



Sparse Sliced Inverse Regression via Lasso

Qian Lin, Zhigen Zhao & Jun S. Liu

To cite this article: Qian Lin, Zhigen Zhao & Jun S. Liu (2019) Sparse Sliced Inverse Regression via Lasso, Journal of the American Statistical Association, 114:528, 1726-1739, DOI: [10.1080/01621459.2018.1520115](https://doi.org/10.1080/01621459.2018.1520115)

To link to this article: <https://doi.org/10.1080/01621459.2018.1520115>



View supplementary material [↗](#)



Accepted author version posted online: 11 Sep 2018.
Published online: 09 Mar 2019.



Submit your article to this journal [↗](#)



Article views: 1433



View related articles [↗](#)



View Crossmark data [↗](#)



Citing articles: 1 View citing articles [↗](#)



Sparse Sliced Inverse Regression via Lasso

Qian Lin^a, Zhigen Zhao^{*b}, and Jun S. Liu^{a,c}

^aCenter for Statistical Science and Department of Industrial Engineering, Tsinghua University, Beijing, China; ^bDepartment of Statistical Science, Temple University, Philadelphia, PA; ^cDepartment of Statistics, Harvard University, Cambridge, MA

ABSTRACT

For multiple index models, it has recently been shown that the sliced inverse regression (SIR) is consistent for estimating the sufficient dimension reduction (SDR) space if and only if $\rho = \lim_{n \rightarrow \infty} \frac{p}{n} = 0$, where p is the dimension and n is the sample size. Thus, when p is of the same or a higher order of n , additional assumptions such as sparsity must 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, we introduce a simple Lasso regression method to obtain an estimate of the SDR space. The resulting algorithm, Lasso-SIR, is shown to be consistent and achieves the optimal convergence rate under certain sparsity conditions when p is of order $o(n^2\lambda^2)$, where λ is the generalized signal-to-noise ratio. We also demonstrate the superior performance of Lasso-SIR compared with existing approaches via extensive numerical studies and several real data examples. Supplementary materials for this article are available online.

ARTICLE HISTORY

Received January 2017
Revised November 2017

KEYWORDS

Dimension reduction; High dimensional statistics; Minimax; Theory of large deviation.

1. Introduction

Dimension reduction and variable selection have become indispensable steps for modern-day data analysts in dealing with the “big data,” where thousands or even millions of features are often available for only hundreds or thousands of samples. With these ultrahigh-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. One can consider the following *multiple index model* (Li 1991):

$$y = f(\beta_1^T x, \beta_2^T x, \dots, \beta_d^T x, \epsilon), \quad (1)$$

where x follows a p -dimensional elliptical distribution with mean zero and covariance matrix Σ , 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 x and has mean 0. When p is very large, it is reasonable to further restrict each β_i to be a sparse vector.

Since the introduction of the sliced inverse regression (SIR) method (Li 1991), many methods have been proposed to estimate the space spanned by $(\beta_1, \dots, \beta_d)$ with few assumptions on the link function $f(\cdot)$. Assume the multiple index model (1), the objective of all the sufficient dimension reduction (SDR, Cook 1998) methods is to find the minimal subspace $\mathcal{S} \subseteq \mathbb{R}^p$ such that $y \perp x \mid P_{\mathcal{S}}x$, where $P_{\mathcal{S}}$ stands for the projection operator to the subspace \mathcal{S} . When the dimension of x is moderately large, all the SDR methods, including SIR, are proven to be successful (Xia et al. 2002; Ni, Cook, and Tsai 2005; Li and Nachtsheim 2006; Li 2007; Zhu, Miao, and Peng 2006).


However, these methods were previously known to work well when the sample size n grows much faster than the dimension p , an assumption that becomes inappropriate for many modern-day datasets, such as those from biomedical researches. It is important to have a thorough investigation of “the behavior of these SDR estimators when n is not large relative to p ,” as raised by Cook, Forzani, and Rothman (2012).

Lin et al. (2018) 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 $\rho_n \stackrel{\text{def}}{=} p/n$ 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 greater than n , certain constraints must be imposed in order for SIR to be consistent. The sparsity assumption, that is, 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, Lin, and Liu (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 s is the number of active covariates and λ is the smallest nonzero eigenvalue of $\text{var}(\mathbb{E}[x|y])$. They further showed that the diagonal thresholding (DT) algorithm proposed in Lin et al. (2018) achieves the optimal rate for the single index model with the identity covariance matrix.

CONTACT Jun S. Liu  jliu@stat.harvard.edu  Department of Statistics, Harvard University, 1 Oxford Street, Cambridge, MA 02138

*Co-first author.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/JASA.

 Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

© 2019 American Statistical Association

1.1. 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 Σ . Consider first the single index model: $y = f(\beta^\tau x, \epsilon)$. Let η be the eigenvector associated with the largest eigenvalue of $\text{var}(\mathbb{E}[x|y])$. Since $\beta \propto \Sigma^{-1}\eta$, there are two immediate ways to estimate the space spanned by β . The first approach, as discussed in Lin et al. (2018), estimates Σ^{-1} and η separately (see Algorithm 1). The second one avoids a direct estimation of Σ^{-1} by solving the following penalized least square problem: $\|\frac{1}{n}XX^\tau\beta - \eta\|_2^2 + \mu\|\beta\|_1$, where X is the $p \times n$ covariate matrix formed by the n samples (see Algorithm 2). However, similar to most L_1 -penalization methods for nonlinear models, theoretical underpinning of this approach has not been well understood. Since these two approaches provide good estimates compared with earlier approaches (e.g., Li 1991; Li and Nachtsheim 2006; Li 2007) as shown in Lin et al. (2018) and supplementary materials, we set the two approaches as benchmarks for comparisons.

We note that an eigenvector $\hat{\eta}$ of $\widehat{\text{var}}(\mathbb{E}[x|y])$, where $\widehat{\text{var}}(\mathbb{E}[x|y])$ is an estimate of the conditional covariance matrix $\text{var}(\mathbb{E}[x|y])$ using SIR (Li 1991), must be a linear combination of the column vectors of X . Thus, we can construct an artificial response vector $\tilde{y} \in \mathbb{R}^n$ such that $\hat{\eta} = \frac{1}{n}X\tilde{y}$, and estimate β by solving another penalized least square problem: $\frac{1}{2n}\|\tilde{y} - X^\tau\beta\|_2^2 + \mu\|\beta\|_1$ (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.

1.2. 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 $y = x^\tau\beta + \epsilon$ for HDLSS data. By imposing the L_1 penalty on the regression coefficients, the Lasso approach (Tibshirani 1996) produces a sparse estimator of β , which turns out to be rate optimal (Raskutti, Wainwright, and Yu 2011). Because of apparent limitations of linear models, there are many attempts to build flexible and computationally friendly semiparametric models, such as the projection pursuit regression (Friedman and Stuetzle 1981; Chen 1991), sliced inverse regression (Li 1991), 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, Wainwright, and Yu 2011) and submatrix detections (Butucea and Ingster

2013), and are not applicable to index models. In this article, we provide a new framework for the theoretical investigation of regularized SDR methods for HDLSS data.

The rest of the article 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. Notations

We adopt the following notations throughout this article. For a matrix V , we call the space generated by its column vectors the column space and denote it by $\text{col}(V)$. The i th row and j th column of the matrix are denoted by $V_{i,*}$ and $V_{*,j}$, respectively. For (column) vectors x and $\beta \in \mathbb{R}^p$, we denote their inner product $\langle x, \beta \rangle$ by $x(\beta)$, and the k th entry of x by $x(k)$. 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 $a_n \geq Cb_n$ 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 as $\|A\|_{1,\infty} = \max_{1 \leq j \leq p} \sum_{i=1}^n |A_{i,j}|$ 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 B formed by the vectors β_1, \dots, β_d is not identifiable. However, **col(B), the space spanned by the columns of B is uniquely defined.** Given n iid samples (y_i, x_i) , $i = 1, \dots, n$, SIR (Li 1991) first divides the data into H equal-sized slices according to the order statistics $y_{(i)}$, $i = 1, \dots, n$. To ease notations and arguments, we assume that $n = cH$ and $\mathbb{E}[x] = 0$, and re-express the data as $y_{h,j}$ and $x_{h,j}$, where h refers to the slice number and j refers to the order number of a sample in the h th slice, that is, $y_{h,j} = y_{(c(h-1)+j)}$, $x_{h,j} = x_{(c(h-1)+j)}$. Here $x_{(k)}$ is the concomitant of $y_{(k)}$. Let the sample mean in the h th slice be denoted by \bar{x}_h , then $\Lambda \triangleq \text{var}(\mathbb{E}[x|y])$ can be estimated by

$$\hat{\Lambda}_H = \frac{1}{H} \sum_{h=1}^H \bar{x}_h \bar{x}_h^\tau = \frac{1}{H} X_H X_H^\tau, \quad (2)$$

where X_H is a $p \times H$ matrix formed by the H sample means, that is, $X_H = (\bar{x}_1, \dots, \bar{x}_H)$. Thus, $\text{col}(\Lambda)$ is estimated by $\text{col}(\hat{V}_H)$, where \hat{V}_H is the matrix formed by the top d eigenvectors of $\hat{\Lambda}_H$. The $\text{col}(\hat{V}_H)$ was shown to be a consistent estimator of $\text{col}(\Lambda)$ under a few technical conditions when p is fixed (Duan and

Li 1991; Hsing and Carroll 1992; Zhu, Miao, and Peng 2006; Li 1991; Lin et al. 2018), which are summarized in the online supplementary file. Recently, Lin et al. (2016, 2018) showed that $\text{col}(\widehat{\mathbf{V}}_H)$ is consistent for $\text{col}(\mathbf{\Lambda})$ if and only if $\rho_n = \frac{p}{n} \rightarrow 0$ as $n \rightarrow \infty$, when 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 the distribution of \mathbf{x} s is elliptically symmetric, Li (1991) showed that

$$\Sigma \text{col}(\mathbf{B}) = \text{col}(\mathbf{\Lambda}), \quad (3)$$

and thus our goal is to recover $\text{col}(\mathbf{B})$ by solving the above equation. It is shown in Lin et al. (2018) that when $\rho_n \rightarrow 0$, $\widehat{\text{col}}(\mathbf{B}) = \widehat{\Sigma}^{-1} \text{col}(\widehat{\mathbf{V}}_H)$ consistently estimate $\text{col}(\mathbf{B})$ where $\widehat{\Sigma} = \frac{1}{n} \mathbf{X} \mathbf{X}^\tau$ is the sample covariance matrix of \mathbf{X} . However, this simple approach breaks down when $\rho_n \not\rightarrow 0$, 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 nonstepwise methods for HDLSS scenarios, which will be used as benchmarks in our simulation studies to measure the performance of newly proposed SDR algorithms.

2.3. Diagonal Thresholding SIR

When $p \gg n$, the DT screening method (Lin et al. 2018) proceeds by marginally screening all the variables via the diagonal elements of $\widehat{\mathbf{\Lambda}}_H$ and then applying SIR to those retained variables to obtain an estimate of $\text{col}(\mathbf{B})$. The procedure is shown to be consistent if the number of nonzero entries in each row of Σ is bounded.

Algorithm 1 (DT-SIR)

- 1: Use the magnitudes of the diagonal elements of $\widehat{\mathbf{\Lambda}}_H$ to select the set of important predictors \mathcal{I} , with $|\mathcal{I}| = o(n)$;
 - 2: Apply SIR to the data $(\mathbf{y}, \mathbf{x}_{\mathcal{I}})$ 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.
-

2.4. 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, a reasonable estimate of $\text{col}(\mathbf{B})$ can be obtained by solving a sample-version of the equation with an appropriate regularization term to cope with the high dimensionality. Let $\widehat{\boldsymbol{\eta}}_1, \dots, \widehat{\boldsymbol{\eta}}_d$ be the eigenvectors associated with the largest d eigenvalues of $\widehat{\mathbf{\Lambda}}_H$. Replacing Σ by its sample version $\frac{1}{n} \mathbf{X} \mathbf{X}^\tau$ and imposing an L_1 penalty, we obtain a penalized sample version of (3):

$$\left\| \frac{1}{n} \mathbf{X} \mathbf{X}^\tau \boldsymbol{\beta} - \widehat{\boldsymbol{\eta}}_i \right\|_2^2 + \mu_i \|\boldsymbol{\beta}\|_1 \quad (4)$$

for some appropriate μ_i 's.

Algorithm 2 (Matrix Lasso)

- 1: Let $\widehat{\boldsymbol{\eta}}_1, \dots, \widehat{\boldsymbol{\eta}}_d$ be the eigenvectors associated with the largest d eigenvalues of $\widehat{\mathbf{\Lambda}}_H$;
 - 2: For $1 \leq i \leq d$, let $\widehat{\boldsymbol{\beta}}_i$ be the minimizer of Equation (4);
 - 3: Estimate the central space $\text{col}(\mathbf{B})$ by $\text{col}(\widehat{\boldsymbol{\beta}}_1, \dots, \widehat{\boldsymbol{\beta}}_d)$.
-

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.

2.5. The Lasso-SIR Algorithm

First consider the single index model

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

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

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

Thus, by defining

$$\widetilde{\mathbf{y}} = \frac{1}{c\widehat{\lambda}} \mathbf{M} \mathbf{M}^\tau \mathbf{X}^\tau \widehat{\boldsymbol{\eta}} \quad (6)$$

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

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

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

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

As shown by Bickel, Ritov, and Tsybakov (2009), the Dantzig selector is asymptotically equivalent to the Lasso for linear regressions. We thus propose and study the Lasso-SIR algorithm in this article.

Algorithm 3 (Lasso-SIR-1: for single index models)

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

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

where $\mu = C\sqrt{\frac{\log(p)}{n\lambda}}$ for sufficiently large constant C ;

- 3: Estimate P_β by $P_{\hat{\beta}(\mu)}$.

There is no need to estimate the inverse of Σ in Lasso-SIR. Moreover, since the optimization problem (7) is well-studied for linear regression models (Tibshirani 1996; Efron et al. 2004; Friedman, Hastie, and Tibshirani 2010), we may formally “transplant” their results to the index models. Practically, we use the R package *glmnet* to solve the optimization problem where the tuning parameter μ is chosen using cross-validation.

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

$$\tilde{\mathbf{Y}} = \frac{1}{c} \mathbf{M}\mathbf{M}^\tau \mathbf{X}^\tau \hat{\eta} \text{diag}\left(\frac{1}{\hat{\lambda}_1}, \dots, \frac{1}{\hat{\lambda}_d}\right). \quad (8)$$

We then apply the Lasso on each column of the pseudo response matrix to produce the corresponding estimate.

Algorithm 4 (Lasso-SIR: for multiple index models)

- 1: Let $\hat{\lambda}_i$ and $\hat{\eta}_i, i = 1, \dots, d$ be the top d eigenvalues and eigenvectors of $\hat{\Lambda}_H$, respectively;
- 2: Let $\tilde{\mathbf{Y}} = \frac{1}{c} \mathbf{M}\mathbf{M}^\tau \mathbf{X}^\tau \hat{\eta} \text{diag}\left(\frac{1}{\hat{\lambda}_1}, \dots, \frac{1}{\hat{\lambda}_d}\right)$. For each $1 \leq i \leq d$, solve the Lasso optimization problem

$$\hat{\beta}_i = \arg \min \mathcal{L}_{\beta,i} \text{ where } \mathcal{L}_{\beta,i} = \frac{1}{2n} \|\tilde{\mathbf{Y}}_{*,i} - \mathbf{X}^\tau \beta\|_2^2 + \mu_i \|\beta\|_1,$$

where $\mu_i = C\sqrt{\frac{\log(p)}{n\lambda_i}}$ for sufficiently large constant C ;

- 3: Let $\hat{\mathbf{B}}$ be the matrix formed by $\hat{\beta}_1, \dots, \hat{\beta}_d$. The estimate of P_B is given by $P_{\hat{\mathbf{B}}}$.

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 $\hat{\Lambda}_H$, 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 $\hat{\lambda}_i^a$ is the estimated value of d .

Remark 1. In another article that the authors are preparing, it is shown that the Lasso-SIR algorithm works on the joint distribution of (\mathbf{X}, Y) and is thus not tied to the single or multiple index models. We choose the single/multiple index models to have a clear representation of the central subspace \mathcal{S} , that is, $\mathcal{S} = \text{span}\{\beta_1, \dots, \beta_d\}$.

Remark 2. When dealing with real data, we suggest that the users employ quantile normalization to transform each covariate when \mathbf{X} is not normally distributed. When p is too large and beyond our bound of $n = O(\sqrt{p})$, as required by our provided R-package (see the “Acknowledgments” section for its downloading information), the user can first conduct variable screening based on DT-SIR, which is also included in this package.

3. Consistency of Lasso-SIR

For simplicity, we assume that $\mathbf{x} \sim N(0, \Sigma)$. 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 article. **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 a large enough but finite integer** (Lin et al. 2016, 2018). In order to prove the consistency, we need the following technical conditions:

- (Q1) There exist constants C_{\min} and C_{\max} such that $0 < C_{\min} < \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) < C_{\max}$;
- (Q2) There exists a constant $\kappa \geq 1$, such that

$$0 < \lambda = \lambda_d(\text{var}(\mathbb{E}[\mathbf{x}|y])) \leq \dots \leq \lambda_1(\text{var}(\mathbb{E}[\mathbf{x}|y])) \leq \kappa \lambda \leq \lambda_{\max}(\Sigma);$$

- (Q3) The central curve $\mathbf{m}(y) = \mathbb{E}[\mathbf{x}|y]$ satisfies the sliced stability condition.

Condition A1 is commonly imposed in the analyses of high-dimensional linear regression models. Condition A2 is merely a refinement of the coverage condition that is commonly imposed in the SIR literature, that is, $\text{rank}(\text{var}(\mathbb{E}[\mathbf{x}|y])) = d$. For single index models, there is a more intuitive explanation of condition A2. Since $\text{rank}(\text{var}(\mathbb{E}[\mathbf{x}|y])) = 1$, condition A2 is simplified to $0 < \lambda = \lambda_1 \leq \lambda_{\max}(\Sigma)$ which is a direct corollary of the total variance decomposition identity (i.e., $\text{var}(\mathbf{x}) = \text{var}(\mathbb{E}[\mathbf{x}|y]) + \mathbb{E}[\text{var}(\mathbf{x}|y)]$). We may treat λ as a generalized signal-to-noise ratio (SNR) and A2 simply requires that the generalized SNR is nonzero. Condition A3 is a property of the central curve, or equivalently, a regularity condition on the link function $f(\cdot)$ and the noise ϵ introduced in Lin et al. (2018).

Remark 3 (Generalized SNR and eigenvalue bound). Recall that the SNR for the linear model $y = \beta^\tau \mathbf{x} + \epsilon$, where $\mathbf{x} \sim N(0, \Sigma)$ and $\epsilon \sim N(0, 1)$, is defined as

$$\text{SNR} = \frac{E[(\beta^\tau \mathbf{x})^2]}{E[y^2]} = \frac{\|\beta\|_2^2 \beta_0^\tau \Sigma \beta_0}{1 + \|\beta\|_2^2 \beta_0^\tau \Sigma \beta_0},$$

where $\beta_0 = \beta / \|\beta\|_2$. A simple calculation shows that

$$\begin{aligned} \text{var}(\mathbb{E}[\mathbf{x}|y]) &= \frac{\Sigma \beta \beta^\tau \Sigma}{\beta_0^\tau \Sigma \beta_0 \|\beta\|_2^2 + 1} \\ \text{and } \lambda(\text{var}(\mathbb{E}[\mathbf{x}|y])) &= \frac{\beta_0^\tau \Sigma \Sigma \beta_0 \|\beta\|_2^2}{\beta_0^\tau \Sigma \beta_0 \|\beta\|_2^2 + 1}, \end{aligned}$$

where $\lambda(\text{var}(\mathbb{E}[\mathbf{x}|y]))$ is the unique nonzero eigenvalue of $\text{var}(\mathbb{E}[\mathbf{x}|y])$. This leads to the following identity for the linear model:

$$\lambda(\text{var}(\mathbb{E}[\mathbf{x}|y])) = \frac{\beta_0^\tau \Sigma \Sigma \beta_0}{\beta_0^\tau \Sigma \beta_0} \text{SNR}.$$

Thus, in a multiple index model we call λ , the smallest nonzero eigenvalue of $\text{var}(\mathbb{E}[\mathbf{x}|y])$, the model's generalized SNR.

Theorem 1 (Consistency of Lasso-SIR for Single Index Models).

Assume that $n\lambda = p^\alpha$ for some $\alpha > 1/2$ and that conditions A1–A3 hold for the single index model, $y = f(\beta_0^\tau \mathbf{x}, \epsilon)$, where β_0 is a unit vector. Let $\hat{\beta}(\mu)$ be the output of Algorithm 3, then

$$\|\hat{P}_{\hat{\beta}} - P_{\beta_0}\|_F \leq C_1 \sqrt{\frac{s \log(p)}{n\lambda}}$$

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

When no sparsity on η is assumed, the condition $\alpha > 1/2$ is necessary. This condition can be relaxed if a certain sparsity structure is imposed on η or Σ such that $\Sigma \beta$ becomes sparse. Next, we state the theoretical result regarding the multiple index model (1).

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

$$\|\hat{P}_{\hat{B}} - P_B\|_F \leq C_1 \sqrt{\frac{s \log(p)}{n\lambda}}$$

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) showed that the lower bound of the risk $\mathbb{E}\|\hat{P}_{\hat{B}} - P_B\|_F^2$ is $\frac{s \log(p/s)}{n\lambda}$ when (i) $d = 1$, or (ii) $d(> 1)$ is finite and $\lambda > c_0 > 0$. This implies that if $s = O(p^{1-\delta})$ for some positive constant δ , the Lasso-SIR algorithm achieves the optimal rate, that is, we have the following corollary.

Corollary 1. Assume that conditions A1–A3 hold. If $n\lambda = p^\alpha$ for some $\alpha > 1/2$ and $s = O(p^{1-\delta})$, then Lasso-SIR estimate $\hat{P}_{\hat{B}}$ achieves the minimax rate when (i) $d = 1$, or (ii) $d(> 1)$ is finite and $\lambda > c_0 > 0$.

Remark 4. Consider the linear regression $y = \beta^\tau \mathbf{x} + \epsilon$, where $\mathbf{x} \sim N(0, \Sigma)$, $\epsilon \sim N(0, 1)$. It is shown in Raskutti, Wainwright, and Yu (2011) that the lower bound of the minimax rate of the l_2 distance between any estimator and the true β is $\frac{s \log(p/s)}{n}$ and the convergence rate of Lasso estimator $\hat{\beta}_{\text{Lasso}}$ 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 $\lambda(\text{var}(\mathbb{E}[\mathbf{x}|y])) \sim \|\beta\|_2^2$, if $\|\beta\|_2$ is bounded away from ∞ . Consequently,

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

holds with high probability. In other words, if we treat Lasso as a dimension reduction method (where $d = 1$ and the link function is linear), the projection matrix $\hat{P}_{\hat{\beta}_{\text{Lasso}}}$ based on Lasso is rate optimal. The Lasso-SIR has extended the Lasso to the nonlinear multiple index models. This justifies a statement 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 challenging 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 $\hat{\lambda}_i$ of $\hat{\Lambda}_H$, $i = 1, \dots, H$, we face the problem that all these $\hat{\lambda}_i$'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 $\hat{\lambda}_i^a = \hat{\lambda}_i \|\beta_i\|_2$, $i = 1, \dots, H$. To this end, we have the following theorem.

Theorem 3. Let $\hat{\beta}_i$ be the output of Algorithm 4 for $i = 1, \dots, H$. 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 ,

$$\begin{aligned} \hat{\lambda}_i^a &\geq C_1 \sqrt{\lambda} - C_2 \sqrt{\frac{s \log(p)}{n}}, \text{ for } 1 \leq i \leq d, \text{ and} \\ \hat{\lambda}_i^a &\leq C_3 \sqrt{\frac{p \log(p)}{n\lambda}} \sqrt{\lambda} + C_4 \sqrt{\frac{s \log(p)}{n}}, \text{ for } d+1 \leq i \leq H, \end{aligned}$$

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

Theorem 3 states that, if $s \log(p) \vee (p \log(p))^{1/2} = o(n\lambda)$, there is a clear gap between signals and noise. The Lasso-SIR algorithm then provides us the rate optimal estimation of the central space. It can be easily verified that $p^{1/2}$ dominants $s \log(p)$ if $s < p^{1/2}$ and $s \log(p)$ dominants $p^{1/2}$ if $s > p^{1/2}$. The region $s^2 = o(p)$ and the region $p = o(s^2)$ are often referred to as the “highly sparse” and “moderately sparse” regions (Ingster,

Tsybakov, and Verzelen 2010), respectively. These two scenarios should be treated differently in high-dimensional SIR and SDR frameworks, just like what has been done in high-dimensional linear regression (Ingster, Tsybakov, and Verzelen 2010).

4. Simulation Studies

4.1. Single Index Models

Let β 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 \mathbf{x} be the design matrix with each row following $N(0, \Sigma)$. We consider two types of covariance matrices: (i) $\Sigma = (\sigma_{ij})$ where $\sigma_{ii} = 1$ and $\sigma_{ij} = \rho^{|i-j|}$; and (ii) $\sigma_{ii} = 1, \sigma_{ij} = \rho$ when $i, j \in \mathcal{S}$ or $i, j \in \mathcal{S}^c$, and $\sigma_{ij} = 0.1$ 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 $n = 1000$ and let p vary among 100, 1000, 2000, and 4000. For all the settings, the random error ϵ follows $N(0, \mathbf{I}_n)$. For single index models, we consider the following model settings:

- (I) $y = \mathbf{x}\beta + \epsilon$ where $\mathcal{S} = \{1, 2, \dots, 10\}$;
- (II) $y = (\mathbf{x}\beta)^3/2 + \epsilon$ where $\mathcal{S} = \{1, 2, \dots, 20\}$;
- (III) $y = \sin(\mathbf{x}\beta) * \exp(\mathbf{x}\beta) + \epsilon$ where $\mathcal{S} = \{1, 2, \dots, 10\}$;
- (IV) $y = \exp(\mathbf{x}\beta/10) + \epsilon$ where $\mathcal{S} = \{1, 2, \dots, 50\}$;
- (V) $y = \exp(\mathbf{x}\beta + \epsilon)$ where $\mathcal{S} = \{1, 2, \dots, 7\}$.

The goal is to estimate $\text{col}(\beta)$, the space spanned by β . As in Lin et al. (2018), the estimation error is defined as $\mathcal{D}(\widehat{\text{col}(\beta)}, \text{col}(\beta))$, 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 \hat{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 simulation results in Table 1 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 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 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. We observed that Lasso performed better than Lasso-SIR for Setting V when $\rho=0.8$ (supplementary materials) or when 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 known $d = 1$, Lasso-SIR's estimation error is smaller than that of Lasso.

The results, reported in the supplementary material, for the other values of ρ are similar to what we observed when $\rho = 0.5$.

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 (0.02)	0.47 (0.11)	0.11 (0.02)	0.19 (0.08)	0.12 (0.02)	1
	1000	0.18 (0.02)	0.65 (0.14)	0.15 (0.02)	0.26 (0.02)	0.18 (0.02)	1
	2000	0.2 (0.02)	0.74 (0.15)	0.16 (0.02)	0.3 (0.03)	0.2 (0.02)	1
	4000	0.23 (0.09)	0.9 (0.17)	0.18 (0.01)	0.39 (0.09)	0.23 (0.03)	1
II	100	0.07 (0.01)	0.6 (0.1)	0.23 (0.03)	0.27 (0.31)	0.07 (0.01)	1
	1000	0.12 (0.02)	0.78 (0.11)	0.31 (0.04)	0.17 (0.02)	0.12 (0.02)	1
	2000	0.15 (0.02)	0.86 (0.13)	0.34 (0.05)	0.2 (0.03)	0.15 (0.02)	1
	4000	0.2 (0.04)	0.99 (0.15)	0.37 (0.05)	0.28 (0.06)	0.19 (0.03)	1
III	100	0.21 (0.03)	0.55 (0.12)	1.25 (0.19)	0.26 (0.11)	0.21 (0.03)	1
	1000	0.28 (0.04)	0.74 (0.14)	1.32 (0.18)	0.51 (0.04)	0.27 (0.04)	1
	2000	0.35 (0.17)	0.87 (0.17)	1.34 (0.14)	0.66 (0.14)	0.31 (0.05)	1.1
	4000	0.46 (0.28)	1 (0.25)	1.33 (0.16)	0.83 (0.22)	0.39 (0.1)	1.1
IV	100	0.46 (0.05)	0.92 (0.09)	0.78 (0.12)	0.58 (0.06)	0.45 (0.04)	1
	1000	0.62 (0.22)	1.07 (0.18)	0.87 (0.11)	0.78 (0.22)	0.59 (0.04)	1.1
	2000	0.71 (0.34)	1.22 (0.26)	0.89 (0.12)	0.94 (0.31)	0.59 (0.04)	1.3
	4000	0.71 (0.26)	1.3 (0.18)	0.91 (0.13)	1 (0.22)	0.63 (0.04)	1.2
V	100	0.12 (0.02)	0.37 (0.1)	0.42 (0.18)	0.15 (0.02)	0.12 (0.02)	1
	1000	0.2 (0.03)	0.55 (0.15)	0.55 (0.22)	0.41 (0.05)	0.2 (0.05)	1
	2000	0.38 (0.34)	0.8 (0.29)	0.6 (0.24)	0.67 (0.27)	0.29 (0.18)	1.2
	4000	0.78 (0.51)	1.22 (0.31)	0.77 (0.25)	1.06 (0.41)	0.48 (0.31)	1.5

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 (0.06)	0.57 (0.15)	0.31 (0.05)	0.26 (0.05)	2
	1000	0.33 (0.07)	0.74 (0.17)	0.62 (0.04)	0.33 (0.07)	2
	2000	0.36 (0.11)	0.92 (0.18)	0.73 (0.07)	0.38 (0.08)	2
	4000	0.44 (0.14)	1.12 (0.25)	0.87 (0.1)	0.42 (0.09)	2
VII	100	0.32 (0.04)	0.67 (0.11)	0.42 (0.04)	0.32 (0.04)	2
	1000	0.6 (0.28)	0.93 (0.22)	1.02 (0.2)	0.66 (0.3)	2.1
	2000	0.95 (0.44)	1.18 (0.27)	1.35 (0.32)	0.83 (0.35)	2.3
	4000	1.17 (0.38)	1.43 (0.31)	1.47 (0.33)	1.08 (0.33)	2.1
VIII	100	0.29 (0.09)	0.61 (0.11)	0.34 (0.08)	0.25 (0.03)	2
	1000	0.37 (0.08)	0.82 (0.14)	0.69 (0.13)	0.35 (0.07)	2
	2000	0.54 (0.35)	1 (0.25)	0.92 (0.28)	0.47 (0.22)	2.2
	4000	0.88 (0.45)	1.37 (0.26)	1.27 (0.31)	0.71 (0.37)	2.5
IX	100	0.43 (0.06)	0.74 (0.12)	0.48 (0.05)	0.43 (0.07)	2
	1000	0.47 (0.09)	0.91 (0.15)	0.91 (0.05)	0.48 (0.09)	2
	2000	0.58 (0.23)	1.11 (0.23)	1.12 (0.16)	0.5 (0.1)	2.1
	4000	0.57 (0.18)	1.25 (0.22)	1.23 (0.1)	0.56 (0.11)	2

The Lasso-SIR performed the best when compared to its competitors.

4.2. Multiple Index Models

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

- (VI) $y_i = |z_{i2}/4 + 2|^3 * \text{sgn}(z_{i1}) + \epsilon_i$ where $\mathcal{S} = \{1, 2, \dots, 7\}$ and $\beta_{1:4,1} = 1, \beta_{5:7,2} = 1$, and $\beta_{ij} = 0$ otherwise;
- (VII) $y_i = z_{i1} * \exp(z_{i2}) + \epsilon_i$ where $\mathcal{S} = \{1, 2, \dots, 12\}$ and $\beta_{1:7,1}, \beta_{8:12,2} \sim N(0, 1)$, and $\beta_{ij} = 0$ otherwise;
- (VIII) $y_i = z_{i1} * \exp(z_{i2} + \epsilon_i)$ where $\mathcal{S} = \{1, 2, \dots, 12\}$ and $\beta_{1:7,1}, \beta_{8:12,2} \sim N(0, 1)$, and $\beta_{ij} = 0$ otherwise;
- (IX) $y_i = z_{i1} * (2 + z_{i2}/3)^2 + \epsilon_i$ where $\mathcal{S} = \{1, 2, \dots, 12\}$ and $\beta_{1:8,1} = 1, \beta_{9:12,2} = 1$ and $\beta_{ij} = 0$ 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 Σ_2 , Algorithm 5 tended to underestimate d , which is worthy of further investigation.

The results, reported in the supplementary material, for the other values of ρ are similar to what we observed when $\rho = 0.5$. The Lasso-SIR performs the best when compared to its competitors.

There are other sparse inverse regression method, such as the Sparse SIR, given in Li and Nachtsheim (2006). In Lin et al. (2018), we have shown that the DT-SIR outperforms this method. We thus did not include the numerical comparison. For the reason of completeness, we have included the numerical results of comparing Lasso-SIR and Sparse SIR in Section D of

the online supplementary file, showing that Lasso-SIR is better than Sparse-SIR.

4.3. Discrete Responses

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

- (X) $y = 1(\mathbf{x}\beta + \epsilon > 0)$ where $\mathcal{S} = \{1, 2, \dots, 10\}$;
- (XI) $y = 1(\exp(\mathbf{x}\beta) + \epsilon > 0)$ where $\mathcal{S} = \{1, 2, \dots, 7\}$;
- (XII) $y = 1((\mathbf{x}\beta)^3/2 + \epsilon)$ where $\mathcal{S} = \{1, 2, \dots, 20\}$;
- (XIII) Let $\mathbf{z} = \mathbf{x}\beta$ where $\mathcal{S} = \{1, 2, \dots, 12\}$, β is a p by 2 matrix with $\beta_{1:7,1}, \beta_{8:12,2} \sim N(0, 1)$ and $\beta_{ij} = 0$ 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 Dataset

We first apply the methods to a two-class classification problem, which aims to distinguish between cancer patients and normal

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

	p	Lasso-SIR	DT-SIR	M-Lasso	Lasso
X	100	0.22 (0.03)	0.66 (0.05)	0.26 (0.03)	0.2 (0.03)
	1000	0.26 (0.04)	1.21 (0.03)	0.52 (0.03)	0.28 (0.03)
	2000	0.27 (0.03)	1.33 (0.02)	0.59 (0.02)	0.29 (0.04)
	4000	0.28 (0.04)	1.39 (0.02)	0.65 (0.03)	0.3 (0.04)
XI	100	0.32 (0.07)	0.83 (0.07)	0.6 (0.17)	0.33 (0.07)
	1000	0.43 (0.1)	1.32 (0.02)	1.07 (0.05)	0.45 (0.09)
	2000	0.45 (0.09)	1.38 (0.01)	1.15 (0.04)	0.46 (0.09)
	4000	0.49 (0.12)	1.41 (0.01)	1.2 (0.05)	0.51 (0.12)
XII	100	0.24 (0.03)	0.63 (0.05)	0.52 (0.35)	0.22 (0.03)
	1000	0.33 (0.03)	1.18 (0.04)	0.53 (0.03)	0.32 (0.03)
	2000	0.37 (0.05)	1.3 (0.04)	0.62 (0.03)	0.35 (0.03)
	4000	0.4 (0.04)	1.38 (0.03)	0.68 (0.03)	0.39 (0.04)
XIII	100	0.38 (0.06)	1.09 (0.06)	0.61 (0.05)	1.07 (0.02)
	1000	0.39 (0.07)	1.79 (0.02)	1.12 (0.05)	1.08 (0.02)
	2000	0.38 (0.07)	1.91 (0.02)	1.24 (0.04)	1.09 (0.03)
	4000	0.42 (0.07)	1.98 (0.01)	1.32 (0.03)	1.1 (0.03)

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

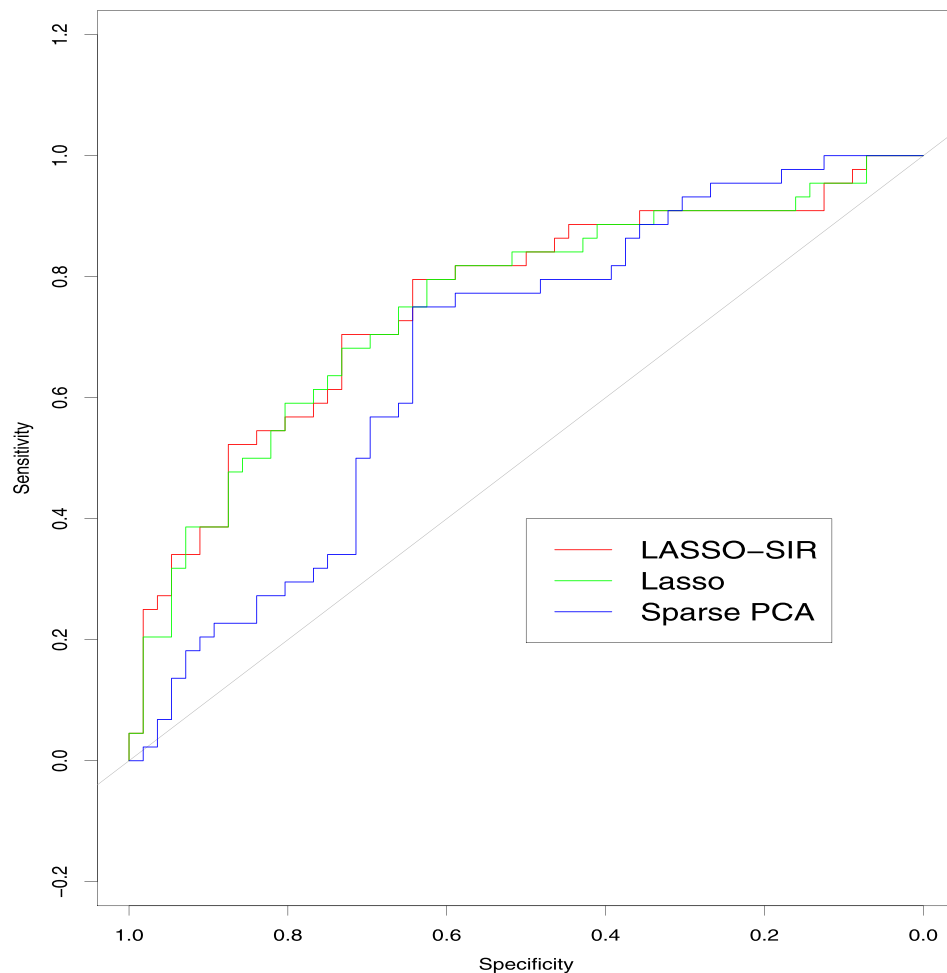
has also been used in the NIPS 2003 feature selection challenge (Guyon et al. 2004). For each subject, there are 10,000 features where 7000 of them are real variables and 3000 of them are random probes. There are 100 subjects in the validation set.

After standardizing \mathbf{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 β 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 dataset with a continuous response. We consider the gene expression data from 45 Japanese and 45 Chinese from the international “HapMap”

**Figure 1.** ROC curve of various methods for Arcene dataset.

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, Shao, and Zhou (2018), 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 dataset with d being chosen as 1 according to Algorithm 5. The number of selected variables was 13. Based on the estimated coefficients β and X , we calculated the first component and the scatterplot between the response Y and this component, showing a moderate linear relationship between them. We then fitted a linear regression between them. The R^2 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 β . 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 β , the R^2 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, Hastie, and Tibshirani 2006). We then fitted a regression model. The R^2 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 investigate the popular wine dataset which has been 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.

6. Discussion

Researchers have made some attempts to extend Lasso to nonlinear regression models in recent years (e.g., Plan and Vershynin 2016; Neykov, Cai, and Liu 2016). However, these

approaches are not efficient enough for SDR problems. In comparison, Lasso-SIR introduced in this article is an efficient high-dimensional variant of SIR (Li 1991) for obtaining a sparse solution to the estimation of the SDR subspace for multiple index models. We showed that Lasso-SIR is rate optimal if $n\lambda = p^\alpha$ for some $\alpha > 1/2$, where λ is the smallest nonzero eigenvalue of $\text{var}(\mathbb{E}[x|y])$. This technical assumption on n , λ , and p is slightly disappointing from the ultrahigh-dimensional perspective. We believe that this technical assumption arises from an intrinsic limitation in estimating the central subspace, that is, some further sparsity assumptions on either Σ or $\text{var}(\mathbb{E}[x|y])$ 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 reformulating 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 multiple index models. 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 semiparametric index models.

The R-package, *LassoSIR*, is available on CRAN (<https://cran.r-project.org/package=LassoSIR>).

Appendix A. Sketch of Proof of Theorems 1–3

We assume the condition (A1), (A2), and (A3) hold throughout of the rest of the article. In particular, the sliced stability condition (A3) requires that $H > d$ is a large enough but finite integer. We denote the SIR estimate of $\Lambda = \text{var}(\mathbb{E}[x|y])$ by $\hat{\Lambda}_H = \frac{1}{H} X_H X_H^T$ (see, e.g., (2)) and its eigenvector of unit length associated to the j th eigenvalue $\hat{\lambda}_j$ by $\hat{\eta}_j$. To avoid unnecessary confusion, we assume that $\frac{s \log(p)}{n\lambda}$ and $\frac{\sqrt{p}}{n\lambda}$ are sufficiently small. We call an event Ω happens with high probability if $\mathbb{P}(\Omega^c) \leq C_1 \exp(-C_2 \log(p))$ for some absolute constants C_1 and C_2 .

A.1. Assistant Lemmas

A.1.1. Concentration Inequalities

Lemma 1. Let d_1, \dots, d_p be positive constants. We have the following statements:

- (i) For p iid standard normal random variables x_1, \dots, x_p , there exist constants C_1 and C_2 such that for any sufficiently small a , we have

$$\mathbb{P}\left(\left|\frac{1}{p} \sum_i d_i (x_i^2 - 1)\right| > a\right) \leq C_1 \exp\left(-\frac{p^2 a^2}{C_2 \sum_j d_j^4}\right). \quad (10)$$

- (ii) For $2p$ iid standard normal random variables $x_1, \dots, x_p, y_1, \dots, y_p$, there exist constants C_1 and C_2 such that for any sufficiently small a , we have

$$\mathbb{P}\left(\left|\frac{1}{p} \sum_i d_i x_i y_i\right| > a\right) \leq C_1 \exp\left(-\frac{p^2 a^2}{C_2 \sum_j d_j^4}\right). \quad (11)$$

Proof. (ii) is a direct corollary of (i). We put the proof of (i) in the supplementary materials. \square

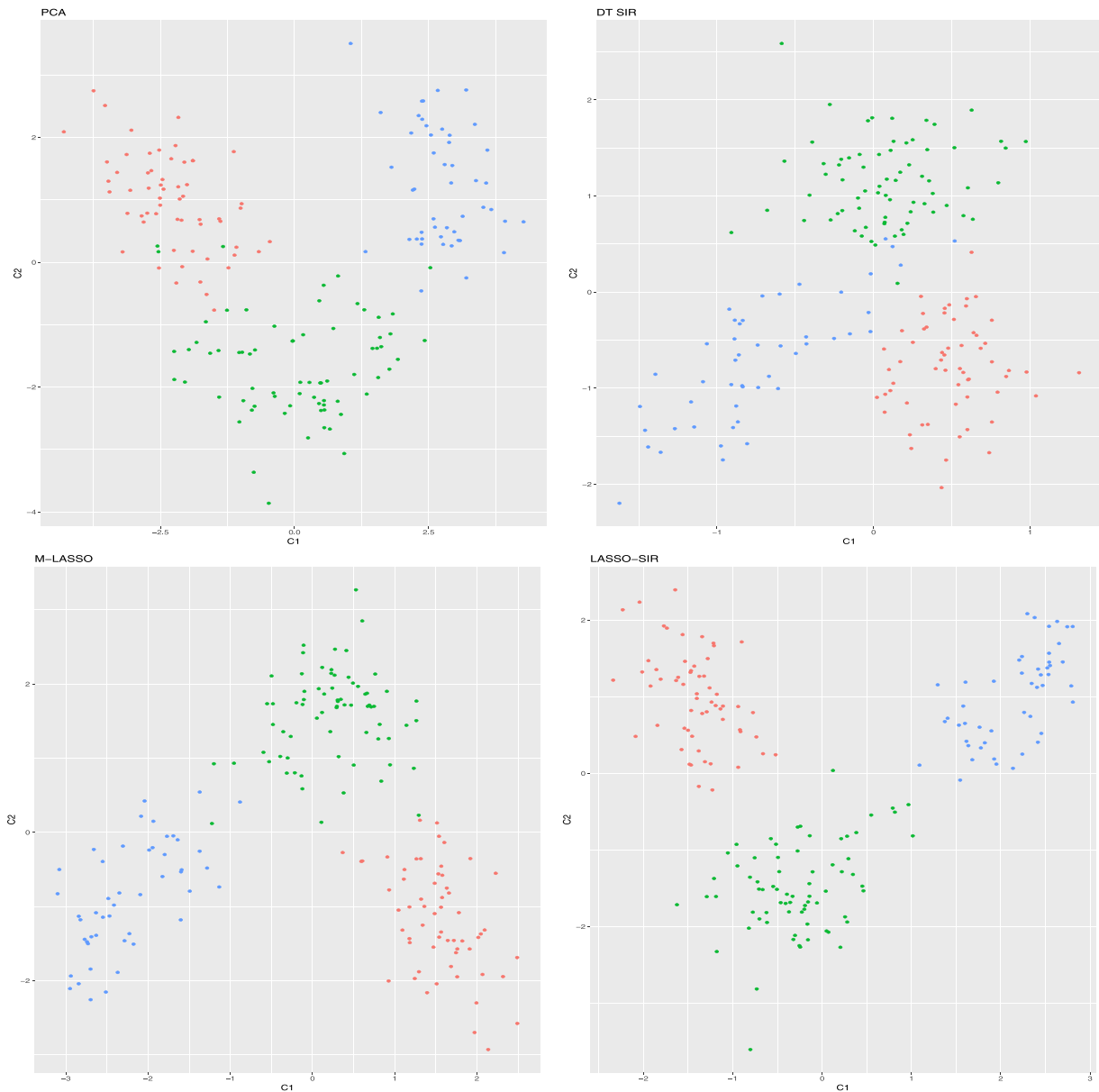


Figure 2. We plotted the second component versus the first component for all the wines, which are labeled with different colors, representing different cultivars (1, red; 2, green; 3, blue). The four methods for calculating the directions are PCA, DT-SIR, M-Lasso, and Lasso-SIR from top-left to bottom-right.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

A.1.2. Sine-Theta Theorem

Lemma 2 (Sine-Theta Theorem). Let A and $A + E$ be symmetric matrices satisfying

$$A = [F_0, F_1] \begin{bmatrix} A_0 & 0 \\ 0 & A_1 \end{bmatrix} \begin{bmatrix} F_0^\tau \\ F_1^\tau \end{bmatrix},$$

$$A + E = [G_0, G_1] \begin{bmatrix} \Lambda_0 & 0 \\ 0 & \Lambda_1 \end{bmatrix} \begin{bmatrix} G_0^\tau \\ G_1^\tau \end{bmatrix},$$

where $[F_0, F_1]$ and $[G_0, G_1]$ are orthogonal matrices. If the eigenvalues of A_0 are contained in an interval (a, b) , and the eigenvalues of Λ_1 are excluded from the interval $(a - \delta, b + \delta)$ for some $\delta > 0$, then

$$\|F_0 F_0^\tau - G_0 G_0^\tau\| \leq \frac{\min(\|F_1^\tau E G_0\|, \|F_0^\tau E G_1\|)}{\delta},$$

and

$$\frac{1}{\sqrt{2}} \|F_0 F_0^\tau - G_0 G_0^\tau\|_F \leq \frac{\min(\|F_1^\tau E G_0\|_F, \|F_0^\tau E G_1\|_F)}{\delta}.$$

A.1.3. Restricted Eigenvalue Properties

We briefly review the restricted eigenvalue (RE) property, which was first introduced in Raskutti, Wainwright, and Yu (2010). Given a set $S \subset [p] = \{1, \dots, p\}$, for any positive number α , 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 $\mathbf{X}\mathbf{X}^\tau/n$ satisfies the restricted eigenvalue condition over S with parameter $(\alpha, \kappa) \in [1, \infty) \times (0, \infty)$ if

$$\frac{1}{n} \theta^\tau \mathbf{X}\mathbf{X}^\tau \theta \geq \kappa^2 \|\theta\|_2^2, \quad \forall \theta \in \mathcal{C}(S, \alpha). \quad (12)$$

If (12) holds uniformly for all the subsets S with cardinality s , we say that $\mathbf{X}\mathbf{X}^\tau/n$ satisfies the restricted eigenvalue condition of order s with parameter (α, κ) . Similarly, we say that the covariance matrix Σ satisfies the RE condition over S with parameter (α, κ) if $\|\Sigma^{1/2}\theta\|_2 \geq \kappa \|\theta\|$ 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 Σ satisfies the restricted

eigenvalue condition of order s with parameter (α, κ) . The following corollary is borrowed from Raskutti, Wainwright, and Yu (2010).

Corollary 2. Suppose that Σ satisfies the RE condition of order s with parameter (α, κ) . Let \mathbf{X} be the $p \times n$ matrix formed by n iid samples from $N(0, \Sigma)$. For some universal positive constants α_1, α_2 , and α_3 , if the sample size satisfies

$$n > \alpha_3 \frac{(1 + \alpha)^2 \max_{i \in [p]} \Sigma_{ii}}{\kappa^2} s \log(p),$$

then the matrix $\frac{1}{n} \mathbf{X} \mathbf{X}^\tau$ satisfies the RE condition of order s with parameter $(\alpha, \frac{\kappa}{8})$ with probability at least $1 - \alpha_1 \exp(-\alpha_2 n)$.

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

Proposition 1. For some universal constants α_1, α_2 and α_3 , if the sample size satisfies that $n > \alpha_1 s \log(p)$, then the matrix $\frac{1}{n} \mathbf{X} \mathbf{X}^\tau$ satisfies the RE condition for any order s with parameter $(3, \sqrt{C_{\min}}/8)$ with probability at least $1 - \alpha_2 \exp(-\alpha_3 n)$.

A.1.4. The Sliced Approximation Lemma

Let $\mathbf{x} \in \mathbb{R}^p$ be a sub-Gaussian random variable. For any unit vector $\boldsymbol{\beta} \in \mathbb{R}^p$, let $\mathbf{x}(\boldsymbol{\beta}) = \langle \mathbf{x}, \boldsymbol{\beta} \rangle$ and $\mathbf{m}(\boldsymbol{\beta}) = \langle \mathbf{m}, \boldsymbol{\beta} \rangle = \mathbb{E}[\mathbf{x}(\boldsymbol{\beta}) | y]$. In order to get the deviation properties of the statistics $\text{var}_H(\mathbf{x}(\boldsymbol{\beta}))$, Lin et al. (2018) has introduced the sliced stable condition, that is, the condition A3 in this article. For the exact definition and more discussion, we refer to Lin et al. (2018).

Lemma 3. Let $\mathbf{x} \in \mathbb{R}^p$ be a sub-Gaussian random variable. Assume that $\mathbb{E}[\mathbf{x}|y]$ is sliced table with respect to y . For any unit vector $\boldsymbol{\beta} \in \mathbb{R}^p$, let $\mathbf{x}(\boldsymbol{\beta}) = \langle \mathbf{x}, \boldsymbol{\beta} \rangle$ and $\mathbf{m}(\boldsymbol{\beta}) = \langle \mathbf{m}, \boldsymbol{\beta} \rangle = \mathbb{E}[\mathbf{x}(\boldsymbol{\beta}) | y]$, we have the following:

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

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

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

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

with probability at most

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

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

The following proposition is a direct corollary.

Proposition 2. There exist positive constants C_1, C_2 and C_3 , such that

$$\|\boldsymbol{\beta}^\tau \Lambda_H \boldsymbol{\beta} - \boldsymbol{\beta}^\tau \text{var}(\mathbb{E}[\mathbf{x}|y]) \boldsymbol{\beta}\|_2 \geq \frac{1}{2\nu} \boldsymbol{\beta}^\tau \text{var}(\mathbb{E}[\mathbf{x}|y]) \boldsymbol{\beta} \quad (13)$$

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 3 and the fact that for any $\boldsymbol{\beta} \in \text{col}(\text{var}(\mathbb{E}[\mathbf{x}|y]))$, $\text{var}(\mathbf{m}(\boldsymbol{\beta})) \geq \lambda$. \square

A.1.5. Properties of $\hat{\boldsymbol{\eta}}_j$ s

Proposition 3. Recall that $\hat{\boldsymbol{\eta}}_j$ is the eigenvector associated to the j th eigenvalue of $\hat{\Lambda}_H, j = 1, \dots, H$. If $n\lambda = p^\alpha$ for some $\alpha > 1/2$, there exist positive constants C_1 and C_2 such that

- (i) for $j = 1, \dots, d$, one has

$$\|P_{\Lambda} \hat{\boldsymbol{\eta}}_j\|_2 \geq C_1 \sqrt{\frac{\lambda}{\hat{\lambda}_j}} \quad (14)$$

- (ii) for $j = d + 1, \dots, H$, one has

$$\|P_{\Lambda} \hat{\boldsymbol{\eta}}_j\|_2 \leq C_2 \frac{\sqrt{p \log(p)}}{n\lambda} \sqrt{\frac{\lambda}{\hat{\lambda}_j}} \quad (15)$$

hold with high probability.

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. (2018) have shown that for single index models, $\mathbb{E}[\angle(\boldsymbol{\eta}_1, \hat{\boldsymbol{\eta}}_1)] = 0$ if and only if $\lim_{n \rightarrow \infty} \frac{p}{n} = 0$. Proposition 3 states that the projection of $\sqrt{\hat{\lambda}_j} \hat{\boldsymbol{\eta}}_j, j = 1, \dots, d$, onto the true direction is nonzero if $n\lambda > p^\alpha$ where $\alpha > 1/2$.

Proof. Let $\mathbf{x} = \mathbf{z} + \mathbf{w}$ be the orthogonal decomposition with respect to $\text{col}(\text{var}(\mathbb{E}[\mathbf{x}|y]))$ and its orthogonal complement. We define two $p \times H$ matrices $\mathbf{Z}_H = (\mathbf{z}_{1,\cdot}, \dots, \mathbf{z}_{H,\cdot})$ and $\mathbf{W}_H = (\mathbf{w}_{1,\cdot}, \dots, \mathbf{w}_{H,\cdot})$ whose definition are similar to the definition of \mathbf{X}_H . We then have the following decomposition

$$\mathbf{X}_H = \mathbf{Z}_H + \mathbf{W}_H. \quad (16)$$

By definition, we know that $\mathbf{Z}_H^\tau \mathbf{W}_H = 0$ and $y \perp \mathbf{w}$. Let Σ_1 be the covariance matrix of \mathbf{w} , then $\mathbf{W}_H = \frac{1}{\sqrt{c}} \Sigma_1^{1/2} \mathbf{E}_H$ where \mathbf{E}_H is a $p \times H$ matrix with iid standard normal entries.

For sufficiently large ν_1 and α , Lemma 3 implies that

$$\begin{aligned} \Omega_1 &= \left\{ \omega \mid \left(1 - \frac{\kappa}{2\nu_1}\right) \lambda \leq \lambda_{\min}\left(\frac{1}{H} \mathbf{Z}_H^\tau \mathbf{Z}_H\right) \right. \\ &\quad \left. \leq \lambda_{\max}\left(\frac{1}{H} \mathbf{Z}_H^\tau \mathbf{Z}_H\right) \leq \left(1 + \frac{1}{2\nu_1}\right) \kappa \lambda \right\} \end{aligned} \quad (17)$$

happens with high probability and Lemma 1 implies

$$\Omega_2 = \left\{ \omega \mid \left\| \frac{1}{H} \mathbf{W}_H^\tau \mathbf{W}_H - \frac{\text{tr}(\Sigma_1)}{n} \mathbf{I}_H \right\|_F \leq \alpha \frac{\sqrt{p \log(p)}}{n} \right\} \quad (18)$$

happens with high probability.

For any $\omega \in \Omega = \Omega_1 \cap \Omega_2$, we can choose a $p \times p$ orthogonal matrix T and an $H \times H$ orthogonal matrix S such that

$$\begin{aligned} \frac{1}{H} T \mathbf{Z}_H(\omega) S &= \begin{pmatrix} B_1 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} \\ \text{and } \frac{1}{H} T \mathbf{W}_H(\omega) S &= \begin{pmatrix} 0 & 0 \\ B_2 & B_3 \\ 0 & B_4 \end{pmatrix} \end{aligned} \quad (19)$$

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

$$\begin{aligned} (1 - \frac{\kappa}{2\nu_1}) \lambda &\leq \lambda_{\min}(B_1^\tau B_1) \leq \lambda_{\max}(B_1^\tau B_1) \leq (1 + \frac{1}{2\nu_1}) \kappa \lambda \\ \left\| \begin{pmatrix} B_2^\tau B_2 & B_2^\tau B_3 \\ B_3^\tau B_2 & B_3^\tau B_3 + B_4^\tau B_4 \end{pmatrix} - \frac{\text{tr}(\Sigma_1)}{n} \mathbf{I}_H \right\|_F &\leq \alpha \frac{\sqrt{p \log(p)}}{n} \end{aligned} \quad (20)$$

Proposition 3 follows from the linear algebraic lemma:

Lemma 4. Assume that $n\lambda = p^\alpha$ for some $\alpha > 1/2$. To avoid unnecessary confusion, we also assume that $\frac{\sqrt{p \log(p)}}{n\lambda}$ is sufficiently small. Let $\mathbf{M} = \begin{pmatrix} B_1 & 0 \\ B_2 & B_3 \\ 0 & B_4 \end{pmatrix}$ 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 (20). Let $\hat{\eta}_j$ be the eigenvector associated with the j th eigenvalue $\hat{\lambda}_j$ of $\mathbf{M}\mathbf{M}^\tau$, $j = 1, \dots, H$. Then the length of the projection of $\hat{\eta}_j$ onto its first d -coordinates is at least $C\sqrt{\frac{\lambda}{\lambda_j}}$ for $j = 1, \dots, d$ and is at most $C\frac{\sqrt{p \log(p)}}{n\lambda}\sqrt{\frac{\lambda}{\lambda_j}}$ for $j = d+1, \dots, H$.

Proof. Let us consider the eigen-decompositions of

$$Q_1 \triangleq \mathbf{M}^\tau \mathbf{M} = \begin{pmatrix} E_1 & E_2 \\ E_3 & E_4 \end{pmatrix} \begin{pmatrix} D_1 & 0 \\ 0 & D_2 \end{pmatrix} \begin{pmatrix} E_1^\tau & E_3^\tau \\ E_2^\tau & E_4^\tau \end{pmatrix}$$

where D_1 (resp. D_2) is a $d \times d$ (resp. $(H-d) \times (H-d)$) diagonal matrices satisfying that $\lambda_{\min}(D_1) \geq \lambda_{\max}(D_2)$. Equation (20) implies that

$$\begin{aligned} \left(1 - \frac{\kappa}{2v_1}\right)\lambda + \frac{\text{tr}(\Sigma_1)}{n} - \alpha \frac{\sqrt{p \log(p)}}{n} &\leq \lambda_{\min}(D_1) \\ &\leq \lambda_{\max}(D_1) \leq \left(1 + \frac{1}{2v_1}\right)\kappa\lambda + \frac{\text{tr}(\Sigma_1)}{n} + \alpha \frac{\sqrt{p \log(p)}}{n}. \end{aligned}$$

On the other hand, we could consider the eigen-decomposition of

$$\begin{aligned} Q_2 &\triangleq \begin{pmatrix} B_1^\tau B_1 + \frac{\text{tr}(\Sigma_1)}{n} I_d & 0 \\ 0 & \frac{\text{tr}(\Sigma_1)}{n} I_{H-d} \end{pmatrix} \\ &= \begin{pmatrix} F_1 & 0 \\ 0 & F_2 \end{pmatrix} \begin{pmatrix} D'_1 & 0 \\ 0 & D'_2 \end{pmatrix} \begin{pmatrix} F_1^\tau & 0 \\ 0 & F_2^\tau \end{pmatrix} \end{aligned}$$

where D'_1 (resp. D'_2) is a $d \times d$ (resp. $(H-d) \times (H-d)$) diagonal matrices satisfying that $\lambda_{\min}(D'_1) \geq \lambda_{\max}(D'_2)$. Equation (20) implies that

$$\begin{aligned} \frac{\text{tr}(\Sigma_1)}{n} - \alpha \frac{\sqrt{p \log(p)}}{n} &\leq \lambda_{\min}(D'_2) \leq \lambda_{\max}(D'_2) \leq \frac{\text{tr}(\Sigma_1)}{n} \\ &+ \alpha \frac{\sqrt{p \log(p)}}{n}. \end{aligned}$$

Thus the eigen-gap is of order $\lambda - \alpha \frac{\sqrt{p \log(p)}}{n}$ (which is of order λ , since $n\lambda = p^\alpha$ for some $\alpha > 1/2$). From (20), we know that $\|Q_1 - Q_2\|_F \leq C\frac{\sqrt{p \log(p)}}{n}$. The Sine-Theta theorem (see, e.g., Lemma 2) implies that

$$\left\| \begin{pmatrix} E_1 \\ E_3 \end{pmatrix} \begin{pmatrix} E_1^\tau & E_3^\tau \end{pmatrix} - \begin{pmatrix} I_d & 0 \\ 0 & 0 \end{pmatrix} \right\|_F \leq C \frac{\sqrt{p \log(p)}}{n\lambda}, \quad (21)$$

that is, $\|E_3 E_3^\tau\|_F \leq C \frac{\sqrt{p \log(p)}}{n\lambda}$. Similar argument gives us that $\|E_2 E_2^\tau\|_F \leq C \frac{\sqrt{p \log(p)}}{n\lambda}$.

Let η be the (unit) eigenvector associated to the nonzero eigenvalue $\hat{\lambda}$ of $\mathbf{M}\mathbf{M}^\tau$. Let us write $\eta^\tau = (\eta_1^\tau, \eta_2^\tau, \eta_3^\tau)$ where $\eta_1, \eta_2 \in \mathbb{R}^d$ and $\eta_3 \in \mathbb{R}^{p-2d}$. Let $\alpha = (\alpha_1^\tau, \alpha_2^\tau)$ where $\alpha_1 = B_1^\tau \eta_1 + B_2^\tau \eta_2 \in \mathbb{R}^d$ and $\alpha_2 = B_3^\tau \eta_2 + B_4^\tau \eta_3 \in \mathbb{R}^{H-d}$. It is easy to verify that $\alpha/\sqrt{\hat{\lambda}}$ is the (unit) eigenvector associated to the eigenvalue $\hat{\lambda}$ of $\mathbf{M}^\tau \mathbf{M}$ and

$$\begin{aligned} \eta_1 &= \frac{B_1}{\sqrt{\hat{\lambda}}} \frac{\alpha_1}{\sqrt{\hat{\lambda}}}, \quad \eta_2 = \frac{B_2}{\sqrt{\hat{\lambda}}} \frac{\alpha_1}{\sqrt{\hat{\lambda}}} + \frac{B_3}{\sqrt{\hat{\lambda}}} \frac{\alpha_2}{\sqrt{\hat{\lambda}}}, \\ \text{and } \eta_3 &= \frac{B_4}{\sqrt{\hat{\lambda}}} \frac{\alpha_2}{\sqrt{\hat{\lambda}}}. \end{aligned}$$

If $\hat{\lambda}$ is among the first d eigenvalues of $\mathbf{M}^\tau \mathbf{M}$, then $\|\alpha_1/\sqrt{\hat{\lambda}}\|_2$ is bounded below by some positive constant. Thus $\|\eta_1\|_2 \geq C\sqrt{\frac{\lambda}{\hat{\lambda}}}$. If $\hat{\lambda}$ is among the last $H-d$ eigenvalues of $\mathbf{M}^\tau \mathbf{M}$, then $\|\alpha_1/\sqrt{\hat{\lambda}}\|_2 = O\left(\frac{\sqrt{p \log(p)}}{n\lambda}\right)$. Thus $\|\eta_1\|_2 \leq O\left(\kappa\sqrt{\frac{\lambda}{\hat{\lambda}}} \frac{\sqrt{p \log(p)}}{n\lambda}\right)$. \square

A.2. Sketch of Proof of Theorem 1

We only sketch some key points of the proof here and leave the details in the online supplementary files. Recall that for single index model $y = f(\beta_0^\tau x, \epsilon)$ where β_0 is a unit vector, we have denoted by $\hat{\eta}$ the eigenvector of $\hat{\Lambda}_H$ associated to the largest eigenvalue $\hat{\lambda}$. Let $\hat{\beta}$ be the minimizer of

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

where $\tilde{y} \in \mathbb{R}^n$ such that $\hat{\eta} = \frac{1}{n} X \tilde{y}$. Let $\eta_0 = \Sigma \beta_0$, $\tilde{\eta} = P_{\eta_0} \hat{\eta}$ and $\tilde{\beta} = \Sigma^{-1} \tilde{\eta} \propto \beta_0$. Since we are interested in the distance between the directions of $\hat{\beta}$ and β_0 , we consider the difference $\delta = \hat{\beta} - \tilde{\beta}$. A slight modification of the argument in Bickel, Ritov, and Tsybakov (2009)

implies that, if we choose $\mu = C\sqrt{\frac{\log(p)}{n\lambda}}$ for sufficiently large constant

C , we have $\|\delta\|_2 \leq C_1 \sqrt{\frac{s \log(p)}{n\lambda}}$ with high probability. The detailed arguments are put in the online supplementary file. Proposition 3, Condition (A1), and $\tilde{\beta} = \Sigma^{-1} \tilde{\eta}$, imply that $C_1 \sqrt{\frac{\lambda}{\hat{\lambda}}} \leq \|\tilde{\beta}\|_2 \leq C_2$ holds with high probability for some constants C_1 and C_2 . Thus, we have

$$\begin{aligned} \|P_{\hat{\beta}} - P_{\beta_0}\|_F &= \|P_{\hat{\beta}} - P_{\tilde{\beta}}\|_F \leq 4 \frac{\|\hat{\beta} - \tilde{\beta}\|_2}{\|\tilde{\beta}\|_2} \\ &= 4 \|\delta\|_2 / \|\tilde{\beta}\|_2 \leq C \sqrt{\frac{s \log(p)}{n\lambda}} \end{aligned} \quad (22)$$

holds with high probability. \square

A.3. Proof of Theorem 2

Recall that $\hat{\eta}_j$ s are the (unit) eigenvectors associated to the j th eigenvalues of $\hat{\Lambda}_H$, $j = 1, \dots, d$. We introduce the following notations,

$$\tilde{\eta}_j = P_{\Lambda} \hat{\eta}_j, \quad \tilde{\beta}_j = \Sigma^{-1} \tilde{\eta}_j \quad \text{and} \quad \gamma_j = \tilde{\beta}_j / \|\tilde{\beta}_j\|_2. \quad (23)$$

Applying the argument in Theorem 1 on these eigenvectors, we have

$$\|\hat{\beta}_j - \tilde{\beta}_j\|_2 \leq C \sqrt{\frac{s \log(p)}{n\hat{\lambda}_j}} \quad \text{and} \quad \|P_{\hat{\beta}_j} - P_{\tilde{\beta}_j}\|_F \leq C \sqrt{\frac{s \log(p)}{n\lambda}} \quad (24)$$

for some constant C hold with high probability. Since we assume that d is fixed, if we can prove that

- (I) the lengths of $\tilde{\beta}_j$, $j = 1, \dots, d$, are bounded below by $C\sqrt{\frac{\lambda}{\hat{\lambda}_j}}$,
- (II) the angles between any two vectors of $\tilde{\beta}_j$, $j = 1, \dots, d$, are bounded below by some constant,

hold with probability, then the Gram-Schmit process implies that $\|P_{\hat{\beta}} - P_{\beta}\|_F \leq C\sqrt{\frac{s \log(p)}{n\lambda}}$ holds with high probability from (24). It is easy to verify that (I) follows from the Proposition 3, the Condition (A1) and the definition of $\tilde{\beta}_j (= \Sigma^{-1} \tilde{\eta}_j)$, $j = 1, \dots, d$. (II) is a direct corollary of the following two statements.

A.3.1. Statement A. The Angles Between Any Two Vectors in $\tilde{\eta}_j, j = 1, \dots, d$ Are Nearly $\pi/2$

Since $n\lambda = p^\alpha$ for some $\alpha > 1/2$, we only need to prove that

$$\left| \cos(\angle(\tilde{\eta}_j, \tilde{\eta}_i)) \right| \leq C \frac{\sqrt{p \log(p)}}{n\lambda} \quad (25)$$

holds with high probability for any $i \neq j$. Recall that we have the following decomposition $\mathbf{X}_H = \mathbf{Z}_H + \mathbf{W}_H$. It is easy to see that $\text{col}(\mathbf{Z}_H) = \text{col}(\text{var}(\mathbb{E}[\mathbf{x}|y]))$ and $\sqrt{n} \text{cov}(\mathbf{w})^{-1/2} \mathbf{W}_H$ is identically distributed to a matrix, \mathcal{E}_1 , with all the entries are iid standard normal random variables. Let us choose an orthogonal matrix T such that $\frac{1}{\sqrt{H}} T \mathbf{Z}_H = (\mathbf{A}^\tau, 0)^\tau$ and $\frac{1}{\sqrt{H}} T \mathbf{W}_H = (0, \mathbf{B}^\tau)^\tau$ where \mathbf{A} is a $d \times H$ matrix and \mathbf{B} is a $(p-d) \times H$ matrix. Thus, $T\hat{\eta}_j$ is the eigenvector of $\frac{1}{H} T \mathbf{X}_H \mathbf{X}_H^\tau T^\tau$ associated with the j th eigenvalue $\hat{\lambda}_j, j = 1, \dots, d$. If we have a) $\lambda_{\min}(\mathbf{A} \mathbf{A}^\tau) \geq \lambda$, b) $\|P_{\text{col}(T \mathbf{Z}_H)}(T\hat{\eta}_j)\|_2 \geq C \sqrt{\frac{\lambda}{\hat{\lambda}_j}}$ and c)

$\|\mathbf{B}^\tau \mathbf{B} - \mu \mathbf{I}_H\|_F \leq C \frac{\sqrt{p \log(p)}}{n}$ for some scalar $\mu > 0$, then the statement (I) is reduced to the following linear algebra lemma.

Lemma 5. Let \mathbf{A} be a $d \times H$ matrix ($d < H$) with $\lambda_{\min}(\mathbf{A} \mathbf{A}^\tau) = \lambda$. Let \mathbf{B} be a $(p-d) \times H$ matrix such that $\|\mathbf{B}^\tau \mathbf{B} - \mu \mathbf{I}_H\|_F^2 \leq C \frac{\sqrt{p \log(p)}}{n}$. Let $\hat{\xi}_j$ be the j th (unit) eigenvector of $\mathbf{C} \mathbf{C}^\tau$ associated with the j th eigenvalue $\hat{\lambda}_j$ where $\mathbf{C}^\tau = (\mathbf{A}^\tau, \mathbf{B}^\tau)$ and $\tilde{\xi}_j$ be the projection of $\hat{\xi}_j$ onto its first d -coordinates. If $\|\tilde{\xi}_j\|_2 \geq C \sqrt{\frac{\lambda}{\hat{\lambda}_j}}$, then for any $i \neq j$,

$$\left| \cos(\angle(\tilde{\xi}_i, \tilde{\xi}_j)) \right| \leq C \frac{\sqrt{p \log(p)}}{n\lambda}. \quad (26)$$

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

Proof. Let $\hat{\alpha}_j = \mathbf{C}^\tau \hat{\xi}_j$, then $\hat{\xi}_j = \frac{1}{\hat{\lambda}_j} \mathbf{C} \alpha_j$ and $\tilde{\xi}_j = \frac{1}{\hat{\lambda}_j} \mathbf{A} \alpha_j$. It is easy to see that $\|\hat{\alpha}_j\|_2 = \sqrt{\hat{\lambda}_j}$ and $\|\mathbf{C} \hat{\alpha}_j\|_2 \geq \hat{\lambda}_j$. Since $\hat{\alpha}_j / \sqrt{\hat{\lambda}_j}$ is also the (unit) eigenvector of

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

for any $i \neq j$, we have

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

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

$$\begin{aligned} \left| \frac{\tilde{\xi}_j^\tau \tilde{\xi}_i}{\|\tilde{\xi}_i\|_2 \|\tilde{\xi}_j\|_2} \right| &\leq C \left| \frac{\hat{\lambda}_j^{1/2} \hat{\lambda}_i^{1/2}}{\hat{\lambda}_j \hat{\lambda}_i} \right| \\ &= C \left| \frac{1}{\hat{\lambda}_j} \frac{\hat{\alpha}_j^\tau}{\hat{\lambda}_j^{1/2}} (\mathbf{B}^\tau \mathbf{B} - \mu \mathbf{I}) \frac{\hat{\alpha}_i}{\hat{\lambda}_i^{1/2}} \right| \leq C \frac{\sqrt{p \log(p)}}{n\lambda}. \end{aligned}$$

□

Note that (a) follows from Lemma 2, (b) follows from Proposition 3, and (c) follows from Lemma 1. Thus statement A holds.

A.3.2. Statement B. The Angles Between Any Two Vectors in $\tilde{\beta}_j$ s Are Bounded Away From 0

Since $\tilde{\beta}_j = \mathbf{\Sigma}^{-1} \tilde{\eta}_j$, we only need to prove that there exists a positive constant $\zeta < 1$ such that

$$\left| \frac{\tilde{\eta}_i^\tau \mathbf{\Sigma}^{-1} \mathbf{\Sigma}^{-1} \tilde{\eta}_j}{\|\mathbf{\Sigma}^{-1} \tilde{\eta}_i\|_2 \|\mathbf{\Sigma}^{-1} \tilde{\eta}_j\|_2} \right| \leq \zeta. \quad (27)$$

Let $(\tilde{\eta}_1 / \|\tilde{\eta}_1\|_2, \dots, \tilde{\eta}_d / \|\tilde{\eta}_d\|_2) = T \mathbf{M}$, where T is a $p \times d$ orthogonal matrix. Since $\tilde{\eta}_j / \|\tilde{\eta}_j\|_2$ s are nearly mutually orthogonal, we know that $\mathbf{M}^\tau \mathbf{M}$ is nearly an identity matrix. Thus, by some continuity argument, the statement is reduced to the following linear algebra lemma.

Lemma 6. Let \mathbf{A} be a $p \times p$ positive definite matrix such that $C_{\min} \leq \lambda_{\min}(\mathbf{A}) \leq \lambda_{\max}(\mathbf{A}) \leq C_{\max}$ for some positive constants C_{\min} and C_{\max} . There exists constant $0 < \zeta < 1$ such that for any $p \times d$ orthogonal matrix \mathbf{B} , we have

$$\left| \frac{\mathbf{B}_{*,i}^\tau \mathbf{A}^\tau \mathbf{A} \mathbf{B}_{*,j}}{\|\mathbf{A} \mathbf{B}_{*,i}\|_2 \|\mathbf{A} \mathbf{B}_{*,j}\|_2} \right| \leq \zeta \quad \forall i \neq j. \quad (28)$$

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

A.4. Proof of Theorem 3

Recall that $\hat{\eta}_i$ is the eigenvector associated with the i th eigenvalue $\hat{\lambda}_i$ of $\hat{\Lambda}_H$, $\tilde{\eta}_i = P_{\Lambda} \hat{\eta}_i$ and $\tilde{\beta}_i = \mathbf{\Sigma}^{-1} \tilde{\eta}_i, i = 1, \dots, H$ (see, e.g., (23)). The argument in Theorem 1 implies that, for any $1 \leq i \leq H$,

$$\|\tilde{\beta}_i - \tilde{\beta}_j\|_2 \leq C \sqrt{\frac{s \log(p)}{n\lambda}}. \quad (29)$$

Proposition 1 implies that

$$\begin{aligned} \|\tilde{\beta}_i\|_2 &\geq C_1 \sqrt{\frac{\lambda}{\hat{\lambda}_i}}, 1 \leq i \leq d \\ \text{and } \|\tilde{\beta}_i\|_2 &\leq C_2 \sqrt{\frac{\lambda}{\hat{\lambda}_i} \frac{\sqrt{p \log(p)}}{n\lambda}}, d+1 \leq i \leq H. \end{aligned} \quad (30)$$

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

Supplementary Materials

The technical details and additional simulation results are put in the supplementary materials.

Funding

Jun S. Liu is partially supported by the NSF Grants DMS-1613035 and DMS-1713152, and NIH Grant R01 GM113242-01. Zhigen Zhao is partially supported by the NSF Grant IIS-1633283.

References

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009), ‘‘Simultaneous Analysis of Lasso and Dantzig Selector,’’ *The Annals of Statistics*, 37, 1705–1732. [1728,1737]

- Butucea, C., and Ingster, Y. I. (2013), "Detection of a Sparse Submatrix of a High-Dimensional Noisy Matrix," *Bernoulli*, 19, 2652–2688. [1727]
- Candes, E., and Tao, T. (2007), "The Dantzig Selector: Statistical Estimation When p is Much Larger Than n ," *The Annals of Statistics*, 35, 2313–2351. [1728]
- Chen, C. H., and Li, K. C. (1998), "Can SIR be as Popular as Multiple Linear Regression?," *Statistica Sinica*, 8, 289–316. [1727,1730]
- Chen, H. (1991), "Estimation of a Projection-Pursuit Type Regression Model," *The Annals of Statistics*, 19, 142–157. [1727]
- Cook, D. R. (1998), *Regression Graphics*. Wiley Series in Probability and Statistics: Probability and Statistics, New York: Wiley. [1726]
- (2000), "SAVE: A Method for Dimension Reduction and Graphics in Regression," *Communications in Statistics—Theory and Methods*, 29, 2109–2121. [1727]
- Cook, D. R., Forzani, L., and Rothman, A. J. (2012), "Estimating Sufficient Reductions of the Predictors in Abundant High-Dimensional Regressions," *The Annals of Statistics*, 40, 353–384. [1726]
- Duan, N., and Li, K. C. (1991), "Slicing Regression: A Link-Free Regression Method," *The Annals of Statistics*, 19, 505–530. [1728]
- Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least Angle Regression," *The Annals of Statistics*, 32, 407–499. [1729]
- Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360. [1727]
- Fan, J., Shao, Q., and Zhou, W. (2018), "Are Discoveries Spurious? Distributions of Maximum Spurious Correlations and Their Applications," *The Annals of Statistics*, 46, 989–1017. [1734]
- Friedman, J., Hastie, T., and Tibshirani, R. (2010), "Regularization Paths for Generalized Linear Models via Coordinate Descent," *Journal of Statistical Software*, 33, 1–22. [1729]
- Friedman, J. H., and Stuetzle, W. (1981), "Projection Pursuit Regression," *Journal of the American Statistical Association*, 76, 817–823. [1727]
- Guyon, I., Gunn, S., Ben-Hur, A., and Dror, G. (2005), "Result Analysis of the NIPS 2003 Feature Selection Challenge," in *Advances in Neural Information Processing Systems*, Cambridge, MA: MIT Press, pp. 545–552. [1733]
- Hsing, T., and Carroll, R. J. (1992), "An Asymptotic Theory for Sliced Inverse Regression," *The Annals of Statistics*, 20, 1040–1061. [1728]
- Ingster, Y. I., Tsybakov, A. B., and Verzelen, N. (2010), "Detection Boundary in Sparse Regression," *Electronic Journal of Statistics*, 4, 1476–1526. [1731]
- Jiang, B., and Liu, J. S. (2014), "Variable Selection for General Index Models via Sliced Inverse Regression," *The Annals of Statistics*, 42, 1751–1786. [1728]
- Jung, S., and Marron, J. S. (2009), "PCA Consistency in High Dimension, Low Sample Size Context," *The Annals of Statistics*, 37, 4104–4130. [1727]
- Li, K. C. (1991), "Sliced Inverse Regression for Dimension Reduction," *Journal of the American Statistical Association*, 86, 316–327. [1726,1727,1728,1734]
- Li, L. (2007), "Sparse Sufficient Dimension Reduction," *Biometrika*, 94, 603–613. [1726,1727,1728]
- Li, L., and Nachtsheim, C. J. (2006), "Sparse Sliced Inverse Regression," *Technometrics*, 48, 503–510. [1726,1727,1728,1732]
- Lichman, M. (2013), "UCI Machine Learning Repository," available at <http://archive.ics.uci.edu/ml>. [1733,1734]
- Lin, Q., Li, X., Dong, H., and Liu, J. S. (2016), "On the Optimality of Sliced Inverse Regression in High Dimensions," arXiv no. 1701.06009. [1726,1727,1728,1729,1730]
- Lin, Q., Zhao, Z., and Liu, J. S. (2018), "On Consistency and Sparsity for Sliced Inverse Regression in High Dimensions," *The Annals of Statistics*, 46, 580–610. [1726,1727,1728,1729,1731,1732,1736]
- Neykov, M., Cai, T., and Liu, J. S. (2016), "L1-Regularized Least Squares for Support Recovery of High Dimensional Single Index Models With Gaussian Designs," *Journal of Machine Learning Research*, 17, 1–37. [1734]
- Neykov, M., Lin, Q., and Liu, J. S. (2016), "Signed Support Recovery for Single Index Models in High-Dimensions," *Annals of Mathematical Sciences and Applications*, 1, 379–426. [1726]
- Ni, L., Cook, D. R., and Tsai, C. L. (2005), "A Note on Shrinkage Sliced Inverse Regression," *Biometrika*, 92, 242–247. [1726]
- Plan, Y., and Vershynin, R. (2016), "The Generalized Lasso With Nonlinear Observations," *IEEE Transactions on Information Theory*, 62, 1528–1537. [1734]
- Raskutti, G., Wainwright, M. J., and Yu, B. (2010), "Restricted Eigenvalue Properties for Correlated Gaussian Designs," *The Journal of Machine Learning Research*, 11, 2241–2259. [1735,1736]
- Raskutti, G., Wainwright, M. J., and Yu, B. (2011), "Minimax Rates of Estimation for High-Dimensional Linear Regression Over-Balls," *IEEE Transactions on Information Theory*, 57, 6976–6994. [1727,1730]
- Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013), "A Sparse-Group Lasso," *Journal of Computational and Graphical Statistics*, 22, 231–245. [1727]
- Thorgeirsson, T. E., Gudbjartsson, D. F., Surakka, I., Vink, J. M., Amin, N., Geller, F., Sulem, P., Rafnar, T., Esko, T., Walter, S. and Gieger, C. (2010), "Sequence Variants at CHRNA3-CHRNA6 and CYP2A6 Affect Smoking Behavior," *Nature Genetics*, 42, 448–453. [1734]
- Thorisson, G. A., Smith, A. V., Krishnan, L., and Stein, L. D. (2005), "The International HapMap Project Web Site," *Genome Research*, 15, 1592–1593. [1734]
- Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society, Series B*, 58, 267–288. [1727,1729]
- Xia, Y., Tong, H., Li, W. K., and Zhu, L. X. (2002), "An Adaptive Estimation of Dimension Reduction Space," *Journal of the Royal Statistical Society, Series B*, 64, 363–410. [1726,1727]
- Yuan, M., and Lin, Y. (2006), "Model Selection and Estimation in Regression With Grouped Variables," *Journal of the Royal Statistical Society, Series B*, 68, 49–67. [1727]
- Zhong, W., Zhang, T., Zhu, Y., and Liu, J. S. (2012), "Correlation Pursuit: Forward Stepwise Variable Selection for Index Models," *Journal of the Royal Statistical Society, Series B*, 74, 849–870. [1728]
- Zhu, L., Miao, B., and Peng, H. (2006), "On Sliced Inverse Regression With High-Dimensional Covariates," *Journal of the American Statistical Association*, 101, 640–643. [1726,1728]
- Zou, H., Hastie, T., and Tibshirani, R. (2006), "Sparse Principal Component Analysis," *Journal of Computational and Graphical Statistics*, 15, 265–286. [1734]