boldsymbol{\beta}_0}\|_F = \|P_{\widehat{\boldsymbol{\beta}}} - P_{\widetilde{\boldsymbol{\beta}}}\|_F \leq 4\frac{\|\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}\|_2}{\|\widetilde{\boldsymbol{\beta}}\|_2} \leq C\sqrt{\frac{s\log(p)}{n\lambda}}.$$

□

LEMMA 5. *Assume that condition* **A1**), **A2**) *and* **A3**) *hold. Let* $\mu = A\sqrt{\frac{\log(p)}{n\widehat{\lambda}}}$. *For sufficiently large A, we have that*

$$\|\frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\tau\boldsymbol{\delta}\|_\infty \leq \mu \tag{A.1}$$

holds with high probability.

PROOF. *Since* $\boldsymbol{\delta} = \widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{\eta}} = \frac{1}{n}\boldsymbol{X}\widetilde{\boldsymbol{y}}$ *and* $\boldsymbol{\Sigma}\widetilde{\boldsymbol{\beta}} = \widetilde{\boldsymbol{\eta}}$, *we have*

$$\frac{1}{n}\|\boldsymbol{X}\boldsymbol{X}^\tau\boldsymbol{\delta}\|_\infty \leq \underbrace{\frac{1}{n}\|\boldsymbol{X}\boldsymbol{X}^\tau\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\widetilde{\boldsymbol{y}}\|_\infty}_{I} + \underbrace{\frac{1}{n}\|\boldsymbol{X}\widetilde{\boldsymbol{y}} - n\widetilde{\boldsymbol{\eta}}\|_\infty}_{II} + \underbrace{\|(\frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\tau - \boldsymbol{\Sigma})\widetilde{\boldsymbol{\beta}}\|_\infty}_{III}. \tag{A.2}$$

For I. By the definition of $\widehat{\boldsymbol{\beta}}$, *we have* $0 \in \frac{1}{n}\left(\boldsymbol{X}\boldsymbol{X}^\tau\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\widetilde{\boldsymbol{y}}\right) + \mu sgn\left(\widehat{\boldsymbol{\beta}}\right)$, *i.e.,* $\frac{1}{n}\|\boldsymbol{X}\boldsymbol{X}^\tau\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\widetilde{\boldsymbol{y}}\|_\infty \leq \mu$.

For II. Let $\boldsymbol{x} = \boldsymbol{z}+\boldsymbol{w}$ *be the orthogonal decomposition with respect to* $col(var(\mathbb{E}[\boldsymbol{x}|y]))$ *and its orthogonal complement. Recall that we have introduced the decomposition:* $\boldsymbol{X}_H = \boldsymbol{Z}_H + \boldsymbol{W}_H$. *It is easy to see that* $col(\boldsymbol{Z}_H) = col(\boldsymbol{\eta}_0)$ *and* $\sqrt{c}\ cov(\boldsymbol{w})^{-1/2}\boldsymbol{W}_H$ *is identically distributed to a matrix,* $\mathcal{E}_1$, *with all the entries are i.i.d. standard normal random variables. Let*

$$\boldsymbol{\alpha}_1 = \frac{1}{\sqrt{H}}\boldsymbol{Z}_H^\tau\widehat{\boldsymbol{\eta}}, \quad \boldsymbol{\alpha}_2 = \frac{1}{\sqrt{H}}\boldsymbol{W}_H^\tau\widehat{\boldsymbol{\eta}} \quad and \quad \boldsymbol{\alpha} = \frac{1}{\sqrt{H}}\boldsymbol{X}_H^\tau\widehat{\boldsymbol{\eta}} = \boldsymbol{\alpha}_1 + \boldsymbol{\alpha}_2.$$

Since $\frac{1}{H}\boldsymbol{X}_H\boldsymbol{X}_H^\tau\widehat{\boldsymbol{\eta}} = \widehat{\lambda}\widehat{\boldsymbol{\eta}}$, *we know that* $\|\boldsymbol{\alpha}\|_2^2 = \widehat{\lambda}$ *and* $\widehat{\boldsymbol{\eta}} = \frac{1}{\sqrt{H\widehat{\lambda}}}\boldsymbol{Z}_H\boldsymbol{\alpha} + \frac{1}{\sqrt{H\widehat{\lambda}}}\boldsymbol{W}_H\boldsymbol{\alpha}$. *From this, we know* $\widetilde{\boldsymbol{\eta}} = P_{\eta_0}\widehat{\boldsymbol{\eta}} = \frac{1}{\sqrt{H\widehat{\lambda}}}\boldsymbol{Z}_H\boldsymbol{\alpha}$ *and* $\widehat{\boldsymbol{\eta}} - P_{\eta_0}\widehat{\boldsymbol{\eta}} = \frac{1}{\sqrt{H\widehat{\lambda}}}\boldsymbol{W}_H\boldsymbol{\alpha}$. *Since* $\|\boldsymbol{\alpha}\|_1 \leq \sqrt{H}\|\boldsymbol{\alpha}\|_2 = \sqrt{H\widehat{\lambda}}$, *we know that* $\|\widehat{\boldsymbol{\eta}} - P_{\eta_0}\widehat{\boldsymbol{\eta}}\|_\infty \leq \frac{1}{\sqrt{\widehat{\lambda}}}\|\boldsymbol{W}_H\|_{\infty,\infty}$. *It is easy to see that for positive constant* $A_2(>1)$, *we have*

$$\mathbb{P}\left(\|\frac{cov(\boldsymbol{w})^{1/2}}{\sqrt{c}}\mathcal{E}_1\|_{\infty,\infty} > \lambda_{\max}^{1/2}(\boldsymbol{\Sigma})\sqrt{\frac{A_2 H\log(pH)}{n}}\right) \leq 2\exp^{-(A_2-1)\log(pH)},$$

i.e., by letting $A > 2\lambda_{\max}^{1/2}(\boldsymbol{\Sigma})\sqrt{A_2 H}$, *we have that* $\|\widehat{\boldsymbol{\eta}} - P_{\eta_0}\widehat{\boldsymbol{\eta}}\|_\infty \leq A\sqrt{\frac{\log(p)}{n\widehat{\lambda}}}$ *holds with high probability.*

For III. Let $\boldsymbol{x} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\epsilon}$ *where* $\boldsymbol{\epsilon} \sim N(0, \boldsymbol{I}_p)$, *then* $\boldsymbol{X} = \boldsymbol{\Sigma}^{1/2}\mathcal{E}$ *where* $\mathcal{E}$ *is a* $p \times n$ *matrix with i.i.d standard normal entries. Let*

$$\mathsf{F}(t) = \left\{\omega \mid \|\boldsymbol{\Sigma}^{1/2}\left(\frac{1}{n}\mathcal{E}\mathcal{E}^\tau - \boldsymbol{I}_p\right)\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\eta}_0\|_\infty > t\lambda_{max}(\boldsymbol{\Sigma})^{1/2}\lambda_{\min}(\boldsymbol{\Sigma})^{-1/2}\right\}.$$

We have the following elementary bound on this event.

LEMMA 6.

$$\mathbb{P}(\mathsf{F}(\mathsf{t})) \leq 4\exp^{-\frac{3nt^2}{16}+\log(p)}. \tag{A.3}$$

PROOF. *For any two deterministic vectors* $\boldsymbol{\alpha}_1$ *and* $\boldsymbol{\alpha}_2$, *let*

$$\widetilde{\mathbf{E}}(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2) = \{\ \omega \mid |\boldsymbol{\alpha}_1\left(\frac{1}{n}\mathcal{E}\mathcal{E}^\tau - \boldsymbol{I}_p\right)\boldsymbol{\alpha}_2| > t\|\boldsymbol{\alpha}_1\|_2\|\boldsymbol{\alpha}_2\|_2\ \}. \tag{A.4}$$

We know that for $0 < t < 1/2$,

$$\mathbb{P}\left(\widetilde{\mathbf{E}}(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)\right) \leq 4\exp^{-\frac{3}{16}nt^2}. \tag{A.5}$$

Combining with the facts $\|\boldsymbol{\Sigma}_{i,*}^{1/2}\| \leq \lambda_{max}(\boldsymbol{\Sigma})^{1/2}$ *and* $\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\eta}_0\|_2 \leq \lambda_{\min}(\boldsymbol{\Sigma})^{-1/2}$, *we know that*

$$\|\boldsymbol{\Sigma}^{1/2}\left(\frac{1}{n}\mathcal{E}\mathcal{E}^\tau - \boldsymbol{I}_p\right)\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\eta}_0\|_\infty > t\lambda_{max}(\boldsymbol{\Sigma})^{1/2}\lambda_{\min}(\boldsymbol{\Sigma})^{-1/2} \tag{A.6}$$

happens with probability at most $4\exp^{-\frac{3}{16}nt^2+\log(p)}$. □

Since $\widetilde{\boldsymbol{\beta}} = \boldsymbol{\Sigma}^{-1}\widetilde{\boldsymbol{\eta}}$, $\widetilde{\boldsymbol{\eta}} = \|\widetilde{\boldsymbol{\eta}}\|\boldsymbol{\eta}_0$ *and* $\|\widetilde{\boldsymbol{\eta}}\| \leq 1$, *we know that*

$$\frac{1}{n}\|\left(\boldsymbol{X}\boldsymbol{X}^\tau - n\boldsymbol{\Sigma}\right)\widetilde{\boldsymbol{\beta}}\|_\infty = \|\widetilde{\boldsymbol{\eta}}\|\|\boldsymbol{\Sigma}^{1/2}\left(\frac{1}{n}\mathcal{E}\mathcal{E}^\tau - \boldsymbol{I}_p\right)\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\eta}_0\|_\infty. \tag{A.7}$$

Conditioning on $\mathsf{F}(\sqrt{\frac{16A_3\log(p)}{3n}}) \cap \mathsf{E}_4$, *we have*

$$\frac{1}{n}\|\left(\boldsymbol{X}\boldsymbol{X}^\tau - n\boldsymbol{\Sigma}\right)\widetilde{\boldsymbol{\beta}}\|_\infty > C\sqrt{\frac{\lambda}{\widetilde{\lambda}}\frac{16A_3\log(p)}{3n}}\lambda_{max}(\boldsymbol{\Sigma})^{1/2}\lambda_{\min}(\boldsymbol{\Sigma})^{-1/2}. \tag{A.8}$$

Combining the Lemma 6 and Proposition 1, we know that (A.8) *happens with high probability.*

To summarize, we know that, for sufficiently large constant A, $\|\frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\tau\delta\|_\infty \leq A\sqrt{\frac{\log(p)}{n\widehat{\lambda}}}$ *holds with high probability.* □

LEMMA 7. *Let* $\mu = A\sqrt{\frac{\log(p)}{n\widehat{\lambda}}}$. *For sufficiently large constant* A, *we have* $\|\delta_{T^c}\|_1 \leq 3\|\delta_T\|_1$ *holds with high probability.*

PROOF. *Since* $\mathcal{L}_{\widehat{\boldsymbol{\beta}}} \leq \mathcal{L}_{\widetilde{\boldsymbol{\beta}}}$, *we have*

$$-\|\delta\|_1\|\frac{1}{n}\boldsymbol{X}\widetilde{y} - \frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\tau\widetilde{\boldsymbol{\beta}}\|_\infty \leq \mu(\|\widetilde{\boldsymbol{\beta}}\|_1 - \|\widehat{\boldsymbol{\beta}}\|_1). \tag{A.9}$$

Since $\|\widetilde{\boldsymbol{\beta}}\|_1 - \|\widehat{\boldsymbol{\beta}}\|_1 \leq \|\delta_T\|_1 - \|\delta_{T^c}\|_1$ *and* $\|\delta\|_1 = \|\delta_T\|_1 + \|\delta_{T^c}\|_1$, *if* $\|\frac{1}{n}\boldsymbol{X}\widetilde{y} - \frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\tau\widetilde{\boldsymbol{\beta}}\|_\infty \leq A_1\sqrt{\frac{\log(p)}{n\widehat{\lambda}}}$ *and* $A > 2A_1$, *we have* $\|\delta_{T^c}\|_1 \leq \frac{A+A_1}{A-A_1}\|\delta_T\|_1 \leq 3\|\delta_T\|_1$.

Since $\frac{1}{n}\|\boldsymbol{X}\widetilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{X}^\tau\widetilde{\boldsymbol{\beta}}\|_\infty \leq \|\widehat{\boldsymbol{\eta}} - \widetilde{\boldsymbol{\eta}}\|_\infty + \|(\frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\tau - \boldsymbol{\Sigma})\widetilde{\boldsymbol{\beta}}\|_\infty = II + III$ *where* II *and* III *are introduced in* (A.2). *From the bounds of* II *and* III *in Lemma 5, we know that* $\|\frac{1}{n}\boldsymbol{X}\widetilde{y} - \frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\tau\widetilde{\boldsymbol{\beta}}\|_\infty \leq A_1\sqrt{\frac{\log(p)}{n\widehat{\lambda}}}$ *holds with high probability for some* A_1 *(does not depend on choice of* μ*) and some constants* C_1 *and* C_2.

□

A.2. Proof of Theorem 3

Let $\widehat{\boldsymbol{\eta}}_i$ be the eigenvector associated with the i-th eigenvalue $\widehat{\lambda}_i$ of $\widehat{\boldsymbol{\Lambda}}_H$, $\widetilde{\boldsymbol{\eta}}_i = P_{\boldsymbol{\Lambda}}\widehat{\boldsymbol{\eta}}_i$ and $\widetilde{\boldsymbol{\beta}}_i = \boldsymbol{\Sigma}\widetilde{\boldsymbol{\eta}}_i$, $i = 1, \ldots, H$. The argument in Theorem 1 implies that, for any $1 \le i \le H$,

$$\|\widehat{\boldsymbol{\beta}}_i - \widetilde{\boldsymbol{\beta}}_i\|_2 \le C\sqrt{\frac{s\log(p)}{n\widehat{\lambda}}}. \tag{A.10}$$

The Proposition 1 implies that

$$\|\widetilde{\boldsymbol{\beta}}_i\|_2 \ge C_1\sqrt{\frac{\lambda}{\widetilde{\lambda}_i}}, 1 \le i \le d \text{ and } \|\widetilde{\boldsymbol{\beta}}_i\|_2 \le C_2\sqrt{\frac{\lambda}{\widetilde{\lambda}_i}}\frac{\sqrt{p\log(p)}}{n\lambda}, d+1 \le i \le H. \tag{A.11}$$

The above two statements give us the desried result in Theorem 3. □

A.3. Proof of Proposition 1

Proof of Lemma 1 By Lemma 10, one of main contributions of Lin et al. [2016], we know that Ω_1 happens with high probability follows. By the following lemma, we know that Ω_2 happens with high probability.

LEMMA 8. *Let $d_1, \ldots, d_p$ be positive constants. We have the following statements:*

i) For p i.i.d standard normal random variables $x_1, \ldots, x_p$, there exist constants C_1 and C_2 such that for any sufficiently small a , we have

$$\mathbb{P}\left(|\frac{1}{p}\sum_i d_i(x_i^2 - 1)| > a\right) \le C_1 \exp^{-\frac{p^2a^2}{C_2\sum_j d_j^2}}. \tag{A.12}$$

ii) For $2p$ i.i.d standard normal random variables $x_1, \ldots, x_p$ $y_1, \cdots, y_p$, there exist constants C_1 and C_2 such that for any sufficiently small a , we have

$$\mathbb{P}\left(|\frac{1}{p}\sum_i d_i x_i y_i| > a\right) \le C_1 \exp^{-\frac{p^2a^2}{C_2\sum_j d_j^2}}. \tag{A.13}$$

PROOF. *ii) is a direct corollary of i). And i) follows from a typical deviation argument, i.e., all we need to do is choosing proper t such that $e^{-pt}\prod_{j=1}^p \frac{1}{\sqrt{1-2td_j}}$ is bounded by the RHS of* (A.12)*. This could be done by simple Taylor expansion argument.* □

□

Proof of Lemma 2 The proof needs the Sin-Theta theroem, which we reviewed below:

LEMMA 9. *Let $\boldsymbol{A}$ and $\boldsymbol{A} + \boldsymbol{E}$ be symmetric matrices satisfying*

$$\boldsymbol{A} = [\boldsymbol{F}_0, \boldsymbol{F}_1]\begin{bmatrix} \boldsymbol{A}_0 & 0 \\ 0 & \boldsymbol{A}_1 \end{bmatrix}\begin{bmatrix} \boldsymbol{F}_0^\tau \\ \boldsymbol{F}_1^\tau \end{bmatrix}, \quad \boldsymbol{A} + \boldsymbol{E} = [\boldsymbol{G}_0, \boldsymbol{G}_1]\begin{bmatrix} \boldsymbol{\Lambda}_0 & 0 \\ 0 & \boldsymbol{\Lambda}_1 \end{bmatrix}\begin{bmatrix} \boldsymbol{G}_0^\tau \\ \boldsymbol{G}_1^\tau \end{bmatrix}$$

where $[\boldsymbol{F}_0, \boldsymbol{F}_1]$ *and* $[\boldsymbol{G}_0, \boldsymbol{G}_1]$ *are orthogonal matrices. If the eigenvalues of* $\boldsymbol{A}_0$ *are contained in an interval (a,b) , and the eigenvalues of* $\boldsymbol{\Lambda}_1$ *are excluded from the interval* $(a-\delta, b+\delta)$ *for some* $\delta > 0$*, then*

$$\frac{1}{\sqrt{2}}\|\boldsymbol{F}_0\boldsymbol{F}_0^\tau - \boldsymbol{G}_0\boldsymbol{G}_0^\tau\|_F \leq \frac{\min(\|\boldsymbol{F}_1^\tau \boldsymbol{E}\boldsymbol{G}_0\|_F, \|\boldsymbol{F}_0^\tau \boldsymbol{E}\boldsymbol{G}_1\|_F)}{\delta}.$$

Let us consider the eigen-decompositions: $Q_1 \triangleq \boldsymbol{M}^\tau\boldsymbol{M} = \begin{pmatrix} \boldsymbol{E}_1 & \boldsymbol{E}_2 \\ \boldsymbol{E}_3 & \boldsymbol{E}_4 \end{pmatrix}\begin{pmatrix} \boldsymbol{D}_1 & 0 \\ 0 & \boldsymbol{D}_2 \end{pmatrix}\begin{pmatrix} \boldsymbol{E}_1^\tau & \boldsymbol{E}_3^\tau \\ \boldsymbol{E}_2^\tau & \boldsymbol{E}_4^\tau \end{pmatrix}$ $Q_2 \triangleq \begin{pmatrix} B_1^\tau B_1 + \frac{p\mu}{n}\boldsymbol{I}_d & 0 \\ 0 & \frac{p\mu}{n}\boldsymbol{I}_{H-d} \end{pmatrix} = \begin{pmatrix} \boldsymbol{F}_1 & 0 \\ 0 & \boldsymbol{F}_2 \end{pmatrix}\begin{pmatrix} \boldsymbol{D}_1' & 0 \\ 0 & \boldsymbol{D}_2' \end{pmatrix}\begin{pmatrix} \boldsymbol{F}_1^\tau & 0 \\ 0 & \boldsymbol{F}_2^\tau \end{pmatrix}$ where$\boldsymbol{D}_1, \boldsymbol{D}_1'$ (*resp.* $\boldsymbol{D}_2, \boldsymbol{D}_2'$) are $d \times d$ (*resp.* $(H-d)\times(H-d)$) diagonal matrices. From (15), we know that

$$\lambda + \frac{p\mu}{n} - C\frac{\sqrt{p\log(p)}}{n} \leq \lambda_{\min}(D_1) \leq \lambda_{\max}(D_1) \leq \kappa\lambda + \frac{p\mu}{n} + C\frac{\sqrt{p\log(p)}}{n}$$
$$\text{and} \qquad \frac{p\mu}{n} - C\frac{\sqrt{p\log(p)}}{n} \leq \lambda_{\min}(D_2) \leq \lambda_{\max}(D_2) \leq \frac{p\mu}{n} + C\frac{\sqrt{p\log(p)}}{n}.$$

Thus the eigengap is of order $\lambda - \frac{\sqrt{p\log(p)}}{n}$ (which is of order λ, since $n\lambda = p^\alpha$ for some $\alpha > 1/2$). Let us apply the Sin-Theta theorem on them. From (15) , we know that $\|Q_1 - Q_2\|_F \leq C\frac{\sqrt{p\log(p)}}{n}$. Thus, we have

$$\left\| \begin{pmatrix} \boldsymbol{E}_1 \\ \boldsymbol{E}_3 \end{pmatrix}\begin{pmatrix} \boldsymbol{E}_1^\tau & \boldsymbol{E}_3^\tau \end{pmatrix} - \begin{pmatrix} \boldsymbol{I}_d & 0 \\ 0 & 0 \end{pmatrix} \right\|_F \leq C\frac{\sqrt{p\log(p)}}{n\lambda}, \tag{A.14}$$

i.e., $\|\boldsymbol{E}_3\boldsymbol{E}_3^\tau\|_F \leq C\frac{\sqrt{p\log(p)}}{n\lambda}$. Similar argument gives us $\|\boldsymbol{E}_2\boldsymbol{E}_2^\tau\|_F \leq C\frac{\sqrt{p\log(p)}}{n\lambda}$.

Let $\boldsymbol{\eta}$ be the (unit) eigenvector associated with the non-zero eigenvalue $\widehat{\lambda}$ of $\boldsymbol{M}\boldsymbol{M}^\tau$. Let us write $\boldsymbol{\eta} = \begin{pmatrix} \boldsymbol{\eta}_1 \\ \boldsymbol{\eta}_2 \\ \boldsymbol{\eta}_3 \end{pmatrix}$ where $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2 \in \mathbb{R}^d$ and $\boldsymbol{\eta}_3 \in \mathbb{R}^{p-2d}$. Let $\alpha = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = \boldsymbol{M}^\tau\boldsymbol{\eta}$, where $\alpha_1 = B_1^\tau\boldsymbol{\eta}_1 + B_2^\tau\boldsymbol{\eta}_2 \in \mathbb{R}^d$ and $\alpha_2 = B_3^\tau\boldsymbol{\eta}_2 + B_4^\tau\boldsymbol{\eta}_3 \in \mathbb{R}^{H-d}$. It is easy to verify that $\frac{\alpha}{\sqrt{\widehat{\lambda}}}$ is the (unit) eigenvector associted with the eigenvalue $\widehat{\lambda}$ of $\boldsymbol{M}^\tau\boldsymbol{M}$ and

$$\boldsymbol{\eta}_1 = \frac{B_1}{\sqrt{\widehat{\lambda}}}\frac{\alpha_1}{\sqrt{\widehat{\lambda}}}, \quad \boldsymbol{\eta}_2 = \frac{B_2}{\sqrt{\widehat{\lambda}}}\frac{\alpha_1}{\sqrt{\widehat{\lambda}}} + \frac{B_3}{\sqrt{\widehat{\lambda}}}\frac{\alpha_2}{\sqrt{\widehat{\lambda}}}, \text{ and } \boldsymbol{\eta}_3 = \frac{B_4}{\sqrt{\widehat{\lambda}}}\frac{\alpha_2}{\sqrt{\widehat{\lambda}}}.$$

If $\widehat{\lambda}$ is among the first d eigenvalues of $\boldsymbol{M}^\tau\boldsymbol{M}$, then $\|\alpha_1/\sqrt{\widehat{\lambda}}\|_2$ is bounded below by some positive constant. Thus $\|\boldsymbol{\eta}_1\|_2 \geq C\sqrt{\frac{\lambda}{\widehat{\lambda}}}$. If $\widehat{\lambda}$ is among the last $H-d$ eigenvalues of $\boldsymbol{M}^\tau\boldsymbol{M}$, then $\|\alpha_1/\sqrt{\widehat{\lambda}}\|_1 = O\left(\frac{\sqrt{p\log(p)}}{n\lambda}\right)$. Thus $\|\boldsymbol{\eta}_1\|_2 \leq O\left(\kappa\sqrt{\frac{\lambda}{\widehat{\lambda}}}\frac{\sqrt{p\log(p)}}{n\lambda}\right)$. □

B. Conditions for SIR to be Consistent

All the results in this appendix are borrowed from Lin et al. [2016] and Lin et al. [2015] with some minor modifications. We state them here mainly for the purpose of self-content. In order that SIR gives a consistent estimate of the central space, researchers (for more discussion, see Li [1991], Hsing and Carroll [1992], Zhu et al. [2006] and Lin et al. [2015]) require the several technical assumptions,

A′) Linearity Condition and Coverage Condition.

$$span\Big\{ \mathbb{E}[\boldsymbol{x}|y] \Big\} = span\Big\{ \boldsymbol{V}_{*,1}, ..., \boldsymbol{V}_{*,d} \Big\} \tag{B.1}$$

where $\boldsymbol{V}_{*,i}$ is the i-th columns of the orthogonal matrix $\boldsymbol{V}$ and a smoothness and tail condition on the central curve $\mathbb{E}[\boldsymbol{x}|y]$, i.e.,

B′) Smoothness and Tail Condition.

Smoothness condition: For $B > 0$ and $n \geq 1$, let $\Pi_n(B)$ be the collection of all the n-point partitions $-B \leq y_{(1)} \leq \cdots \leq y_{(n)} \leq B$ of $[-B, B]$. The central curve $\boldsymbol{m}(y)$ satisfies the following:

$$\lim_{n\to\infty} \sup_{y\in\Pi_n(B)} n^{-1/4} \sum_{i=2}^{n} \|\boldsymbol{m}(y_i) - \boldsymbol{m}(y_{i-1})\|_2 = 0, \forall B > 0.$$

Tail condition: for some $B_0 > 0$, there exists a non-decreasing function $\widetilde{m}(y)$ on (B_0, ∞), such that

$$\begin{gathered} \widetilde{m}^4(y)P(|Y| > y) \to 0 \text{ as } y \to \infty \\ \|\boldsymbol{m}(y) - \boldsymbol{m}(y')\|_2 \leq |\widetilde{m}(y) - \widetilde{m}(y')| \text{ for } y, y' \in (-\infty, -B_0) \cup (B_0, \infty) \end{gathered} \tag{B.2}$$

Recent results in Lin et al. [2015], Lin et al. [2016] and Neykov et al. [2016] suggest that the following sliced stability condition (see Definition 1) is more convenient to build the high dimensional theory of SIR.

Definition 1. *Let Y be a univariate random variable. For $0 < \boldsymbol{\gamma}_1 < 1 < \boldsymbol{\gamma}_2$, let $\mathcal{A}_H(\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2)$ denote all partitions $\{-\infty = a_1 \leq a_2 \leq \ldots \leq a_{H+1} = +\infty\}$ of $\mathbb{R}$, such that*

$$\frac{\boldsymbol{\gamma}_1}{H} \leq \mathbb{P}(a_h \leq Y \leq a_{h+1}) \leq \frac{\boldsymbol{\gamma}_2}{H}.$$

A curve $\boldsymbol{m}(y)$ is ϑ-sliced stable with respect to y, if there exist positive constants $\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \boldsymbol{\gamma}_3$ such that for any partition $\in \mathcal{A}_H(\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2)$ and any $\boldsymbol{\beta} \in \mathbb{R}^p$, one has:

$$\frac{1}{H} \sum_{h=1}^{H+1} var\left(\boldsymbol{\beta}^\tau \boldsymbol{m}(Y) | a_{h-1} \leq Y < a_h\right) \leq \frac{\boldsymbol{\gamma}_3}{H^\vartheta} var\left(\boldsymbol{\beta}^\tau \boldsymbol{m}(Y)\right). \tag{B.3}$$

A curve is sliced stable if it is ϑ-sliced stable for some positive constant ϑ.

Note that the LHS of (B.3) $\approx var(\mathbb{E}[\boldsymbol{\beta}^\tau \boldsymbol{m}(Y)|Y]) = 0$, so the sliced stability condition is merely requiring the convergence rate to be a power of the slice number and is believed to be a general condition. In fact, Neykov et al. [2016] shows that if we replace the tail condition (B.2) by a slightly stronger condition

$$\mathbb{E}[\widetilde{m}(Y)^4] < \infty, \tag{B.4}$$

then $\mathbf{B}'$) implies the sliced stability condition with $\vartheta = \frac{1}{2}$.

One of the main implication of the sliced stability condition is the following lemma (see Lin et al. [2015]), which plays the key role in their sequential works.

LEMMA 10. *Let $\boldsymbol{x} \in \mathbb{R}^p$ be a sub-Gaussian random variable. For any unit vector $\boldsymbol{\beta} \in \mathbb{R}^p$, let $\boldsymbol{x}(\boldsymbol{\beta}) = \langle \boldsymbol{x}, \boldsymbol{\beta} \rangle$ and $\boldsymbol{m}(\boldsymbol{\beta}) = \langle \boldsymbol{m}, \boldsymbol{\beta} \rangle = \mathbb{E}[\boldsymbol{x}(\boldsymbol{\beta}) \mid y]$, we have the following:*

i) If $var(\boldsymbol{m}(\boldsymbol{\beta})) = 0$, there exists positive constants C_1, C_2 and C_3 such that for any $b = O(1)$ and sufficiently large H, we have

$$\mathbb{P}(var_H(\boldsymbol{x}(\boldsymbol{\beta})) > b) \leq C_1 \exp\left(-C_2 \frac{nb}{H^2} + C_3 \log(H)\right).$$

ii) If $var(\boldsymbol{m}(\boldsymbol{\beta})) \neq 0$, there exists positive constants C_1, C_2 and C_3 such that, for any $\nu > 1$, we have

$$|var_H(\boldsymbol{x}(\boldsymbol{\beta})) - var(\boldsymbol{m}(\boldsymbol{\beta}))| \geq \frac{1}{2\nu} var(\boldsymbol{m}(\boldsymbol{\beta}))$$

with probability at most

$$C_1 \exp\left(-C_2 \frac{n\ var(\boldsymbol{m}(\boldsymbol{\beta}))}{H^2 \nu^2} + C_3 \log(H)\right).$$

where we choose H such that $H^\vartheta > C_4 \nu$ for some sufficiently large constant C_4.

We summarize serveral simple implications of this lemma into the following Proposition, the proof of which is simple and thus omited. For interest readers, we refer to Lin et al. [2016].

PROPOSITION 3. *There exist positive constants C_1,C_2 and C_3, such that*

$$\|\boldsymbol{\beta}^\tau \boldsymbol{\Lambda}_H \boldsymbol{\beta} - \boldsymbol{\beta}^\tau var(\mathbb{E}[\boldsymbol{x}|y])\boldsymbol{\beta}\|_2 \geq \frac{1}{2\nu} \boldsymbol{\beta}^\tau var(\mathbb{E}[\boldsymbol{x}|y])\boldsymbol{\beta} \tag{B.5}$$

with probability at most $C_1 \exp\left(-C_2 \frac{n\lambda}{H^2\nu^2} + C_3 \log(H)\right)$.

PROOF. *It follows from Lemma 10 and the fact that for any $\boldsymbol{\beta} \in col(var(\mathbb{E}[\boldsymbol{x}|y]))$, $var(\boldsymbol{m}(\boldsymbol{\beta}))) \geq \lambda$.* □

C. Results of simulations

Table C1. Estimation error: $\mathbf{\Sigma} = \mathbf{\Sigma}_1$, and $\rho = 0$.

	p	Lasso-SIR	DT-SIR	Lasso	M-Lasso	Lasso-SIR(Known d)	$\hat{d}$
I	100	0.09	0.21	0.08	0.09	0.09	1
	1000	0.12	0.21	0.10	0.22	0.12	1
	2000	0.14	0.22	0.10	0.29	0.14	1
	4000	0.18	0.22	0.11	0.39	0.18	1
II	100	0.05	0.29	0.23	0.05	0.05	1
	1000	0.09	0.35	0.30	0.12	0.09	1
	2000	0.12	0.38	0.31	0.18	0.11	1
	4000	0.15	0.41	0.33	0.27	0.15	1
III	100	0.17	0.23	1.14	0.20	0.18	1
	1000	0.27	0.30	1.25	0.63	0.23	1.1
	2000	0.35	0.34	1.31	0.77	0.26	1.1
	4000	0.45	0.42	1.29	0.93	0.34	1.3
IV	100	0.35	0.79	0.41	0.98	0.35	1
	1000	0.59	0.96	0.61	0.84	0.56	1.1
	2000	0.72	1.02	0.67	1.01	0.64	1.2
	4000	0.95	1.14	0.71	1.23	0.82	1.4
V	100	0.10	0.18	0.55	0.11	0.09	1
	1000	0.12	0.19	0.69	0.30	0.13	1
	2000	0.15	0.19	0.72	0.37	0.15	1
	4000	0.18	0.20	0.74	0.47	0.18	1

Table C2. Estimation error: $\mathbf{\Sigma} = \mathbf{\Sigma}_1$ and $\rho = 0.3$.

	p	Lasso-SIR	DT-SIR	Lasso	M-Lasso	Lasso-SIR(Known d)	$\hat{d}$
I	100	0.10	0.435	0.09	0.12	0.1	1
	1000	0.15	0.50	0.12	0.24	0.15	1
	2000	0.18	0.50	0.13	0.31	0.17	1
	4000	0.21	0.49	0.14	0.42	0.22	1
II	100	0.06	0.46	0.22	0.08	0.06	1
	1000	0.11	0.55	0.28	0.14	0.11	1
	2000	0.14	0.55	0.30	0.20	0.14	1
	4000	0.19	0.58	0.32	0.32	0.19	1
III	100	0.19	0.50	1.18	0.24	0.19	1
	1000	0.29	0.60	1.30	0.58	0.25	1
	2000	0.35	0.63	1.32	0.73	0.29	1.1
	4000	0.57	0.75	1.33	0.98	0.4	1.4
IV	100	0.38	0.85	0.56	0.54	0.37	1
	1000	0.56	0.97	0.70	0.77	0.54	1
	2000	0.65	1.03	0.76	0.92	0.58	1.1
	4000	0.79	1.12	0.79	1.09	0.65	1.3
V	100	0.10	0.47	0.48	0.14	0.1	1
	1000	0.14	0.55	0.60	0.35	0.15	1
	2000	0.17	0.56	0.66	0.44	0.18	1
	4000	0.30	0.60	0.72	0.66	0.25	1

Table C3. Estimation error: $\mathbf{\Sigma} = \mathbf{\Sigma}_1$, and $\rho = 0.8$.

	p	Lasso-SIR	DT-SIR	Lasso	M-Lasso	Lasso-SIR(Known d)	$\hat{d}$
I	100	0.18	1.34	0.16	1.01	0.18	1
	1000	0.24	1.38	0.22	0.79	0.24	1
	2000	0.27	1.39	0.23	0.53	0.27	1
	4000	0.32	1.39	0.25	0.45	0.32	1
II	100	0.10	1.34	0.33	1.17	0.11	1
	1000	0.16	1.39	0.55	1.08	0.16	1
	2000	0.19	1.39	0.71	0.92	0.19	1
	4000	0.23	1.40	0.92	0.54	0.23	1
III	100	0.28	1.34	1.26	1.00	0.28	1
	1000	0.45	1.38	1.29	0.92	0.44	1
	2000	0.54	1.39	1.30	0.84	0.54	1
	4000	0.76	1.43	1.29	0.89	0.68	1.1
IV	100	0.74	1.40	1.21	0.91	0.72	1
	1000	0.75	1.41	1.23	0.88	0.76	1
	2000	0.79	1.44	1.26	0.94	0.75	1.1
	4000	0.93	1.52	1.27	1.09	0.76	1.4
V	100	0.19	1.31	0.36	1.10	0.19	1
	1000	0.31	1.38	0.56	0.55	0.32	1
	2000	0.50	1.42	0.74	0.71	0.47	1.1
	4000	1.15	1.66	0.82	1.25	0.80	2.1

Table C4. Estimation error: $\mathbf{\Sigma} = \mathbf{\Sigma}_2$ and $\rho = 0.2$.

	p	Lasso-SIR	DT-SIR	Lasso	M-Lasso	Lasso-SIR(Known d)	$\hat{d}$
I	100	0.13	1.22	0.09	0.16	0.13	1
	1000	0.33	1.37	0.11	0.65	0.26	1.1
	2000	0.30	1.37	0.12	0.74	0.30	1
	4000	0.36	1.38	0.13	0.81	0.30	1.1
II	100	0.10	1.26	0.24	0.11	0.10	1
	1000	0.25	1.37	0.31	0.47	0.25	1
	2000	0.30	1.38	0.33	0.59	0.29	1
	4000	0.31	1.40	0.35	0.65	0.32	1
III	100	0.24	1.24	1.19	0.33	0.23	1
	1000	0.55	1.33	1.30	0.98	0.40	1.3
	2000	0.59	1.35	1.28	1.08	0.45	1.3
	4000	0.58	1.36	1.28	1.14	0.47	1.3
IV	100	0.54	1.37	1.19	0.60	0.54	1
	1000	0.63	1.41	1.25	0.87	0.63	1
	2000	0.64	1.41	1.27	0.99	0.65	1
	4000	0.65	1.41	1.26	1.07	0.66	1
V	100	0.23	1.23	0.49	0.29	0.13	1.1
	1000	1.03	1.10	0.61	1.15	0.79	1.6
	2000	1.09	0.96	0.67	1.20	0.86	1.6
	4000	1.07	0.99	0.71	1.21	0.87	1.7

Table C5. Estimation error: $\mathbf{\Sigma} = \mathbf{\Sigma}_1$ and $\rho = 0$.

	p	Lasso-SIR	DT-SIR	M-Lasso	Lasso-SIR(Known d)	$\hat{d}$
VI	100	0.15	0.18	0.23	0.14	2
	1000	0.18	0.17	0.61	0.17	2
	2000	0.22	0.20	0.72	0.20	2
	4000	0.28	0.20	0.86	0.27	2
VII	100	0.27	0.35	0.32	0.27	2
	1000	0.37	0.40	0.93	0.37	2
	2000	0.44	0.41	1.09	0.45	2
	4000	0.75	0.64	1.40	0.60	2.4
VIII	100	0.87	0.88	0.90	0.23	1.2
	1000	0.45	0.44	0.91	0.31	1.8
	2000	0.34	0.35	0.80	0.34	2
	4000	0.57	0.53	1.04	0.41	1.8
IX	100	0.87	0.89	0.91	0.26	1.2
	1000	0.60	0.54	1.10	0.39	1.7
	2000	0.78	0.71	1.18	0.59	1.6
	4000	0.96	0.83	1.25	0.83	1.5

Table C6. Estimation error: $\mathbf{\Sigma} = \mathbf{\Sigma}_1$ and $\rho = 0.3$.

	p	Lasso-SIR	DT-SIR	M-Lasso	Lasso-SIR(Known d)	$\hat{d}$
VI	100	0.20	0.34	0.25	0.20	2
	1000	0.24	0.30	0.61	0.23	2
	2000	0.26	0.36	0.71	0.26	2
	4000	0.31	0.41	0.84	0.29	2
VII	100	0.28	0.63	0.42	0.28	2
	1000	0.41	0.71	0.95	0.40	2
	2000	0.58	0.78	1.17	0.54	2.1
	4000	0.97	0.92	1.46	0.78	2.3
VIII	100	0.25	0.55	0.35	0.22	2
	1000	0.32	0.59	0.77	0.29	2
	2000	0.34	0.69	0.81	0.34	2
	4000	0.57	0.77	1.11	0.47	2.2
IX	100	0.31	0.50	0.43	0.31	2
	1000	0.35	0.47	0.99	0.35	2
	2000	0.42	0.55	1.17	0.40	2.1
	4000	0.51	0.56	1.28	0.44	2

Table C7. Estimation error: $\mathbf{\Sigma} = \mathbf{\Sigma}_1$ and $\rho = 0.8$.

	p	Lasso-SIR	DT-SIR	M-Lasso	Lasso-SIR(Known d)	$\hat{d}$
VI	100	0.52	1.86	1.01	0.51	2
	1000	0.79	1.92	0.93	0.79	2
	2000	0.96	1.94	1.05	0.94	2
	4000	1.14	2.01	1.26	1.06	2.2
VII	100	0.80	1.77	1.07	0.72	1.6
	1000	1.09	1.78	1.23	1.34	1.3
	2000	1.09	1.76	1.23	1.39	1.3
	4000	1.12	1.76	1.27	1.42	1.2
VIII	100	0.42	1.81	0.79	0.34	2
	1000	1.00	1.97	1.22	0.86	2.2
	2000	1.12	1.93	1.27	1.17	2.1
	4000	1.16	1.89	1.27	1.28	1.8
IX	100	0.78	1.90	0.95	0.79	2
	1000	0.92	1.95	1.08	0.90	2
	2000	0.97	1.97	1.17	0.95	2
	4000	1.12	2.03	1.37	1.01	2.3

Table C8. Estimation error: $\mathbf{\Sigma} = \mathbf{\Sigma}_2$ and $\rho = 0.2$.

	p	Lasso-SIR	DT-SIR	M-Lasso	Lasso-SIR(Known d)	$\hat{d}$
VI	100	0.27	1.73	0.43	0.21	1.9
	1000	1.01	1.73	1.11	0.26	1
	2000	1.01	1.73	1.14	0.29	1
	4000	1.02	1.73	1.18	0.38	1
VII	100	0.39	1.70	0.55	0.31	1.9
	1000	1.03	1.73	1.25	0.47	1
	2000	1.04	1.73	1.30	0.55	1
	4000	1.04	1.73	1.34	0.69	1
VIII	100	0.24	1.69	0.34	0.24	2
	1000	0.97	1.73	1.15	0.33	1.1
	2000	1.03	1.74	1.24	0.35	1
	4000	1.03	1.74	1.26	0.41	1
IX	100	1.00	1.69	1.04	0.40	1
	1000	1.03	1.73	1.22	0.67	1
	2000	1.03	1.73	1.27	0.73	1
	4000	1.04	1.73	1.30	0.89	1

Table C9. Estimation error: $\mathbf{\Sigma} = \mathbf{\Sigma}_1$ and $\rho = 0$.

	p	Lasso-SIR	DT-SIR	M-Lasso	Lasso
X	100	0.19	0.23	0.22	0.18
	1000	0.22	1.13	0.60	0.25
	2000	0.23	1.24	0.67	0.27
	4000	0.24	1.29	0.71	0.29
XI	100	0.36	0.82	0.41	0.37
	1000	0.44	1.29	1.16	0.43
	2000	0.41	1.35	1.22	0.48
	4000	0.45	1.37	1.23	0.47
XII	100	0.23	0.53	0.25	0.21
	1000	0.30	1.12	0.62	0.30
	2000	0.32	1.23	0.71	0.34
	4000	0.34	1.29	0.75	0.36
XIII	100	0.36	0.92	0.43	1.05
	1000	0.44	1.69	1.22	1.08
	2000	0.44	1.80	1.30	1.09
	4000	0.43	1.85	1.33	1.10

Table C10. Estimation error: $\mathbf{\Sigma} = \mathbf{\Sigma}_1$ and $\rho = 0.3$.

	p	Lasso-SIR	DT-SIR	M-Lasso	Lasso
X	100	0.20	0.50	0.26	0.19
	1000	0.24	1.15	0.56	0.26
	2000	0.25	1.25	0.63	0.28
	4000	0.26	1.31	0.68	0.30
XI	100	0.34	0.86	0.57	0.34
	1000	0.43	1.31	1.13	0.43
	2000	0.42	1.36	1.20	0.45
	4000	0.45	1.38	1.22	0.48
XII	100	0.22	0.54	0.28	0.21
	1000	0.29	1.10	0.57	0.30
	2000	0.31	1.22	0.66	0.34
	4000	0.33	1.29	0.71	0.36
XIII	100	0.38	0.99	0.61	1.06
	1000	0.39	1.73	1.19	1.08
	2000	0.41	1.84	1.29	1.09
	4000	0.42	1.88	1.33	1.10

Table C11. Estimation error: $\mathbf{\Sigma} = \mathbf{\Sigma}_1$ and $\rho = 0.8$.

	p	Lasso-SIR	DT-SIR	M-Lasso	Lasso
X	100	0.28	1.36	0.99	0.28
	1000	0.45	1.41	0.90	0.39
	2000	0.55	1.41	0.82	0.46
	4000	0.65	1.41	0.77	0.54
XI	100	0.40	1.38	0.84	0.39
	1000	0.71	1.41	1.04	0.69
	2000	0.89	1.41	1.11	0.88
	4000	1.02	1.4	1.17	1.02
XII	100	0.36	1.37	1.17	0.30
	1000	0.63	1.41	1.16	0.52
	2000	0.83	1.41	1.14	0.71
	4000	1.03	1.41	1.09	0.91
XIII	100	0.48	1.92	0.89	1.10
	1000	0.56	1.99	1.14	1.12
	2000	0.59	2.00	1.21	1.14
	4000	0.61	2.00	1.31	1.15

Table C12. Estimation error: $\mathbf{\Sigma} = \mathbf{\Sigma}_2$ and $\rho = 0.2$.

	p	Lasso-SIR	DT-SIR	M-Lasso	Lasso
X	100	0.19	1.24	0.28	0.19
	1000	0.23	1.41	0.61	0.26
	2000	0.24	1.41	0.68	0.28
	4000	0.26	1.41	0.73	0.31
XI	100	0.36	1.29	0.58	0.35
	1000	0.42	1.41	1.16	0.43
	2000	0.46	1.41	1.22	0.48
	4000	0.46	1.41	1.24	0.49
XII	100	0.24	1.25	0.31	0.22
	1000	0.32	1.40	0.62	0.32
	2000	0.34	1.41	0.71	0.34
	4000	0.35	1.41	0.77	0.38
XIII	100	0.40	1.72	0.67	1.06
	1000	0.44	1.99	1.22	1.09
	2000	0.46	2.00	1.31	1.10
	4000	0.45	2.00	1.37	1.11