The Annals of Statistics
2021, Vol. 49, No. 1, 1–20
https://doi.org/10.1214/19-AOS1813

ON THE OPTIMALITY OF SLICED INVERSE REGRESSION IN HIGH DIMENSIONS

BY QIAN LIN[1], XINRAN LI[2], DONGMING HUANG[3] AND JUN S.LIU[4]

[1]*Center for Statistical Science, Department of Industrial Engineering, Tsinghua University, qianlin@tsinghua.edu.cn*
[2]*Department of Statistics, University of Illinois at Urbana-Champaign, xinranli@illinois.edu*
[3]*Department of Statistics and Applied Probability, National University of Singapore, stahd@nus.edu.sg*
[4]*Department of Statistics, Harvard University, jliu@stat.harvard.edu*

The central subspace of a pair of random variables $(y, \boldsymbol{x}) \in \mathbb{R}^{p+1}$ is the minimal subspace $\mathcal{S}$ such that $y \perp\!\!\!\perp \boldsymbol{x} | P_{\mathcal{S}}\boldsymbol{x}$. In this paper, we consider the minimax rate of estimating the central space under the multiple index model $y = f(\boldsymbol{\beta}_1^\tau \boldsymbol{x}, \boldsymbol{\beta}_2^\tau \boldsymbol{x}, \ldots, \boldsymbol{\beta}_d^\tau \boldsymbol{x}, \epsilon)$ with at most s active predictors, where $\boldsymbol{x} \sim N(0, \boldsymbol{\Sigma})$ for some class of $\boldsymbol{\Sigma}$. We first introduce a large class of models depending on the smallest nonzero eigenvalue λ of $\operatorname{var}(\mathbb{E}[\boldsymbol{x}|y])$, over which we show that an aggregated estimator based on the SIR procedure converges at rate $d \wedge ((sd + s\log(ep/s))/(n\lambda))$. We then show that this rate is optimal in two scenarios, the single index models and the multiple index models with fixed central dimension d and fixed λ. By assuming a technical conjecture, we can show that this rate is also optimal for multiple index models with bounded dimension of the central space.

1. Introduction. Because of rapid advances in information technologies in recent years, it has become a common problem for data analysts that the dimension (p) of data is much larger than the sample size (n), that is, the *"large p, small n problem*." For these problems, variable selection and dimension reduction are often the indispensable first steps. In the early 1990s, a fascinating supervised dimension reduction method, the sliced inverse regression (SIR) (Li (1991)), was proposed to discover how a univariate response relates to a low dimensional projection of the predictors. More precisely, SIR postulates the following *multiple index model* for the data

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

and estimates the subspace $\mathcal{S} = \operatorname{span}\{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d\}$ via an eigenanalysis of the estimated conditional covariance matrix $\operatorname{var}[\mathbb{E}(\boldsymbol{x}|y)]$. Note that the individual $\boldsymbol{\beta}_i$'s are not identifiable, but the space $\mathcal{S}$ can be estimated well. Based on the observation that $y \perp\!\!\!\perp \boldsymbol{x} \mid P_{\mathcal{S}}\boldsymbol{x}$, where $P_{\mathcal{S}}\boldsymbol{x}$ is the projection of $\boldsymbol{x}$ onto $\mathcal{S}$, Dennis Cook (1998) proposed a more general framework for dimension reduction without loss of information, often referred to as the *Sufficient Dimension Reduction* (SDR). Under this framework, researchers look for the minimal subspace $\mathcal{S}' \subset \mathbb{R}^p$ such that $y \perp\!\!\!\perp \boldsymbol{x} \mid P_{\mathcal{S}'}\boldsymbol{x}$, where y is no longer necessarily a scalar response. Although numerous SDR algorithms have been developed in the past decades, SIR is still the most popular one for practitioners because of its simplicity and computational efficiency. Asymptotic theories developed for these SDR algorithms have all focused on scenarios where the data dimension p is either fixed or growing at a much slower rate compared with the sample size n (Dennis Cook (2000), Li (2000), Li and Wang (2007)). The *"large p, small n"* characteristic of modern data raises new challenges to these SDR algorithms.

Received April 2018; revised November 2018.
MSC2020 subject classifications. Primary 62J02; secondary 62H25.
Key words and phrases. Sufficient dimension reduction, optimal rates, sliced inverse regression, semidefinite positive programming.

Algorithm 1 DT-SIR for Single-Index Models

1: Let $S_t = \{i | \widehat{\mathbf{\Lambda}}_H(i,i) > t\}$ for a properly chosen t, where $\widehat{\Lambda}_H$ is an estimate of $var(\mathbb{E}[x \mid y])$ as defined in (3).
2: Let $\widehat{\boldsymbol{\beta}}$ be the principal eigenvector of $\widehat{\mathbf{\Lambda}}_H(S_t, S_t)$.
3: We embed $\widehat{\boldsymbol{\beta}}$ into $\mathbb{R}^p$ by filling the entries outside S_t with 0 and denote it by $\widehat{\boldsymbol{\beta}}_{\mathrm{DT}}$.

Lin, Zhao and Liu (2018) recently showed under mild conditions that the SIR estimate of the central space is consistent if and only if $\lim \frac{p}{n} = 0$. This provides a theoretical justification for the necessity of some structural assumptions for SIR when $p > n$. A commonly employed and also practically meaningful structural assumption made for high-dimensional linear regression problems is the sparsity assumption, that is, only a few predictors among the thousands or millions of candidate ones participate in the model. We will show that this sparsity assumption can also rescue *the curse of dimension* for SDR algorithms such as SIR. Motivated by Lasso and the regularized sparse PCA (Tibshirani (1996), Zou and Hastie (2005)), Li and Nachtsheim (2006) and Li (2007) proposed some regularization approaches for SIR and SDR. However, these approaches often fail in high dimensional numerical examples and are difficult to rectify because little is known about theoretical behaviors of these algorithms in high dimensional problems. The DT-SIR algorithm (i.e., Algorithm 1) in Lin, Zhao and Liu (2018) has been shown to provide consistent estimation. The main objective of the current paper is to understand the fundamental limits of the sparse SIR problem from a decision theoretic point of view. Such an investigation not only is interesting in its own right, but will also provide insights for other SDR algorithms developed for high-dimensional problems.

Neykov, Lin and Liu (2016) considered the (signed)-support recovery problem of the following class of single index models:

$$y = f(\boldsymbol{\beta}^\tau \boldsymbol{x}, \epsilon), \quad \boldsymbol{\beta}_i \in \{\pm 1/\sqrt{s}, 0\}, \qquad \operatorname{supp}(\boldsymbol{\beta}) = s,$$

where $\boldsymbol{x} \sim N(0, \mathbf{I}_p)$, $\epsilon \sim N(0,1)$. Let $\xi = \frac{n}{s\log(p)}$. They proved that (a) if ξ is sufficiently small, any algorithm fails to recover the (signed) support of $\boldsymbol{\beta}$ with probability at least $1/2$; and (b) if ξ is sufficiently large, the DT-SIR algorithm (see Lin, Zhao and Liu (2018) or Algorithm 1) can recover the (signed) support with probability converging to 1 as $n \to \infty$. That is, the minimal sample size required to recover the support of $\boldsymbol{\beta}$ is of order $s\log(p)$. These results shed some light on the possibility of obtaining the optimal rate of SIR-type algorithms in high dimension.

SIR is widely considered as a "generalized eigenvector" problem (Chen and Li (1998)). Inspired by recent advances in sparse PCA (Amini and Wainwright (2008), Birnbaum et al. (2013), Cai, Ma and Wu (2013), Johnstone and Lu (2004), Vu and Lei (2012)), where researchers aim at estimating the principal eigenvectors of the spiked model, it is reasonable to expect a similar phase transition phenomenon (Johnstone and Lu (2004)), the signed support recovery (Amini and Wainwright (2008)), and the optimal rate (Birnbaum et al. (2013), Cai, Ma and Wu (2013), Vu and Lei (2012)) for SIR when $\mathbf{\Sigma} = \mathbf{I}$. However, as was pointed out by Lin, Zhao and Liu (2018), the sample means in the corresponding slices of the SIR algorithm are neither independent nor identically distributed. The usual concentration inequalities are not applicable. This difficulty forced them to develop the corresponding deviation properties, that is, the *"key lemma"* in Lin, Zhao and Liu (2018). On the other hand, the observation that the number H of slices is allowed to be finite when d is bounded (we always require that $H > d$) suggests that a consistent estimate of the central space based on finite (e.g., H) sample means is possible. This is again similar to the so-called *high-dimension, low sample-size* (HDLSS) scenario of PCA, which was first studied in Jung and Marron (2009) by estimating

the principal eigenvectors based on finite samples. These connections suggest that theoretical issues in sparse SIR might be analogous to those in sparse PCA. However, our results in this article suggest that sparse linear regression is a more appropriate prototype for sparse SIR.

The main contribution of this article is the determination of the minimax rate for estimating the central space. The risk of our interest is $\mathbb{E}[\|P_{\boldsymbol{V}} - P_{\widehat{\boldsymbol{V}}}\|_F^2]$, where $\boldsymbol{V}$ is an orthogonal matrix formed by an orthonormal basis of $\mathcal{S}$, and $P_{\widehat{\boldsymbol{V}}}$ is an estimate of $P_{\boldsymbol{V}}$, the projection matrix associated with the orthogonal matrix $\boldsymbol{V}$. We first construct an estimator (computationally unrealistic) such that the risk of this estimator is of order $\frac{ds+s\log(ep/s)}{n\lambda} \wedge d$. We further demonstrate that the risk of any estimator is bounded below by $\frac{s\log(ep/s)}{n\lambda} \wedge 1$ over two classes of models, $\mathfrak{M}(p, d, \lambda, \kappa)$ and $\mathfrak{M}_{s,q}(p, d, \lambda, \kappa)$, defined in (8) and (14), respectively. To the best of our knowledge, this is the first result about the minimax rate of estimating the central space in high dimension. In Section 2.7, we show that the computationally efficient algorithm DT-SIR (Lin, Zhao and Liu (2018)) achieves this optimal rate when $d = 1$ and $s = O(p^{1-\delta})$ for some $\delta > 0$. Furthermore, we investigate the effects of the slice number H in the SIR procedure.

2. Main results. Since the establishment of the SDR framework about two decades ago, estimating the central space has been investigated under different assumptions (Cook, Forzani and Rothman (2012), Dennis Cook (1998), Dennis Cook and Weisberg (1991), Ferré (1998), Hsing and Carroll (1992), Li and Wang (2007), Schott (1994)). Various SDR algorithms have their own advantages and disadvantages for certain classes of link functions (models). For example, SIR only works when both the linearity and coverage conditions are satisfied (Li (1991)); Sliced Average Variance Estimation (SAVE) (Dennis Cook and Weisberg (1991)) works when the coverage condition is slightly violated but requires the constant variance condition. Thus, to discuss the minimax rate of estimating the central space for model (1), it is necessary to first specify the class of models where one or several algorithms are practically used, and then check if these algorithms and their variants can estimate the central space optimally over this class of models. SIR is one of the most widely used and well understood SDR algorithms. It is of special interest to know if it is rate optimal over a large class of models. This will not only improve our understanding of high dimensional behaviors of SIR and its variants, but also bring us insights on behaviors of other SDR algorithms.

2.1. *Notation.* In addition to those that have been used in Section 1, we adopt the following notation throughout the article. For a matrix $\boldsymbol{V}$, we denote its column space by $\mathrm{col}(\boldsymbol{V})$ and its ith row and jth column by $\boldsymbol{V}_{i,*}$ and $\boldsymbol{V}_{*,j}$, respectively. For vectors $\boldsymbol{x}$ and $\boldsymbol{\beta} \in \mathbb{R}^p$, we denote the kth entry of $\boldsymbol{x}$ as $\boldsymbol{x}(k)$ and the inner product $\langle \boldsymbol{x}, \boldsymbol{\beta}\rangle$ as $\boldsymbol{x}(\boldsymbol{\beta})$. 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. For a matrix A, $\|A\|_F = \mathrm{tr}(AA^\tau)^{1/2}$. For a positive integer p, $[p]$ denotes the index set $\{1, 2, \ldots, p\}$. For any positive integers p and d, $\mathbb{O}(p, d)$ denotes the set of all $p \times d$ orthogonal matrices. 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.

2.2. *A brief review of SIR.* Since we are interested in the space spanned by $\boldsymbol{\beta}_i$'s in model (1), without loss of generality, we can assume that $\boldsymbol{V} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d)$ is a $p \times d$ orthogonal matrix (i.e., $\boldsymbol{V}^\tau \boldsymbol{V} = \mathbf{I}_d$) and the models considered in this paper are

$$y = f\big(\boldsymbol{V}^\tau \boldsymbol{x}, \epsilon\big), \quad \boldsymbol{V} \in \mathbb{O}(p, d), \tag{2}$$

where f is an unknown link function, $\boldsymbol{x} \sim N(0, \mathbf{I}_p)$, and $\epsilon \sim N(0,1)$ independent of $\boldsymbol{x}$. Though $\boldsymbol{V}$ is not identifiable, the column space $\operatorname{col}(\boldsymbol{V})$ can be estimated. The *Sliced Inverse Regression* (SIR) procedure proposed in Li (1991) estimate the central space $\operatorname{col}(\boldsymbol{V})$ without knowing $f(\cdot)$, which can be briefly summarized as follows. Given n $i.i.d.$ samples $(y_i, \boldsymbol{x}_i)$, $i = 1, \ldots, n$, SIR first divides them into H equal-sized slices according to the order statistics $y_{(i)}$.[1] We reexpress the data as $y_{h,j}$ and $\boldsymbol{x}_{h,j}$, where (h, j) is the double subscript in which h refers to the slice number and j refers to the order number of a sample in the hth slice, that is,

$$y_{h,j} = y_{(c(h-1)+j)}, \qquad \boldsymbol{x}_{h,j} = \boldsymbol{x}_{(c(h-1)+j)}.$$

Here, $\boldsymbol{x}_{(k)}$ is the concomitant of $y_{(k)}$ (see, e.g., Yang (1977)). Let the sample mean in the hth slice be $\bar{\boldsymbol{x}}_{h,\cdot}$, and the overall sample mean be $\bar{\bar{\boldsymbol{x}}}$. Then SIR uses

$$\widehat{\boldsymbol{\Lambda}}_H = \frac{1}{H}\sum_{h=1}^{H} \bar{\boldsymbol{x}}_{h,\cdot}\bar{\boldsymbol{x}}_{h,\cdot}^{\tau} \tag{3}$$

to estimate $\boldsymbol{\Lambda} \triangleq \operatorname{var}(\mathbb{E}[\boldsymbol{x}|y])$, and $\operatorname{col}(\widehat{\boldsymbol{V}}_H)$ to estimate the central space $\operatorname{col}(\boldsymbol{V})$, where $\widehat{\boldsymbol{V}}_H$ is the matrix formed by the top d eigenvectors of $\widehat{\boldsymbol{\Lambda}}_H$. We assume that the dimension of the central space d is known throughout the article.

In order for SIR to give a consistent estimate of the central space, the following sufficient conditions have been suggested (Hsing and Carroll (1992), Li (1991), Zhu, Miao and Peng (2006)) in addition to the "linearity condition" that is automatically satisfied for Gaussian $\boldsymbol{x}$:

(A') Coverage condition:

$$\operatorname{span}\{\mathbb{E}[\boldsymbol{x}|y]\} = \operatorname{span}\{\boldsymbol{V}_{*,1}, \ldots, \boldsymbol{V}_{*,d}\},$$

where $\boldsymbol{V}_{*,i}$ is the ith columns of the orthogonal matrix $\boldsymbol{V}$.

(B') Smoothness and tail conditions on the central curve $\mathbb{E}[\boldsymbol{x}|y]$.

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 conditions:

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

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

$$\begin{aligned} &\widetilde{m}^4(y)P\big(|Y| > y\big) \to 0 \quad \text{as } y \to \infty, \\ &\|\boldsymbol{m}(y) - \boldsymbol{m}(y')\|_2 \leq |\widetilde{m}(|y|) - \widetilde{m}(|y'|)| \\ &\qquad \text{for } y, y' \in (-\infty, -B_0) \text{ or } y, y' \in (B_0, \infty). \end{aligned} \tag{4}$$

As in Lin, Zhao and Liu (2018), where they demonstrated the phase transition phenomenon of SIR in high dimension, we replace Condition (B') by

(B'') Modified smoothness and tail conditions.

[1] To ease notation and arguments, we assume that $n = cH$.

They are the same as those in (B') except that eqn (4) is replaced by

$$\begin{aligned} &\mathbb{E}\big[\widetilde{m}(y)^4 1_{|y|>B_0}\big] < \infty, \\ &\big\|\boldsymbol{m}(y) - \boldsymbol{m}(y')\big\|_2 \le \big|\widetilde{m}(|y|) - \widetilde{m}(|y'|)\big| \\ &\qquad \text{for } y, y' \in (-\infty, -B_0) \text{ or } y, y' \in (B_0, \infty). \end{aligned} \tag{5}$$

It is easy to see that Condition (B'') is slightly stronger than Condition (B'). A main advantage of Condition (B'') is the following proposition proved in Neykov, Lin and Liu (2016).

PROPOSITION 1. *If Condition (B'') holds, the central curve $\mathbb{E}[\boldsymbol{x}|y]$ satisfies the sliced stable condition (defined below) with $\vartheta = \frac{1}{2}$.*

DEFINITION 1 (Sliced stable condition). Let Y be a random variable. For $0 < \boldsymbol{\gamma}_1 < 1 < \boldsymbol{\gamma}_2$, let $\mathcal{A}_H(\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2)$ denote all partitions $\{-\infty = a_0 \le a_2 \le \cdots \le a_H = +\infty\}$ of $\mathbb{R}$, such that

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

A curve $\boldsymbol{m}(y) \in \mathbb{R}^p$ is ϑ-sliced stable with respect to Y, if $\boldsymbol{m}(y)$ lies in a d-dimensional subspace and there exist positive constants $\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \boldsymbol{\gamma}_3, C$ such that for any $H > Cd$, for any partition $\in \mathcal{A}_H(\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2)$ and any $\boldsymbol{\beta} \in \mathbb{R}^p$, we have

$$\frac{1}{H}\sum_{h=1}^{H} \operatorname{var}\big(\boldsymbol{\beta}^\tau \boldsymbol{m}(Y) | a_{h-1} \le Y < a_h\big) \le \frac{\boldsymbol{\gamma}_3}{H^\vartheta} \operatorname{var}\big(\boldsymbol{\beta}^\tau \boldsymbol{m}(Y)\big). \tag{6}$$

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

Intuitively, $H \to \infty$ implies that the LHS of (6) converges to zero. Definition 1 states that its convergence rate is a power of H, although any function of H that converges to 0 can be used to replace $1/H^\vartheta$ on the RHS of (6). Thus, the sliced stable condition is almost the necessary condition to ensure that the SIR works. A main advantage of the sliced stable condition is that we can easily quantify the deviation properties of the eigenvalues, eigenvectors and each entries of $\widehat{\boldsymbol{\Lambda}}_H$. This is one of the main technical contributions of Lin, Zhao and Liu (2018). We henceforth assume that the central curve satisfies the sliced stable condition. As shown by Proposition 1, Condition (B'') ensures the sliced-stable condition.

2.3. *The class of functions $\mathcal{F}_d(\lambda, \kappa)$.* Let $\boldsymbol{z} = \boldsymbol{V}^\tau \boldsymbol{x}$, then $\boldsymbol{z} \sim N(0, \mathbf{I}_d)$. Let $\boldsymbol{\Lambda}_z = \operatorname{var}(\mathbb{E}[\boldsymbol{z}|y])$. Since $\mathbb{E}[\boldsymbol{x}|y] = P_{\boldsymbol{V}}\mathbb{E}[\boldsymbol{x}|y] = \boldsymbol{V}\mathbb{E}[\boldsymbol{V}^\tau \boldsymbol{x}|y] = \boldsymbol{V}\mathbb{E}[\boldsymbol{z}|y]$, the sliced stability for $\mathbb{E}[\boldsymbol{z}|y]$ implies the sliced stability for $\mathbb{E}[\boldsymbol{x}|y]$ and vice verse. Since we have assumed that $\boldsymbol{x} \sim N(0, \mathbf{I}_p)$, the linearity condition holds automatically.

Inspired by the assumption on the condition number in Cai, Ma and Wu (2013), we consider the following condition:

$$\lambda \le \lambda_d\big(\operatorname{var}(\mathbb{E}[\boldsymbol{x}|y])\big) \le \lambda_1\big(\operatorname{var}(\mathbb{E}[\boldsymbol{x}|y])\big) \le \kappa\lambda \le 1 \tag{7}$$

for some positive constant $\kappa > 1$, which is a refinement of the coverage condition, that is, $\operatorname{rank}(\operatorname{var}(\mathbb{E}[\boldsymbol{x}|y])) = d$. Without loss of generality, we assume thereafter $\lambda \le 1/2$. Since $\boldsymbol{\Lambda} \triangleq \operatorname{var}(\mathbb{E}[\boldsymbol{x}|y]) = \boldsymbol{V}\boldsymbol{\Lambda}_z\boldsymbol{V}^\tau$, we know $\lambda_j(\boldsymbol{\Lambda}) = \lambda_j(\boldsymbol{\Lambda}_z)$, $j = 1, \ldots, d$. In particular, we have $\lambda \le \lambda_d(\operatorname{var}(\mathbb{E}[\boldsymbol{z}|y])) \le \lambda_1(\operatorname{var}(\mathbb{E}[\boldsymbol{z}|y])) \le \kappa\lambda \le 1$, where κ is assumed to be a fixed constant. The class of functions that satisfy the sliced-stable and coverage conditions, denoted as $\mathcal{F}_d(\lambda, \kappa)$, is of our main interest and defined below.

DEFINITION 2. Let $\boldsymbol{z} \sim N(0, \mathbf{I}_d)$ and $\epsilon \sim N(0, 1)$. A function $f(\boldsymbol{z}, \epsilon)$ belongs to the class $\mathcal{F}_d(\lambda, \kappa)$ if the following conditions hold:

(*A*) Coverage condition: $0 < \lambda \le \lambda_d(\mathbf{\Lambda}_z) \le \cdots \le \lambda_1(\mathbf{\Lambda}_z) \le \kappa\lambda \le 1$, where $\mathbf{\Lambda}_z \triangleq \operatorname{var}(\mathbb{E}[z|f(z,\epsilon)])$.

(*B*) Sliced stable condition: $\boldsymbol{m}_z(y) = \mathbb{E}[z|f(z,\epsilon) = y]$ is sliced stable with respect to $f(z,\epsilon)$.

It is easy to see that almost all functions f that make SIR work belong to $\mathcal{F}_d(\lambda,\kappa)$ for some κ and λ.

2.4. *Upper bounds on the risks.* Suppose we have n samples generated from a multiple index model $\mathcal{M}$ with link function f and orthogonal matrix $\boldsymbol{V}$, that is, $y = f(\boldsymbol{V}^\tau \boldsymbol{x}, \epsilon)$. We are interested in the risk $\mathbb{E}_{\mathcal{M}}\|P_{\widehat{\boldsymbol{V}}} - P_{\boldsymbol{V}}\|_F^2$ where $P_{\widehat{\boldsymbol{V}}}$ is an estimate of $P_{\boldsymbol{V}}$ based on these samples. In this subsection, we provide an upper bound on this risk. All detailed proofs are deferred to Section 4 and Supplementary Material (Lin et al. (2020)).

2.4.1. *Oracle risk.* Here, we are interested in estimating the central space over the following class of models parametrized by $(\boldsymbol{V}, f)$:

$$\mathfrak{M}(p,d,\lambda,\kappa) \triangleq \big\{(\boldsymbol{V}, f)|\boldsymbol{V} \in \mathbb{O}(p,d),\, f \in \mathcal{F}_d(\lambda,\kappa)\big\}. \tag{8}$$

We refer to the risk over $\mathfrak{M}$ as the "oracle risk." The first main result of this article is the following.

THEOREM 1 (An upper bound on the minimax oracle risk). *Assuming that κ is fixed, $\frac{dp}{n\lambda}$ is sufficiently small, $d^2 \le p$ and $\log(n\lambda) \prec p$, we have*

$$\inf_{\widehat{\boldsymbol{V}}} \sup_{\mathcal{M}\in\mathfrak{M}(p,d,\lambda,\kappa)} \mathbb{E}_{\mathcal{M}}\|P_{\widehat{\boldsymbol{V}}} - P_{\boldsymbol{V}}\|_F^2 \prec d \wedge \frac{d(p-d)}{n\lambda}. \tag{9}$$

We will show that the estimate $P_{\widehat{\boldsymbol{V}}_H}$ achieves the rate in Theorem 1 where $\widehat{\boldsymbol{V}}_H$ is the $p \times d$ orthogonal matrix forming by the top-d eigenvectors of $\widehat{\mathbf{\Lambda}}_H$ (See equation (3)). This appears to contradict a result in Lin, Zhao and Liu (2018), which states that

$$\|\widehat{\mathbf{\Lambda}}_H - \operatorname{var}(\mathbb{E}[\boldsymbol{x}|y])\|_2 = O_P\Big(\frac{1}{H^\vartheta} + \frac{H^2 p}{n} + \sqrt{\frac{H^2 p}{n}}\Big). \tag{10}$$

Lin, Zhao and Liu (2018) indicates that the convergence rate (i) does not depend on d, the dimension of central subspace; (ii) does not depend on λ, the smallest nonzero eigenvalue of $\operatorname{var}(\mathbb{E}[\boldsymbol{x}|y])$; (iii) depends on H (the number of slices) and seems worse than our upper bound here. The first two differences appear simply because Lin, Zhao and Liu (2018) have assumed that d is bounded and the nonzero eigenvalues of $\operatorname{var}(\mathbb{E}[\boldsymbol{x}|y])$ are bounded below by some positive constant (i.e., the information about eigenvalues and d is absorbed by some constants). The third difference appears because we here are interested in the convergence rate of the SIR estimate of the space $\mathcal{S}$ rather than the convergence rate of the SIR estimate of the matrix $\operatorname{var}(\mathbb{E}[\boldsymbol{x}|y])$. As they have pointed out, the convergence rate of $\widehat{\mathbf{\Lambda}}_H$ might be different (slower) than the convergence rate of $P_{\widehat{\boldsymbol{V}}_H}$. More precisely, we have

$$\widehat{\mathbf{\Lambda}}_H - \mathbf{\Lambda} = (\widehat{\mathbf{\Lambda}}_H - P_{\boldsymbol{V}}\widehat{\mathbf{\Lambda}}_H P_{\boldsymbol{V}}) + (P_{\boldsymbol{V}}\widehat{\mathbf{\Lambda}}_H P_{\boldsymbol{V}} - \mathbf{\Lambda}). \tag{11}$$

From the proof of Theorem 1 of Lin, Zhao and Liu (2018), we can easily check that the first term is of rate $\frac{pH^2}{n} + \sqrt{\frac{pH^2}{n}}$ and the second term is of rate $\frac{1}{H^\vartheta}$. Since $P_{\boldsymbol{V}}\widehat{\mathbf{\Lambda}}_H P_{\boldsymbol{V}}$ and $\mathbf{\Lambda}$ share the same column space and we are interested in estimating $P_{\boldsymbol{V}}$, the convergence rate of the second term in (11) does not matter provided that H is a large enough integer. Thus, Theorem 1 does not contradict the convergence result in Lin, Zhao and Liu (2018).

REMARK 1. On the role of H. Researchers have claimed that the performance of SIR procedure is not sensitive to the choice of H, that is, H can be as large as $\frac{n}{2}$ (Hsing and Carroll (1992)) and can also be a large enough fixed integer when $d = 1$ (Duan and Li (1991)). A direct corollary of Theorem 1 is that if d is fixed, H can be a large enough constant such that $\operatorname{col}(\widehat{\boldsymbol{V}}_H)$ is an optimal estimate of $\operatorname{col}(\boldsymbol{V})$. In the SIR literature, researchers care about the eigenvectors of $\boldsymbol{\Lambda}$ and ignore the eigenvalue information. We show here that when H is relatively small comparing with the sample size, the larger the H, the more accurate the estimate of the eigenvalues of $\boldsymbol{\Lambda}$, and illustrate this phenomenon numerically in Section 3.1.

2.4.2. *Upper bound on the risk of sparse SIR.* Lin, Zhao and Liu (2018) shows that when dimension p is larger than or comparable with the sample size n, the SIR estimate of the central space is inconsistent. Thus, structural assumptions such as sparsity are necessary for high dimensional SIR problems. We here impose the weak l_q sparsity on the loading vectors $\boldsymbol{V}_{*,1}, \ldots, \boldsymbol{V}_{*,d}$. For a $p \times d$ orthogonal matrix $\boldsymbol{V}$ (i.e., $\boldsymbol{V}^\tau \boldsymbol{V} = \mathbf{I}_d$), we order the row norms in decreasing order as $\|\boldsymbol{V}_{(1),*}\|_2 \geq \cdots \geq \|\boldsymbol{V}_{(p),*}\|_2$ and define the weak l_q radius of $\boldsymbol{V}$ to be

$$\|\boldsymbol{V}\|_{q,w} \triangleq \max_{j\in[p]} j\|\boldsymbol{V}_{(j),*}\|^q. \tag{12}$$

Let $\mathbb{O}_{s,q}(p,d) = \{\boldsymbol{V} | \boldsymbol{V} \in \mathbb{O}(p,d) \text{ such that } \|\boldsymbol{V}\|_{q,w} \leq s\}$ be the set of weak l_q sparse orthogonal matrices. Weak l_q-ball is a commonly used condition for sparsity. See, for example, Abramovich et al. (2006) for wavelet estimation and Cai and Zhou (2012) for sparse covariance matrix estimation. Furthermore, we need the notion of *effective support*, which was introduced by Cai, Ma and Wu (2013). The size of *effective support* is defined to be $k_{q,s} \triangleq \lceil x_q(s,d) \rceil$, where

$$x_q(s,d) \triangleq \max\left\{0 \leq x \leq p \middle| x \leq s\left(\frac{n\lambda}{d + \log(\frac{ep}{x})}\right)^{q/2}\right\} \tag{13}$$

and $\lceil a \rceil$ denotes the smallest integer no less than $a \in \mathbb{R}$. See Cai, Ma and Wu (2013) for a more detailed discussion about sparse orthogonal matrices.

In this subsection, we are interested in estimating the central space over the following class of high dimensional models parametrized by $(\boldsymbol{V}, f)$:

$$\mathfrak{M}_{s,q}(p,d,\lambda,\kappa) \triangleq \{(\boldsymbol{V}, f) | \boldsymbol{V} \in \mathbb{O}_{s,q}(p,d), f \in \mathcal{F}_d(\lambda,\kappa)\}. \tag{14}$$

Let $\epsilon_n^2 \triangleq \frac{1}{n\lambda}(dk_{q,s} + k_{q,s}\log\frac{ep}{k_{q,s}})$. We have the following result.

THEOREM 2 (The upper bound on optimal rates). *Assume that κ is fixed, $d^2 \leq k_{q,s}$, $\log(n\lambda) \prec k_{q,s}$ and ϵ_n^2 is sufficiently small. We have*

$$\inf_{\widehat{\boldsymbol{V}}} \sup_{\mathcal{M}\in\mathfrak{M}_{s,q}(p,d,\lambda,\kappa)} \mathbb{E}_{\mathcal{M}}\|P_{\widehat{\boldsymbol{V}}} - P_{\boldsymbol{V}}\|_F^2 \prec d \wedge \frac{dk_{q,s} + k_{q,s}\log\frac{ep}{k_{q,s}}}{n\lambda}. \tag{15}$$

In order to establish the upper bound in Theorem 2, we need to construct an estimator that attains it. Let $\mathcal{B}(k_{q,s})$ be the set of all subsets of $[p]$ with size $k_{q,s}$. To ease the notation, we often drop the subscript (q,s) of $k_{q,s}$ below and assume that there are $n = 2Hc$ samples. Let us divide the samples into two equal-sized sets at random. Let $\widehat{\boldsymbol{\Lambda}}_H^{(1)}$ and $\widehat{\boldsymbol{\Lambda}}_H^{(2)}$ be the SIR estimates of $\boldsymbol{\Lambda} = \operatorname{var}(\mathbb{E}[\boldsymbol{x}|y])$ based on the first and second sets of samples, respectively. Inspired by the idea in Cai, Ma and Wu (2013), we introduce the following aggregation estimator $\widehat{\boldsymbol{V}}_E$ of $\boldsymbol{V}$.

Aggregation estimator $\widehat{\boldsymbol{V}}_E$. For each $B \in \mathcal{B}_k$, we let

$$\begin{aligned} \widehat{\boldsymbol{V}}_B &\triangleq \arg\max_{\boldsymbol{V}} \langle \widehat{\boldsymbol{\Lambda}}_H^{(1)}, \boldsymbol{V}\boldsymbol{V}^\tau \rangle = \arg\max_{\boldsymbol{V}} \operatorname{Tr}(\boldsymbol{V}^\tau \widehat{\boldsymbol{\Lambda}}_H^{(1)} \boldsymbol{V}) \\ &\text{s.t. } \boldsymbol{V}^\tau \boldsymbol{V} = \mathbf{I}_d, \|\boldsymbol{V}\|_{q,w} = k \text{ and } \operatorname{supp}(\widehat{\boldsymbol{V}}_B) \subset B \end{aligned} \tag{16}$$

and

$$B^* \triangleq \arg\max_{B \in \mathcal{B}(k)} \langle \widehat{\boldsymbol{\Lambda}}_H^{(2)}, \widehat{\boldsymbol{V}}_B \widehat{\boldsymbol{V}}_B^\tau \rangle = \arg\max_{B \in \mathcal{B}(k)} \operatorname{Tr}(\widehat{\boldsymbol{V}}_B^\tau \widehat{\boldsymbol{\Lambda}}_H^{(2)} \widehat{\boldsymbol{V}}_B).$$

Our aggregation estimator $\widehat{\boldsymbol{V}}_E$ is defined to be $\widehat{\boldsymbol{V}}_{B*}$.

B^* is a stochastic set and, for any fixed B, $\widehat{\boldsymbol{V}}_B$ is independent of the second set of samples. From the definition of $\widehat{\boldsymbol{V}}_E$, it is easy to see

$$\langle \boldsymbol{\Lambda}_H^{(2)}, \widehat{\boldsymbol{V}}_E \widehat{\boldsymbol{V}}_E^\tau - \widehat{\boldsymbol{V}}_B \widehat{\boldsymbol{V}}_B^\tau \rangle \geq 0 \tag{17}$$

for any $\widehat{\boldsymbol{V}}_B$ where $B \in \mathcal{B}$. We have shown in Lin et al. (2020) that the aggregation estimator $\widehat{\boldsymbol{V}}_E$ achieves the convergence rate on the right-hand side of (15).

2.5. *Lower bound and minimax risk.* We assume that dimension d of the central space is bounded in this subsection. This is a reasonable assumption since most numerical studies in existing literature have only $d \leq 2$ except that Ferré (1998) performed a numerical study for a model with $d = 4$ and reported that the 4th direction was difficult to discover. We have also observed from extensive numerical studies that the 4th direction is difficult to detect for $p = 10$ even with the sample size greater than 10^6. To the best of our knowledge, the optimal rate of estimating the central space depending only on n, s and p in high dimensions has never been discussed in the literature.

The semiparametric characteristic of the multiple index model brings us additional difficulties in determining the lower bound of the minimax rate. Because of our ignorance on the function class $\mathcal{F}_d(\lambda, \kappa)$, we can only establish the lower bound in two restrictive cases: (i) λ, the smallest nonzero eigenvalue of $\operatorname{var}(\mathbb{E}[\boldsymbol{x}|y])$, is a bounded below by a sufficiently small positive constant; and (ii) single index models where $d = 1$.

2.5.1. λ *is bounded below by a sufficiently small positive constant.* Assume that λ, the smallest nonzero eigenvalues of $\operatorname{var}(\mathbb{E}[\boldsymbol{x}|y])$, is bounded below by a sufficiently small positive constant and κ is a sufficiently large positive constant. We begin with the following optimal convergence rate of the *oracle risk*.

THEOREM 3 (Oracle risk). *Assume that d is bounded, λ is bounded below by a sufficiently small constant, and κ is a sufficiently large constant. If $\frac{dp}{n}$ sufficiently small and $\log(n\lambda) \prec p$, we have*

$$\inf_{\widehat{\boldsymbol{V}}} \sup_{\mathcal{M} \in \mathfrak{M}(p,d,\lambda,\kappa)} \mathbb{E}_{\mathcal{M}} \|P_{\widehat{\boldsymbol{V}}} - P_{\boldsymbol{V}}\|_F^2 \asymp d \wedge \frac{dp}{n}. \tag{18}$$

REMARK 2. Although we have assumed that the dimension of the central space d is bounded, we include it in the convergence rate to emphasize that the result holds for multiple index models.

Because of Theorem 1, we only need to establish the lower bound. We defer the detailed proof to the online Supplementary Material (Lin et al. (2020)) and briefly sketch its key steps here. One of the key steps in obtaining the lower bound is to construct a finite family of distributions that are distant from each other in the parameter space and close to each

other in terms of the KL-divergence. Recall that, for any sufficiently small $\epsilon > 0$ and any positive constant $\alpha < 1$, Cai, Ma and Wu (2013) have constructed a subset $\Theta \subset \mathbb{G}(p, d)$, the Grassmannian manifold consisting of all the d dimensional subspaces in $\mathbb{R}^p$, such that

$$|\Theta| \geq \left(\frac{c_0}{\alpha c_1}\right)^{d(p-d)} \quad \text{and}$$
$$\alpha^2\epsilon^2 \leq \|\theta_i - \theta_j\|_F^2 \leq \epsilon^2 \quad \text{for any } \theta_i, \theta_j \in \Theta$$

for some absolute constants c_0 and c_1. For any $\theta_j \in \Theta$, if we can choose a $p \times d$ orthogonal matrix $\boldsymbol{B}_j$ such that the column space of $\boldsymbol{B}_j$ corresponds to $\theta_j \in \mathbb{G}(p, d)$, we may consider the following finite class of models:

$$y = f\big(\boldsymbol{B}_j^\tau \boldsymbol{x}\big) + \epsilon, \quad \boldsymbol{x} \sim N(0, \mathbf{I}_p) \text{ and } \epsilon \sim N(0, 1).$$

Here, f is a d-variates function with bounded first derivative such that these models belong to $\mathfrak{M}(p, d, \lambda, \kappa)$ where λ is sufficiently small and κ is sufficiently large (cf. Lemma 15 in Lin et al. (2020)). Let $p_{f,\mathbf{B}}$ denote the joint density of $(y, \boldsymbol{x})$. Simple calculation shows (cf. Lemma 14 in Lin et al. (2020)) that

$$\mathrm{KL}(p_{f,\boldsymbol{B}_1}, p_{f,\boldsymbol{B}_2}) \leq C\big(\max \|\nabla f\|^2\big)\|\boldsymbol{B}_1 - \boldsymbol{B}_2\|_F^2 \leq C\|\boldsymbol{B}_1 - \boldsymbol{B}_2\|_F^2. \tag{19}$$

If we have

$$\|\boldsymbol{B}_1 - \boldsymbol{B}_2\|_F^2 \leq \|P_{\boldsymbol{B}_1} - P_{\boldsymbol{B}_2}\|_F^2, \tag{20}$$

we may apply the standard Fano type argument (e.g., Cai, Ma and Wu (2013)) to obtain the essential rate $\frac{dp}{n}$ of the lower bound.

However, (20) is not always true (e.g., it fails if $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ are two different orthogonal matrices sharing the same column space). We need to carefully specify $\boldsymbol{B}_j$ for each $\theta_j \in \Theta \subset \mathbb{G}(p, d)$ such that they satisfy the inequality (20) (cf. Lemma 22 in Lin et al. (2020). Thus we know that the rate in Theorem 1 is optimal if d is bounded, λ is a sufficiently small constant and κ is a sufficiently large constant. Once the "*oracle risk*" has been established, the standard argument in Cai, Ma and Wu (2013) leads us the following.

THEOREM 4 (Optimal rates). *Assume that d is bounded, λ is bounded below by a sufficiently small constant, and κ is a sufficiently large constant. If $n^{-1}(dk_{q,s} + k_{q,s}\log(ep/k_{q,s}))$ is sufficiently small and $\log(n\lambda) \prec p$, we have*

$$\inf_{\widehat{V}} \sup_{\mathcal{M} \in \mathfrak{M}_{s,q}(p,d,\lambda,\kappa)} \mathbb{E}_{\mathcal{M}} \big\|\widehat{\boldsymbol{V}}\widehat{\boldsymbol{V}}^\tau - \boldsymbol{V}\boldsymbol{V}^\tau\big\|_F^2 \asymp d \wedge \frac{dk_{q,s} + k_{q,s}\log\frac{ep}{k_{q,s}}}{n} \tag{21}$$

PROOF. See the Supplementary Material (Lin et al. (2020)). □

2.5.2. *Single index models.* If we restrict our consideration to single index models (i.e., $d = 1$), we have a convergence rate optimally depending on n, λ, s and p.

THEOREM 5 (Oracle risk for single index models). *Assuming that the conditions of Theorem 1 hold and that $d = 1$, we have*

$$\inf_{\widehat{V}} \sup_{\mathcal{M} \in \mathfrak{M}(p,d,\lambda,\kappa)} \mathbb{E}_{\mathcal{M}} \big\|\widehat{\boldsymbol{V}}\widehat{\boldsymbol{V}}^\tau - \boldsymbol{V}\boldsymbol{V}^\tau\big\|_F^2 \asymp 1 \wedge \frac{p}{n\lambda}. \tag{22}$$

With the upper bound in Theorem 1, all we need to do is to establish a suitable lower bound. Let us consider the following linear model:

$$y = f_\lambda(\boldsymbol{\beta}^\tau \boldsymbol{x}) = \sqrt{2\lambda}\boldsymbol{\beta}^\tau \boldsymbol{x} + \epsilon,$$

where $\boldsymbol{\beta}$ is a unit vector, $\boldsymbol{x} \sim N(0, \mathbf{I})$ and $\epsilon \sim N(0, 1)$ and $\lambda \le 1/2$. Simple calculation shows that

$$\operatorname{var}\big(\mathbb{E}[\boldsymbol{\beta}^\tau \boldsymbol{x}|y]\big) = \frac{2\lambda}{1+2\lambda} \ge \lambda \quad \text{and} \quad \nabla f_\lambda \le C\sqrt{\lambda}.$$

Thus, inequality (19) becomes

$$\mathrm{KL}(p_{f,\boldsymbol{\beta}_1}, p_{f,\boldsymbol{\beta}_2}) \le C\big(\max \|\nabla f\|^2\big)\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_F^2 \le C\lambda\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_F^2 \tag{23}$$

and the desired lower bound follows from the same argument as that of Theorem 3. Once the oracle risk has been established, the standard argument in Cai, Ma and Wu (2013) leads us to the following result.

THEOREM 6 (Optimal rates: $d = 1$). *Assume that the conditions of Theorem 2 hold and that $d = 1$. We have*

$$\inf_{\widehat{V}} \sup_{\mathcal{M} \in \mathfrak{M}_{s,q}(p,d,\lambda,\kappa)} \mathbb{E}_{\mathcal{M}} \big\| \widehat{\boldsymbol{V}}\widehat{\boldsymbol{V}}^\tau - \boldsymbol{V}\boldsymbol{V}^\tau \big\|_F^2 \asymp 1 \wedge \frac{k_{q,s} \log \frac{ep}{k_{q,s}}}{n\lambda}. \tag{24}$$

PROOF. It is similar to the proof of Theorem 4, and thus omitted. □

2.5.3. *Multiple index models with d bounded.* The arguments in the Section 2.5.2 motivate us to propose the following (conjectural) property for the function class $\mathcal{F}_d(\lambda, \kappa)$.

CONJECTURE 1. *If d is bounded, there is a constant C such that for any $0 < \lambda \le 1$, there exists a d-variate function f_λ such that $f_\lambda(x_1, \ldots, x_d) + x_{d+1} \in \mathcal{F}_d(\lambda, \kappa)$ and*

$$\big\|\nabla f_\lambda(x_1, \ldots, x_d)\big\| \le C\sqrt{\lambda}. \tag{25}$$

REMARK 3. Inequality (25) can be slightly relaxed to that it holds with high probability for $x \sim N(0, \mathbf{I}_d)$.

The construction in Section 2.5.2 shows that this conjecture holds for $d = 1$. For $d > 1$, suppose that there exists a function f such that $f(x_1, \ldots, x_d) + x_{d+1} \in \mathcal{F}_d(\mu, \kappa)$. We expect that, for $y = \sqrt{\lambda} f(\boldsymbol{x}) + \epsilon$, there exist constants C_1 and C_2 such that

$$C_1\lambda \le \lambda_d\big(\operatorname{var}(\mathbb{E}_\lambda[\boldsymbol{x}|y])\big) \le \lambda_1\big(\operatorname{var}(\mathbb{E}_\lambda[\boldsymbol{x}|y])\big) \le C_2\kappa\lambda.$$

Note that the density function $p(y)$ of y is the convolution of the density functions of ϵ and $\sqrt{\lambda} f(\boldsymbol{x})$. Heuristically, if $f(\boldsymbol{x})$ is (nearly) normal, by the continuity of the convolution operator, we expect that $\lambda_d(\operatorname{var}(\mathbb{E}[\boldsymbol{x}|y])) \asymp \lambda$. Since we cannot prove it rigorously, we present some supporting numerical evidences in Section 3.2. Assuming this conjecture, we have the following theorems, of which the proofs are similar to those of Theorems 3 and 4.

THEOREM 7 (Oracle risk). *Assuming that the conditions of Theorem 1 hold, d is bounded and Conjecture 1 holds, we have*

$$\inf_{\widehat{V}} \sup_{\mathcal{M} \in \mathfrak{M}(p,d,\lambda,\kappa)} \mathbb{E}_{\mathcal{M}} \big\| \widehat{\boldsymbol{V}}\widehat{\boldsymbol{V}}^\tau - \boldsymbol{V}\boldsymbol{V}^\tau \big\|_F^2 \asymp d \wedge \frac{dp}{n\lambda}. \tag{26}$$

PROOF. It is similar to the proof of Theorem 3, and thus omitted. □

THEOREM 8 (Optimal Rates). *Assuming that the conditions of Theorem 2 hold, d is fixed and Conjecture 1 holds, we have*

$$\inf_{\widehat{V}} \sup_{\mathcal{M}\in\mathfrak{M}_{s,q}(p,d,\lambda,\kappa)} \mathbb{E}_{\mathcal{M}} \big\| \widehat{\boldsymbol{V}}\widehat{\boldsymbol{V}}^{\tau} - \boldsymbol{V}\boldsymbol{V}^{\tau} \big\|_F^2 \asymp d \wedge \frac{dk_{q,s} + k_{q,s}\log\frac{ep}{k_{q,s}}}{n\lambda}. \tag{27}$$

PROOF. It is similar to the proof of Theorem 4, and thus omitted. □

2.6. *Beyond the uncorrelated predictors.* So far we have shown that the lower bound $\frac{s\log(p/s)}{n\lambda} \wedge 1$ is achievable for a quite general class of single-index models with uncorrelated predictors. A natural further question is whether a rate-optimal estimator for the SDR direction with correlated predictors (i.e., when $\boldsymbol{x} \sim N(0, \boldsymbol{\Sigma})$) can achieve the lower bound as well.

A complete answer is beyond the scope of a single paper. In fact, the minimax rate for linear regression with Gaussian design is obtained only for $\boldsymbol{\Sigma}$ with bounded eigenvalues (Raskutti, Wainwright and Yu (2011)); and the minimax rate for sparse PCA is derived only for spiked models where the irrelevant noises are uncorrelated (Cai, Ma and Wu (2013)). Because the semiparametric characteristic of SIR makes it more difficult to analyze, it is within our expectation that the minimax rate results for single or multiple index models are even less complete and concise than those for linear regression and sparse PCA.

We provide here a slightly more general statement regarding the minimax rate with correlated predictors. More precisely, we consider the class $\mathfrak{M}_s(p, d, \lambda, \kappa, \boldsymbol{\Sigma})$ consisting of models $y = f(\boldsymbol{\Gamma}^{\tau}\boldsymbol{x}, \epsilon)$, $\boldsymbol{x} \sim N(0, \boldsymbol{\Sigma})$ and $\epsilon \sim N(0, 1)$ where the covariance matrix $\boldsymbol{\Sigma}$ and the $p \times d$ orthogonal matrix $\boldsymbol{\Gamma}$ satisfying the following condition:

$$\|J_K\boldsymbol{\Gamma}\|_F \le C\|J_K\boldsymbol{\Sigma}\boldsymbol{\Gamma}\|_F \tag{28}$$

for any $K \subset [p]$, where J_K is the diagonal matrix $\operatorname{diag}\{J_1, \ldots, J_p\}$ with $J_j = 1$ if $j \in K$ and 0, otherwise. We further assume the following conditions:

(G1) $\boldsymbol{\Sigma}$ has at most k nonzero entries in each row where k is a fixed integer and $C_1 \le \lambda_{\min}(\boldsymbol{\Sigma}) \le \lambda_{\max}(\boldsymbol{\Sigma}) \le C_2$ for some constants C_1 and C_2.
(G2) $\mathbb{E}[\boldsymbol{x} \mid y]$ satisfying the sliced stability condition.
(G3) $0 < \lambda \le \lambda_d(\operatorname{var}(\mathbb{E}[\boldsymbol{x} \mid y])) \le \cdots \le \lambda_1(\operatorname{var}(\mathbb{E}[\boldsymbol{x} \mid y])) \le \kappa\lambda$ for some constant κ.
(G4) $|\operatorname{supp}(\boldsymbol{\Gamma})| \le s$.

Then we have the following result.

THEOREM 9 (Optimal rates). *Assuming that d is fixed, $s = o(p)$, $s\log(p)/(n\lambda)$ is sufficiently small, Conjecture 1, and conditions* (G1)–(G4) *hold, we have*

$$\inf_{\widehat{\boldsymbol{\Gamma}}} \sup_{\mathcal{M}\in\mathfrak{M}_{s,q}(p,d,\lambda,\kappa,\boldsymbol{\Sigma})} \mathbb{E}_{\mathcal{M}} \big\| \widehat{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\Gamma}}^{\tau} - \boldsymbol{\Gamma}\boldsymbol{\Gamma}^{\tau} \big\|_F^2 \asymp 1 \wedge \frac{s\log(p)}{n\lambda}. \tag{29}$$

PROOF. A sketch of proof is presented in the Supplementary Material. □

In a recent work (Lin, Zhao and Liu (2019)), we proposed the Lasso-SIR algorithm to estimate the central space for general $\boldsymbol{\Sigma}$ and showed that it achieves the optimal rate in certain regions.

REMARK 4. One could easily verify that condition (28) implies the following:

- There is a constant C satisfying that, for any j, $1 \le j \le p$ and i, $1 \le i \le d$, we have

$$\sum_i |\langle \mathbf{\Sigma}_j, \mathbf{\Gamma}_i \rangle|^2 \ge C \sum_i |\mathbf{\Gamma}_i(j)|^2 \|\mathbf{\Sigma}_j\|_2^2, \tag{A}$$

where $\mathbf{\Gamma}_i(j)$ is the jth coordinate of $\mathbf{\Gamma}_i$ and $\mathbf{\Sigma}_j$ is the jth column (or row) vector of $\mathbf{\Sigma}$.

In fact, let K be any integer $\in \{1, 2, \ldots, p\}$. We know that Condition (28) implies Condition (A).

On the other hand, assuming that condition (A) holds for any $1 \le j \le p$. Now, for any $K \subset [p]$, we have

$$\|J_K \mathbf{\Sigma} \mathbf{\Gamma}\|_F^2 = \sum_{i=1}^d \sum_{k \in K} \big(\langle \mathbf{\Sigma}_k, \mathbf{\Gamma}_i \rangle\big)^2 \ge C \sum_i \sum_k \|\mathbf{\Sigma}_k\|_2^2 \|\mathbf{\Gamma}_i(k)\|_2^2 \ge C \lambda_{\min}(\mathbf{\Sigma})^2 \|J_K \mathbf{\Gamma}\|_F^2,$$

where we use $\mathbf{\Gamma}$ to denote the matrix $(\mathbf{\Gamma}_1, \ldots, \mathbf{\Gamma}_d)$.

The condition (A) might be easier to verify than the condition (28) in some cases. For example, if the angles between $\mathbf{\Sigma}_j$'s and $\mathbf{\Gamma}_i$ are away from $\frac{\pi}{2}$, that is, there exists a constant C such that $|\langle \mathbf{\Sigma}_j, \mathbf{\Gamma}_i \rangle| \ge C \|\mathbf{\Gamma}_i\|_2 \|\mathbf{\Sigma}_j\|_2$, then condition (A) holds since $\|\mathbf{\Gamma}_i\|_2 \ge \|\mathbf{\Gamma}_i(j)\|_2$.

2.7. *Optimality of DT-SIR.* In the previous section, we have proved that the aggregation estimator $\widehat{\boldsymbol{V}}_E$ is rate optimal. In practice, however, it is computationally too expensive. The DT-SIR algorithm proposed in Lin, Zhao and Liu (2018) is computationally efficient in general, and can be further simplified when $\mathbf{\Sigma}_{\boldsymbol{x}} = \mathbf{I}$. In this section, we focus on the single index models with the exact sparsity on the loading vector $\boldsymbol{\beta}$, that is, $|\operatorname{supp}(\boldsymbol{\beta})| = s$.

THEOREM 10. *Suppose that $s = O(p^{1-\delta})$ for some $\delta > 0$, $\frac{s \log(p)}{n\lambda}$ is sufficiently small and $n = O(p^C)$ for some constant C. Let $\widehat{\boldsymbol{\beta}}_{\mathrm{DT}}$ be the DT-SIR estimate with threshold level $t = C_1 \frac{\log(p)}{n}$ for some constant C_1, then we have*

$$\mathbb{E}_{\boldsymbol{\beta}} \|P_{\widehat{\boldsymbol{\beta}}_{\mathrm{DT}}} - P_{\boldsymbol{\beta}}\|^2 \le C_2 \frac{s \log(p - s)}{n\lambda}. \tag{30}$$

PROOF. See the Supplementary Material (Lin et al. (2020)). □

From Theorem 10, it is easy to see that, if $s = O(p^{1-\delta})$, the DT-SIR estimator $P_{\widehat{\boldsymbol{\beta}}_{\mathrm{DT}}}$ is rate optimal for $n > s \log(p)$. However, this is not the case for sparse PCA since the diagonal thresholding (DT) algorithm achieves the minimax rate only if $n > s^2 \log(p)$. This leads us to speculate that a more appropriate prototype of sparse SIR should be sparse linear regression instead of sparse PCA (Lin, Zhao and Liu (2019)). The idea of comparing SIR with linear regression dates back to the birth of SIR (Chen and Li (1998)). In support of this viewpoint, we note that the diagonal elements of $\widehat{\mathbf{\Lambda}}$ can be treated as a generalization of the square of $\mathbb{E}[yx_i]$.

EXAMPLE 1. Consider the simple model $y = a_1 x_1 + a_2 x_2 + \epsilon$ where $x_1, x_2 \sim N(0, 1)$ and $\epsilon \sim N(0, 1)$. It is easy to show that

$$\operatorname{var}\big(\mathbb{E}[x_i \mid y]\big) = \frac{a_i^2}{1 + a_1^2 + a_2^2} = \operatorname{cor}(y, x_i)^2 \approx a_i^2 = \big(\mathbb{E}[yx_i]\big)^2$$

if a_1^2 and a_2^2 are sufficiently small.

TABLE 1
The empirical mean (standard error) of the SIR estimate $\widehat{\lambda}(\mu)$ (true λ equals to μ here)

	n/1000	$H=2$	$H=5$	$H=10$	$H=50$	$H=100$	$H=200$	$H=500$
$\mu=0.5$	5	0.319	0.446	0.479	0.503	0.509	0.520	0.551
		(0.013)	(0.017)	(0.017)	(0.016)	(0.017)	(0.018)	(0.017)
	10	0.318	0.448	0.480	0.500	0.505	0.510	0.525
		(0.009)	(0.012)	(0.012)	(0.012)	(0.012)	(0.013)	(0.012)
	50	0.319	0.448	0.479	0.498	0.500	0.501	0.504
		(0.004)	(0.006)	(0.005)	(0.006)	(0.005)	(0.006)	(0.006)
	100	0.319	0.448	0.479	0.498	0.499	0.501	0.503
		(0.003)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
$\mu=0.3$	5	0.190	0.271	0.288	0.307	0.313	0.328	0.371
		(0.011)	(0.012)	(0.014)	(0.015)	(0.015)	(0.014)	(0.016)
	10	0.191	0.27	0.288	0.302	0.307	0.312	0.335
		(0.008)	(0.009)	(0.010)	(0.010)	(0.010)	(0.010)	(0.012)
	50	0.191	0.269	0.288	0.299	0.3	0.302	0.307
		(0.003)	(0.004)	(0.005)	(0.005)	(0.005)	(0.005)	(0.004)
	100	0.191	0.269	0.288	0.299	0.3	0.301	0.303
		(0.002)	(0.003)	(0.003)	(0.004)	(0.003)	(0.003)	(0.003)
$\mu=0.1$	5	0.064	0.091	0.098	0.109	0.117	0.136	0.190
		(0.007)	(0.008)	(0.009)	(0.009)	(0.009)	(0.010)	(0.010)
	10	0.0643	0.0901	0.0973	0.103	0.108	0.117	0.144
		(0.005)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.007)
	50	0.0638	0.0899	0.0963	0.101	0.101	0.103	0.109
		(0.002)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
	100	0.0636	0.0898	0.0961	0.100	0.100	0.102	0.104
		(0.001)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)

3. Numerical studies. We illustrate three aspects of the high dimensional behavior of SIR via numerical experiments. The first experiment focuses on the impacts of the choice of H (assuming that it is small relative to the sample size) in SIR: the larger the H, the more accurate the estimate of eigenvalue of $\operatorname{var}(\mathbb{E}[\boldsymbol{x}|y])$. The second experiment aims at providing supporting evidences for Conjecture 1. The third experiment demonstrates empirical performances of the DT-SIR algorithm.

3.1. *Effects of H.* Our numerical results below show that the accuracy in estimating the eigenvalues of $\operatorname{var}(\mathbb{E}[\boldsymbol{x}|y])$ depends on the choice of H. In general, the larger the H is, the more accurate the estimation, provided that there are a sufficient number of samples within each slice. Let us consider the following linear model:[2]

$$\text{Model } \mu: \quad y = \sqrt{\frac{\mu}{1-\mu}}\boldsymbol{x}_1 + \epsilon, \quad \boldsymbol{x} \sim N(0, \mathbf{I}_p), \epsilon \sim N(0,1). \tag{31}$$

It is easy to see that the only nonzero eigenvalue of $\operatorname{var}(\mathbb{E}[\boldsymbol{x}|y])$ is μ. The results are shown in Table 1, where H ranges in $\{2, 5, 10, 50, 100, 200, 500\}$, μ in $\{0.5, 0.3, 0.1\}$ and n in $\{5000, 10{,}000, 50{,}000, 100{,}000\}$. Each entry is the empirical mean (standard deviation), calculated based on 100 replications, of the SIR estimate of $\widehat{\mu}$ for given μ, n and H.

From Table 1, it is clear that the estimations became quite acceptable when H ranges from 10 to 100. The larger the n is, the more accurate the estimations are. The cautious reader may

[2]Up to a monotone transform, this is the only case that we can give the explicit value of $\lambda(\operatorname{var}(\mathbb{E}[\boldsymbol{x}|y]))$.

TABLE 2
The empirical expectation of $\lambda_1(\mu)/\mu$

$n\mu/H^2$	$\mu=1$	$\mu=0.5$	$\mu=0.1$	$\mu=0.05$	$\mu=0.01$	$\mu=0.005$	$\mu=0.001$	$\mu=0.0005$	$\mu=0.0001$
2	0.352	0.579	1.438	1.732	1.978	1.999	2.009	1.985	1.994
	(0.002)	(0.004)	(0.014)	(0.013)	(0.014)	(0.011)	(0.010)	(0.009)	(0.007)
4	0.341	0.579	1.415	1.697	1.949	1.969	1.991	1.982	1.977
	(0.001)	(0.002)	(0.005)	(0.005)	(0.006)	(0.006)	(0.005)	(0.005)	(0.006)
10	0.333	0.571	1.409	1.706	1.944	1.970	1.970	1.964	1.971
	(4.6e-4)	(1.0e-3)	(1.8e-3)	(2.2e-03)	(2.4e-3)	(2.6e-03)	(1.4e-03)	(1.7e-03)	(1.8e-03)
20	0.329	0.565	1.412	1.702	1.950	1.971	1.963	1.968	1.969
	(3.2e-04)	(3.4e-04)	(1.0e-03)	(1.5e-03)	(1.5-03)	(1.2e-03)	(8.9e-04)	(1.1e-03)	(1.1e-03)

notice that, in the row with $\mu = 0.1$ and $n = 5000$, the empirical mean and the standard error are not behaving as we have expected, for example, when $H = 500$, the empirical mean and standard error are 0.190 and 0.010, respectively, which are worse than the case with $H = 10$. This is not contradicting our theory. Note that in the Lemma 1, the deviation property of $\widehat{\lambda}$ depends on the value $\frac{n\mu}{H^2}$, that is, the larger the $\frac{n\mu}{H^2}$ is, the more concentrated the $\widehat{\lambda}$ is. In particular, for the entry corresponding to $\mu = 0.1$, $n = 5000$ and $H = 500$, the value $\frac{n\mu}{H^2} = 1/500$ is much smaller than the corresponding value, 5, associated with the entry with $\mu = 0.1$, $n/1000 = 5$ and $H = 10$.

3.2. *Supporting evidences for Conjecture* 1. Let us consider the following model with two indexes (i.e., $d = 2$):

$$\text{Model } \mu: \quad y = \sqrt{\mu}\big(1 + g(\boldsymbol{x}_1)\big)\big(g(\boldsymbol{x}_1) + g(\boldsymbol{x}_2)\big) + \epsilon, \tag{32}$$

where $g : \mathbb{R} \mapsto \mathbb{R}$ is a smooth function such that for a small constant $\delta > 0$,

$$g(x) = \begin{cases} x & \text{if } |x| \le 100 - \delta, \\ 0 & \text{if } |x| \ge 100 + \delta \end{cases} \tag{33}$$

and $|g'(x)| \le C$ for some constant C. Let $\lambda_1(\mu)$ and $\lambda_2(\mu)$ be the two eigenvalues of $\operatorname{var}(\mathbb{E}[\boldsymbol{x}|y])$. Since we know that the absolute value of the derivative of the link function $\le C\sqrt{\mu}$, we want to check if $C_1\mu \le \lambda_2(\mu) \le \lambda_1(\mu) \le C_2\mu$ holds for some positive constant C_1 and C_2 and if model (32) belongs to $\mathcal{F}_2(C_1\mu, C_2/C_1)$. We study the boundedness of $\lambda_1(\mu)/\mu$ and $\lambda_2(\mu)/\mu$ via numerical simulation. In the simulation, we choose H to be 20. Let μ range in $\{1, 0.5, 0.1, 0.05, 0.01, 0.005, 0.001, 0.0005, 0.0001\}$ and the ratio $n\mu/H^2$ range in $\{2, 4, 10, 20\}$.

In Tables 2 and 3, each entry is the average of 100 replications. For a fixed μ, the larger the ratio $n\mu/H^2$, the more accurate the estimation of $\lambda_i(\mu)/\mu$, $i = 1, 2$. In particular, it is easy to see from the rows with the ratio $n\mu/H^2 = 20$ that $\lambda_i(\mu)/\mu$, $i = 1, 2.$ are bounded.

3.3. *Performance of DT-SIR*. We assume the exact sparsity $s = O(p^{1-\delta})$ for some $\delta \in (0, 1)$, and consider the following data generating models:

$$\begin{aligned}
&\text{Model 1}: \quad y = \boldsymbol{x}^\tau\boldsymbol{\beta} + \sin\big(\boldsymbol{x}^\tau\boldsymbol{\beta}\big) + \epsilon, \\
&\text{Model 2}: \quad y = 2\arctan\big(\boldsymbol{x}^\tau\boldsymbol{\beta}\big) + \epsilon, \\
&\text{Model 3}: \quad y = \big(\boldsymbol{x}^\tau\boldsymbol{\beta}\big)^3 + \epsilon, \\
&\text{Model 4}: \quad y = \sinh\big(\boldsymbol{x}^\tau\boldsymbol{\beta}\big) + \epsilon,
\end{aligned}$$

TABLE 3
The empirical expectation of $\lambda_2(\mu)/\mu$

$n\mu/H^2$	$\mu=1$	$\mu=0.5$	$\mu=0.1$	$\mu=0.05$	$\mu=0.01$	$\mu=0.005$	$\mu=0.001$	$\mu=0.0005$	$\mu=0.0001$
2	0.110	0.127	0.126	0.099	0.051	0.037	0.027	0.025	0.024
	(4.5e-4)	(6.4e-4)	(6.1e-4)	(3.7e-4)	(1.9e-4)	(1.1e-4)	(6.0e-05)	(7.1e-05)	(6.3e-05)
4	0.097	0.114	0.109	0.091	0.039	0.028	0.014	0.013	0.012
	(2.6e-4)	(2.4e-4)	(2.6e-4)	(2.2e-4)	(7.6e-5)	(5.2e-5)	(2.2e-05)	(2.0e-5)	(1.6e-5)
10	0.094	0.111	0.103	0.083	0.032	0.021	0.008	0.006	0.005
	(1.0e-4)	(1.1e-4)	(1.1e-4)	(5.9e-5)	(2.9e-5)	(1.6e-5)	(4.9e-6)	(4.1e-06)	(2.4e-06)
20	0.092	0.108	0.102	0.080	0.003	0.002	0.005	0.004	0.004
	(4.e-5)	(6.3e-5)	(4.9e-5)	(4.0e-5)	(1.3e-5)	(7.6e-6)	(2.0e-06)	(1.4e-06)	(6.9e-07)

where $\boldsymbol{x} \sim N(\mathbf{0}, \boldsymbol{I}_p)$, $\epsilon \sim N(0,1)$, $\boldsymbol{x} \perp\!\!\!\perp \epsilon$, and $\boldsymbol{\beta}$ is a fixed vector with s nonzero coordinates. Let $\psi = \{s\log(p-s)/n\}^{-1}$. The dimension p of the predictors takes value in $\{100, 200, 300, 600, 1200\}$, the sparsity parameter δ is fixed at 0.5, and ψ takes values in $\{3, 5, 7, \ldots, 61\}$. For each (p, ψ) combination, $s = \lfloor p^{1-\delta} \rfloor$, $n = \lfloor \psi s \log(p-s) \rfloor$, and we simulate data from each model 1000 times. We then get the estimate $\widehat{\boldsymbol{\beta}}_{\mathrm{DT}}$ using DT-SIR algorithm, and the results of the average values of $\|P_{\widehat{\boldsymbol{\beta}}_{\mathrm{DT}}} - P_{\boldsymbol{\beta}}\|^2$ for each model with each (p, ψ) combination are shown in Figure 1, which shows the distance between the estimated projection matrix and the true one becomes smaller as ψ increases for all fixed p.

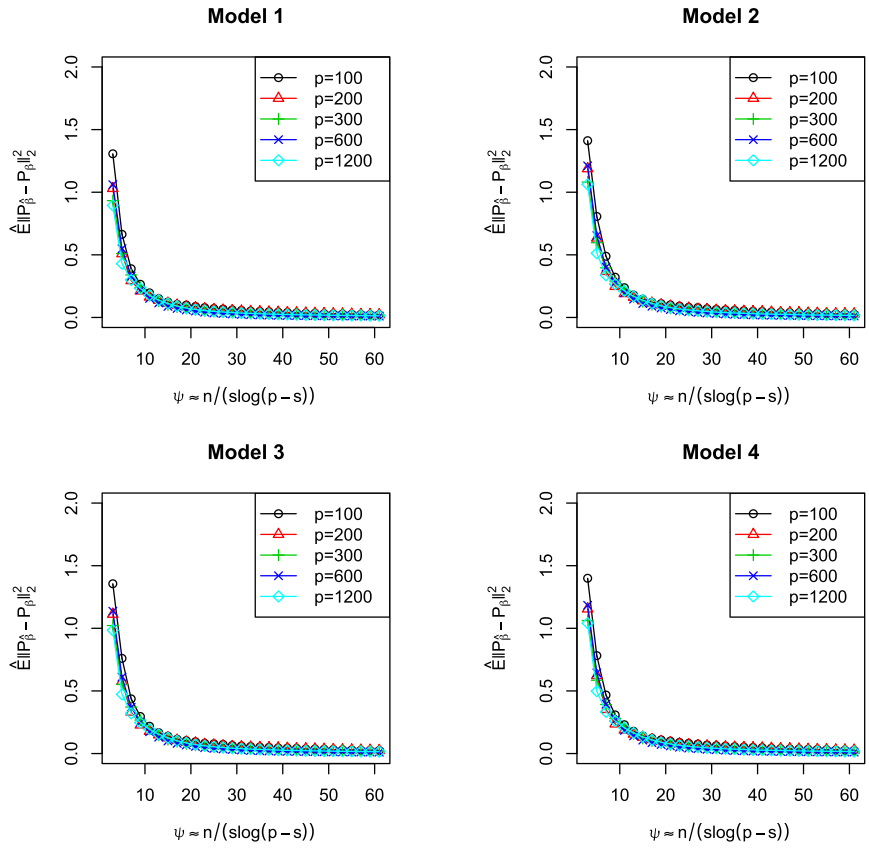

FIG. 1. *Average values of* $\|P_{\widehat{\boldsymbol{\beta}}_{\mathrm{DT}}} - P_{\boldsymbol{\beta}}\|^2$.

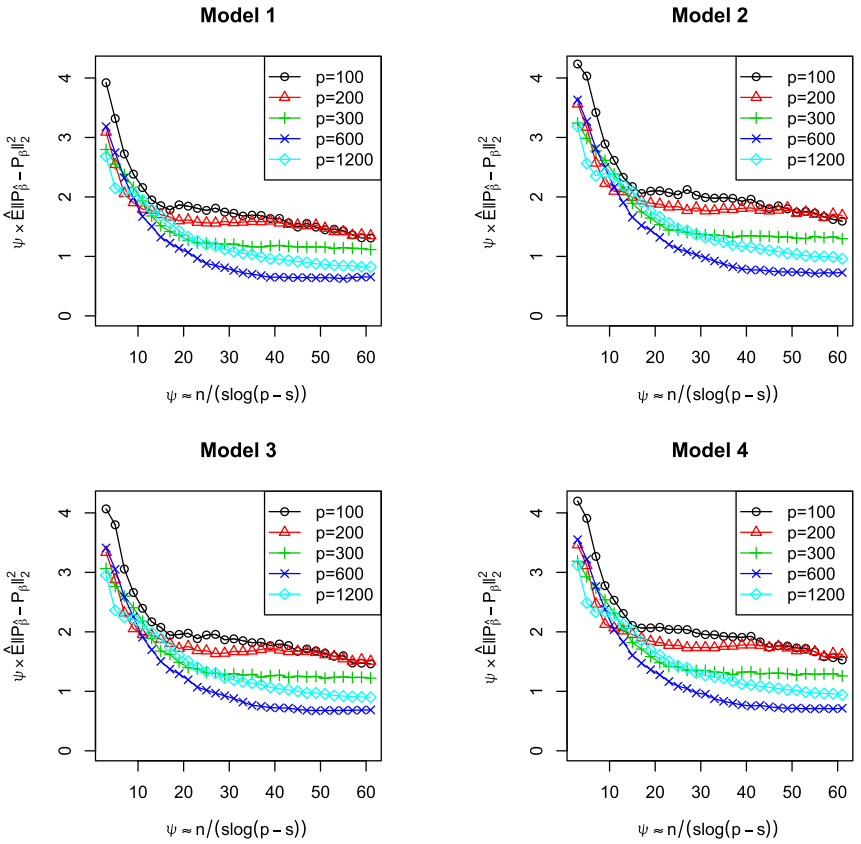

FIG. 2. *Average values of* $\psi \| P_{\widehat{\boldsymbol{\beta}}_{\mathrm{DT}}} - P_{\boldsymbol{\beta}} \|^2$.

According to Theorem 10, $\psi \| P_{\widehat{\boldsymbol{\beta}}_{\mathrm{DT}}} - P_{\boldsymbol{\beta}} \|^2$ is less than a constant with high probability. Therefore, we also display the average values of $\psi \| P_{\widehat{\boldsymbol{\beta}}_{\mathrm{DT}}} - P_{\boldsymbol{\beta}} \|^2$ for these models in Figure 2, which demonstrates that $\psi \| P_{\widehat{\boldsymbol{\beta}}_{\mathrm{DT}}} - P_{\boldsymbol{\beta}} \|^2$ is a decreasing function of ψ and tends to stabilize when ψ becomes large enough. These empirical results also validate Theorem 10.

4. Proofs. We need the following technical lemma, which can be derived from the proof of the *"key lemma"* in Lin, Zhao and Liu (2018).

LEMMA 1. *Assume that* $f \in \mathcal{F}_d(\lambda, \kappa)$ *in the model* (2). *Let* $\widehat{\boldsymbol{\Lambda}}_H$ *be the SIR estimate* (3) *of* $\mathrm{var}(\mathbb{E}[\boldsymbol{x}|y])(= \boldsymbol{\Lambda})$. *There exist positive absolute constants* C, C_1, C_2 *and* C_3 *such that, for any* $f \in \mathcal{F}_d(\lambda, \kappa)$ *and any* $\nu > 1$, *if* $H > C(\nu^{1/\vartheta} \vee d)$ *for sufficiently large constant* C, *then for any unit vector* $\boldsymbol{\beta}$ *that lies in the column space of* $\boldsymbol{\Lambda}$, *we have*

$$|\boldsymbol{\beta}^\tau (\widehat{\boldsymbol{\Lambda}}_H - \boldsymbol{\Lambda})\boldsymbol{\beta}| > \frac{1}{2\nu} \boldsymbol{\beta}^\tau \boldsymbol{\Lambda} \boldsymbol{\beta} \tag{34}$$

with probability at most

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

In particularly, if d *and* ν *are bounded, we can choose* H *to be a large enough finite integer such that* (34) *holds with high probability.*

PROOF. It is a direct corollary of the *"key lamma"* in Lin, Zhao and Liu (2018). □

Notation: Suppose that we have $n = Hc$ samples $(y_i, \boldsymbol{x}_i)$ from the distribution defined by the model $\mathcal{M} = (\boldsymbol{V}, f) \in \mathfrak{M}(p, d, \kappa, \lambda)$. Let $H = H_1 d$ where H_1 is a sufficiently large integer and $\widehat{\boldsymbol{V}} = (\widehat{\boldsymbol{V}}_1, \ldots, \widehat{\boldsymbol{V}}_d)$ where $\widehat{\boldsymbol{V}}_i$ is the eigenvector associated to the ith largest eigenvalue of $\widehat{\boldsymbol{\Lambda}}_H$. We introduce the following decomposition:

$$\boldsymbol{x} = P_{\mathcal{S}}\boldsymbol{x} + P_{\mathcal{S}^\perp}\boldsymbol{x} \triangleq \boldsymbol{\zeta} + \boldsymbol{w},$$

that is, $\boldsymbol{\zeta}$ lies in the central space $\mathcal{S}$ and $\boldsymbol{w}$ lies in the space $\mathcal{S}^\perp$ which is perpendicular to $\mathcal{S}$. Let $\boldsymbol{V}^\perp$ be a $p \times (p-d)$ orthogonal matrix such that $\boldsymbol{V}^\tau \boldsymbol{V}^\perp = 0$. Since $\mathcal{S} = \operatorname{span}\{\boldsymbol{V}\}$ and $\boldsymbol{x} \sim N(0, \mathbf{I}_p)$, we may write $\boldsymbol{w} = \boldsymbol{V}^\perp \boldsymbol{\epsilon}$ for some $\boldsymbol{\epsilon} \sim N(0, \mathbf{I}_{p-d})$. Thus we know that $\boldsymbol{\Sigma}_{\boldsymbol{w}} \triangleq \operatorname{var}(\boldsymbol{w}) = \boldsymbol{V}^\perp \boldsymbol{V}^{\perp,\tau}$. We introduce the notation $\overline{\boldsymbol{\zeta}}_{h,\cdot}$, $\overline{\boldsymbol{w}}_{h,\cdot}$ and $\overline{\boldsymbol{\epsilon}}_{h,\cdot}$ similar to the definition of $\overline{\boldsymbol{x}}_{h,\cdot}$. Let $\mathcal{Z} = \frac{1}{\sqrt{H}}(\overline{\boldsymbol{\zeta}}_{1,\cdot}, \overline{\boldsymbol{\zeta}}_{2,\cdot}, \ldots, \overline{\boldsymbol{\zeta}}_{H,\cdot})$, $\mathcal{W} = \frac{1}{\sqrt{H}}(\overline{\boldsymbol{w}}_{1,\cdot}, \overline{\boldsymbol{w}}_{2,\cdot}, \ldots, \overline{\boldsymbol{w}}_{H,\cdot})$, $\mathcal{E} = \frac{1}{\sqrt{H}}(\overline{\boldsymbol{\epsilon}}_{1,\cdot}, \overline{\boldsymbol{\epsilon}}_{2,\cdot}, \ldots, \overline{\boldsymbol{\epsilon}}_{H,\cdot})$ be three $p \times H$ matrices formed by the vectors $\frac{1}{\sqrt{H}}\overline{\boldsymbol{\zeta}}_{h,\cdot}$, $\frac{1}{\sqrt{H}}\overline{\boldsymbol{w}}_{h,\cdot}$, and $\frac{1}{\sqrt{H}}\overline{\boldsymbol{\epsilon}}_{h,\cdot}$. We have the following decomposition:

$$\begin{aligned}\widehat{\boldsymbol{\Lambda}}_H &= \mathcal{Z}\mathcal{Z}^\tau + \mathcal{Z}\mathcal{W}^\tau + \mathcal{W}\mathcal{Z}^\tau + \mathcal{W}\mathcal{W}^\tau \\ &= \boldsymbol{\Lambda}_u + \mathcal{Z}\mathcal{E}^\tau \boldsymbol{V}^{\perp,\tau} + \boldsymbol{V}^\perp \mathcal{E}\mathcal{Z}^\tau + \boldsymbol{V}^\perp \mathcal{E}\mathcal{E}^\tau \boldsymbol{V}^{\perp,\tau},\end{aligned} \tag{35}$$

where we define $\boldsymbol{\Lambda}_u \triangleq \mathcal{Z}\mathcal{Z}^\tau$ and use the fact $\mathcal{W} = \boldsymbol{V}^\perp \mathcal{E}$. Since $\boldsymbol{\epsilon} \sim N(0, \mathbf{I}_{p-d})$, we know that the entries $\mathcal{E}_{i,j}$ of $\mathcal{E}$ are $i.i.d.$ samples of $N(0, \frac{1}{n})$.

4.1. *Proof of Theorem 1.* First, we have the following lemma.

LEMMA 2. *Let $\rho = \frac{p}{n}$. Assume that $\frac{p}{n\lambda}$ is sufficiently small. We have the following statements*:

(i) *There exist constants C_1, C_2 and C_3 such that*

$$\mathbb{P}\big(\|\mathcal{W}\mathcal{W}^\tau\| > C_1(\rho + t)\big) \le C_2 \exp(-C_3 nt).$$

We will take $t = p/n$ in the late argument.

(ii) *For any vector $\boldsymbol{\beta} \in \mathbb{R}^p$ and any $v > 1$, let $E_{\boldsymbol{\beta}}(v) = \{|\boldsymbol{\beta}^\tau(\boldsymbol{\Lambda}_{\boldsymbol{u}} - \boldsymbol{\Lambda})\boldsymbol{\beta}| > \frac{1}{2v}\boldsymbol{\beta}^\tau \boldsymbol{\Lambda}\boldsymbol{\beta}\}$. Recall that $H = d H_1$. If we choose H_1 sufficiently large such that $H^\vartheta > Cv$ for some positive constant C, there exist positive constants $C_1, \ldots, C_3$ and C_4 such that*

$$\mathbb{P}\Big(\bigcup_{\boldsymbol{\beta}} E_{\boldsymbol{\beta}}(v)\Big) \le C_1 \exp\Big(-C_2 \frac{n\lambda}{H^2 v^2} + C_3 \log(H) + C_4 d\Big).$$

(iii) *For any $v > 1$, there exist positive constants $C_1, \ldots, C_6$ and C_7, such that*

$$\begin{aligned}\mathbb{P}(\|\mathcal{W}\mathcal{Z}^\tau\| > C_7\sqrt{\kappa\lambda\rho}) \le C_1 \exp\Big(-C_2 \frac{n\lambda}{H^2 v^2} + C_3 \log(H) + C_4 d\Big) \\ + C_5 \exp(-C_6 p).\end{aligned}$$

PROOF. (i) is a direct corollary of Lemma 23. (ii) is a direct corollary of Lemma 1 and the usual ϵ-net argument. (iii) is a direct corollary of (i) and (ii). □

Let $\mathrm{E} = \mathrm{E}_1 \cap \mathrm{E}_2 \cap \mathrm{E}_3$ where $\mathrm{E}_1 = \{\|\mathcal{W}\mathcal{W}^\tau\| \le C\rho\}$, $\mathrm{E}_2 = \{\|\mathcal{W}\mathcal{Z}^\tau\| \le C\sqrt{\kappa\lambda\rho}\}$, $\mathrm{E}_3 = \{\|\boldsymbol{\Lambda}_u - \boldsymbol{\Lambda}\| \le \frac{1}{2v}\kappa\lambda\}$, and C is a constant larger than $2C_1$ and C_7 in the lemma.

COROLLARY 1. *Assume* $\log(n\lambda) \prec p$. *Lemma 2 implies the following simple results where C stands for some absolute constant which might be varying in different statements*:

(a) $\mathbb{P}(\mathbb{E}^c) \le \frac{CH^2}{n\lambda}$.
(b) *Conditioning on* $\mathbb{E}_3$, *we have* $\lambda_d(\mathbf{\Lambda_u}) \ge (1 - \frac{\kappa}{2v})\lambda$.
(c) *Conditioning on* $\mathbb{E}$, *if* $\frac{p}{n\lambda}$ *is sufficiently small, we have* $\|\widehat{\mathbf{\Lambda}}_H - \mathbf{\Lambda_u}\| \le C\sqrt{\frac{\kappa\lambda p}{n}}$.
(d) *Conditioning on* $\mathbb{E}$, *If* $\frac{p}{n\lambda}$ *is sufficiently small, we have* $\lambda_{d+1}(\widehat{\mathbf{\Lambda}}_H) < \frac{1}{4}\lambda$.

Now we start the proof of Theorem 1. Note that

$$\begin{aligned}&\mathbb{E}\|\widehat{\boldsymbol{V}}\widehat{\boldsymbol{V}}^\tau - \boldsymbol{V}\boldsymbol{V}^\tau\|_F^2\\ &\quad= \underbrace{\mathbb{E}\|\widehat{\boldsymbol{V}}\widehat{\boldsymbol{V}}^\tau - \boldsymbol{V}\boldsymbol{V}^\tau\|_F^2\mathbf{1}_{\mathbb{E}}}_{I} + \underbrace{\mathbb{E}\|\widehat{\boldsymbol{V}}\widehat{\boldsymbol{V}}^\tau - \boldsymbol{V}\boldsymbol{V}^\tau\|_F^2\mathbf{1}_{\mathbb{E}^c}}_{II}.\end{aligned}$$

For II. It is easy to see that

$$II \le 2\big(d \wedge (p-d)\big)\mathbb{P}\big(\mathbb{E}^c\big) = 2d\mathbb{P}\big(\mathbb{E}^c\big) \le \frac{CdH^2}{n\lambda} = \frac{Cd^3H_1^2}{n\lambda}.$$

For I. Let $\mathbf{\Lambda_u} = \widetilde{\boldsymbol{V}}\boldsymbol{D}_H\widetilde{\boldsymbol{V}}^\tau$ be the spectral decomposition of $\mathbf{\Lambda_u}$, where $\widetilde{\boldsymbol{V}}$ is a $p \times d$ orthogonal matrix and $\boldsymbol{D}_H$ is a $d \times d$ diagonal matrix. Conditioning on $\mathbb{E}$, we know that $\widetilde{\boldsymbol{V}}$ and $\boldsymbol{V}$ are sharing the same column space. Thus we have $\widetilde{\boldsymbol{V}}\widetilde{\boldsymbol{V}}^\tau = \boldsymbol{V}\boldsymbol{V}^\tau$. Let us apply the Sin–Theta theorem (e.g., Lemma 24to the pair of symmetric matrices $(\mathbf{\Lambda_u}, \widehat{\mathbf{\Lambda}}_H = \mathbf{\Lambda_u} + \boldsymbol{Q})$ where $\boldsymbol{Q} \triangleq \widehat{\mathbf{\Lambda}}_H - \mathbf{\Lambda}_u$. Since $\frac{p}{n\lambda}$ is sufficiently small, conditioning on $\mathbb{E}$, we have $\lambda_{d+1}(\widehat{\mathbf{\Lambda}}_H) \le \frac{1}{4}\lambda$ and $\lambda_d(\mathbf{\Lambda_u}) = \lambda_d(\boldsymbol{D}_H) \ge \frac{\lambda}{2}$. Thus, we have

$$\begin{aligned}\mathbb{E}\|\boldsymbol{V}\boldsymbol{V}^\tau - \widehat{\boldsymbol{V}}\widehat{\boldsymbol{V}}^\tau\|_F^2\mathbf{1}_{\mathbb{E}} &= \mathbb{E}\|\widetilde{\boldsymbol{V}}\widetilde{\boldsymbol{V}}^\tau - \widehat{\boldsymbol{V}}\widehat{\boldsymbol{V}}^\tau\|_F^2\mathbf{1}_{\mathbb{E}}\\ &\le \frac{32}{\lambda^2}\min\big(\mathbb{E}\|\widetilde{\boldsymbol{V}}^{\perp,\tau}\boldsymbol{Q}\widehat{\boldsymbol{V}}\|_F^2\mathbf{1}_{\mathbb{E}}, \mathbb{E}\|\widetilde{\boldsymbol{V}}^\tau\boldsymbol{Q}\widehat{\boldsymbol{V}}^\perp\|_F^2\mathbf{1}_{\mathbb{E}}\big)\\ &\le \frac{32}{\lambda^2}\min\big(\mathbb{E}\|\boldsymbol{Q}\widetilde{\boldsymbol{V}}\|_F^2\mathbf{1}_{\mathbb{E}}, \mathbb{E}\|\boldsymbol{Q}\widetilde{\boldsymbol{V}}^\perp\|_F^2\mathbf{1}_{\mathbb{E}}\big).\end{aligned}$$

Since $\widetilde{\boldsymbol{V}}$ and $\boldsymbol{V}$ are sharing the same column space, we have $\widetilde{\boldsymbol{V}}^\tau\mathcal{W} = \boldsymbol{V}^\tau\mathcal{W} = 0$ and $\widetilde{\boldsymbol{V}}^{\perp,\tau}\mathcal{Z} = \boldsymbol{V}^{\perp,\tau}\mathcal{Z} = 0$. Thus, we have

$$\widetilde{\boldsymbol{V}}^\tau\boldsymbol{Q} = \widetilde{\boldsymbol{V}}^\tau\mathcal{Z}\mathcal{W}^\tau, \widetilde{\boldsymbol{V}}^{\perp,\tau}\boldsymbol{Q} = \widetilde{\boldsymbol{V}}^{\perp,\tau}\mathcal{W}\mathcal{W}^\tau + \widetilde{\boldsymbol{V}}^{\perp,\tau}\mathcal{W}\mathcal{Z}^\tau.$$

Conditioning on $\mathbb{E}$, we have $\|\mathbf{\Lambda}_u\|_2 \le 2\kappa\lambda$. Thus

$$\min\big(\mathbb{E}\|\boldsymbol{Q}\widetilde{\boldsymbol{V}}\|_F^2\mathbf{1}_{\mathbb{E}}, \mathbb{E}\|\boldsymbol{Q}\widetilde{\boldsymbol{V}}^\perp\|_F^2\mathbf{1}_{\mathbb{E}}\big) \le 2\mathbb{E}\|\widetilde{\boldsymbol{V}}^\tau\mathcal{Z}\mathcal{W}^\tau\|_F^2\mathbf{1}_{\mathbb{E}} \le 4\kappa\lambda\mathbb{E}\|\mathcal{W}^\tau\|_F^2 \le \frac{4\kappa\lambda}{n}d(p-d).$$

Since κ is assumed to be fixed, we know that if $\frac{p}{n\lambda}$ is sufficiently small and $d^2 \le p$, we have

$$\sup_{\mathcal{M}\in\mathfrak{M}(p,d,\kappa,\lambda)}\mathbb{E}\|\widehat{\boldsymbol{V}}\widehat{\boldsymbol{V}}^\tau - \boldsymbol{V}\boldsymbol{V}^\tau\|_F^2 \prec \frac{d(p-d)}{n\lambda}.$$

5. Discussion. In this paper, we have determined the minimax rate of estimating the central space over a large class of models $\mathfrak{M}_{s,q}(p, d, \lambda, \kappa)$ in two scenarios: (1) single index models, and (2) d and λ are bounded. Here, λ, the smallest nonzero eigenvalue of $\mathrm{var}(\mathbb{E}[\boldsymbol{x}|y])$, plays the role of signal strength in SIR and can be viewed as a generalized notion of the signal-to-noise ratio for multiple index models. Since we have established an upper

bound of convergence rate of estimating the central space for all d and λ, we will attempt to show that this convergence rate is optimal even for diverging d and λ in a future research.

The aggregation estimator we constructed here is actually an estimator of the column space of $\operatorname{var}(\mathbb{E}[\boldsymbol{x}|y])$ rather than that of the central space. Since we have assumed that $\boldsymbol{\Sigma} = \mathbf{I}$ in this paper, the column space of $\operatorname{var}(\mathbb{E}[\boldsymbol{x}|y])$ coincides with the central space in model (1). When there are correlations between predictors, if we assume that the eigenvectors associated with nonzero eigenvalues of $\operatorname{var}(\mathbb{E}[\boldsymbol{x}|y])$ are sparse (with sparsity s) instead of assuming that the loading vectors $\boldsymbol{\beta}_i$'s are sparse, our argument in this paper implies that $\mathbb{E}[\|P_{\widehat{\operatorname{col}(\operatorname{var}(\mathbb{E}[\boldsymbol{x}|y]))}} - P_{\operatorname{col}(\operatorname{var}(\mathbb{E}[\boldsymbol{x}|y]))}\|_F^2]$ converges at the rate $\frac{ds+s\log(ep/s)}{n\lambda}$.

Although our studies of the sparse SIR were inspired by recent advances in sparse PCA, the results in this paper suggest a more intimate connection between SIR and linear regressions. Recall that for the linear regression model $y = \boldsymbol{\beta}^\tau \boldsymbol{x} + \epsilon$ with $\boldsymbol{x} \sim N(0, \mathbf{I})$ and $s = O(p^{1-\delta})$, the minimax rate (Raskutti, Wainwright and Yu (2011)) of estimating $\boldsymbol{\beta}$ is achieved by the simple correlation screening. Analogously, the minimax rate for estimating $P_{\boldsymbol{\beta}}$ is achieved by the DT-SIR algorithm of Lin, Zhao and Liu (2018), which simply screens each variable based on the estimated variance of its conditional means. This fact suggests that a more appropriate prototype of SIR in high dimensions might be linear regression rather than sparse PCA, because there is a computational barrier of the rate optimal estimates for sparse PCA (Berthet and Rigollet (2013)). This possibility further suggests that an efficient (rate optimal) high dimensional variant of SIR with general variance matrix $\boldsymbol{\Sigma}$ might be possible, since it is now well known that lasso (Tibshirani (1996)) and the Dantzig selector (Candes and Tao (2007)) achieve the optimal rate of linear regression (Bickel, Ritov and Tsybakov (2009)) for general $\boldsymbol{\Sigma}$. This speculation warrants further future investigations.

Acknowledgments. The authors thank the Associate Editor and three referees for their constructive comments. Lin's research is supported in part by the NSFC (Grant 11971257), Beijing NSF (Grant Z190001) and Beijing Academy of Artificial Intelligence. Liu's research is supported in part by NIH Grant R01 GM113242-01, NSF grants DMS-1613035 and DMS-1903139.

SUPPLEMENTARY MATERIAL

Supplement to "On the optimality of SIR in high dimensions" (DOI: 10.1214/19-AOS1813SUPP; .pdf). This supplementary material contains technical proofs omitted from main article.

REFERENCES

ABRAMOVICH, F., BENJAMINI, Y., DONOHO, D. L. and JOHNSTONE, I. M. (2006). Adapting to unknown sparsity by controlling the false discovery rate. *Ann. Statist.* **34** 584–653. MR2281879 https://doi.org/10.1214/009053606000000074

AMINI, A. A. and WAINWRIGHT, M. J. (2008). High-dimensional analysis of semidefinite relaxations for sparse principal components. In *Information Theory*, 2008. *ISIT* 2008. *IEEE International Symposium on* 2454–2458. IEEE.

BERTHET, Q. and RIGOLLET, P. (2013). Computational lower bounds for sparse pca. Preprint. Available at arXiv:1304.0828.

BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469 https://doi.org/10.1214/08-AOS620

BIRNBAUM, A., JOHNSTONE, I. M., NADLER, B. and PAUL, D. (2013). Minimax bounds for sparse PCA with noisy high-dimensional data. *Ann. Statist.* **41** 1055–1084. MR3113803 https://doi.org/10.1214/12-AOS1014

CAI, T. T., MA, Z. and WU, Y. (2013). Sparse PCA: Optimal rates and adaptive estimation. *Ann. Statist.* **41** 3074–3110. MR3161458 https://doi.org/10.1214/13-AOS1178

CAI, T. T. and ZHOU, H. H. (2012). Optimal rates of convergence for sparse covariance matrix estimation. *Ann. Statist.* **40** 2389–2420. MR3097607 https://doi.org/10.1214/12-AOS998

CANDES, E. and TAO, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.* **35** 2313–2351. MR2382644 https://doi.org/10.1214/009053606000001523

CHEN, C.-H. and LI, K.-C. (1998). Can SIR be as popular as multiple linear regression? *Statist. Sinica* **8** 289–316. MR1624402

COOK, R. D., FORZANI, L. and ROTHMAN, A. J. (2012). Estimating sufficient reductions of the predictors in abundant high-dimensional regressions. *Ann. Statist.* **40** 353–384. MR3014310 https://doi.org/10.1214/11-AOS962

DENNIS COOK, R. (1998). *Regression Graphics*: *Ideas for Studying Regressions Through Graphics*. *Wiley Series in Probability and Statistics*: *Probability and Statistics*. Wiley, New York. MR1645673 https://doi.org/10.1002/9780470316931

DENNIS COOK, R. (2000). Save: A method for dimension reduction and graphics in regression. *Comm. Statist. Theory Methods* **29** 2109–2121.

DENNIS COOK, R. and WEISBERG, S. (1991). Comment. *J. Amer. Statist. Assoc.* **86** 328–332.

DUAN, N. and LI, K.-C. (1991). Slicing regression: A link-free regression method. *Ann. Statist.* **19** 505–530. MR1105834 https://doi.org/10.1214/aos/1176348109

FERRÉ, L. (1998). Determining the dimension in sliced inverse regression and related methods. *J. Amer. Statist. Assoc.* **93** 132–140. MR1614604 https://doi.org/10.2307/2669610

HSING, T. and CARROLL, R. J. (1992). An asymptotic theory for sliced inverse regression. *Ann. Statist.* **20** 1040–1061. MR1165605 https://doi.org/10.1214/aos/1176348669

JOHNSTONE, I. M. and LU, A. Y. (2004). *Sparse Principal Components Analysis*.

JUNG, S. and MARRON, J. S. (2009). PCA consistency in high dimension, low sample size context. *Ann. Statist.* **37** 4104–4130. MR2572454 https://doi.org/10.1214/09-AOS709

LI, K.-C. (1991). Sliced inverse regression for dimension reduction. *J. Amer. Statist. Assoc.* **86** 316–342. MR1137117

LI, K.-C. (2000). High dimensional data analysis via the SIR/PHD approach.

LI, L. (2007). Sparse sufficient dimension reduction. *Biometrika* **94** 603–613. MR2410011 https://doi.org/10.1093/biomet/asm044

LI, L. and NACHTSHEIM, C. J. (2006). Sparse sliced inverse regression. *Technometrics* **48** 503–510. MR2328619 https://doi.org/10.1198/004017006000000129

LI, B. and WANG, S. (2007). On directional regression for dimension reduction. *J. Amer. Statist. Assoc.* **102** 997–1008. MR2354409 https://doi.org/10.1198/016214507000000536

LIN, Q., ZHAO, Z. and LIU, J. S. (2018). On consistency and sparsity for sliced inverse regression in high dimensions. *Ann. Statist.* **46** 580–610. MR3782378 https://doi.org/10.1214/17-AOS1561

LIN, Q., ZHAO, Z. and LIU, J. S. (2019). Sparse sliced inverse regression via Lasso. *J. Amer. Statist. Assoc.* **114** 1726–1739. MR4047295 https://doi.org/10.1080/01621459.2018.1520115

LIN, Q., LI, X., HUANG, D. and LIU, J. S. (2021). Supplement to "On the optimality of sliced inverse regression in high dimensions." https://doi.org/10.1214/19-AOS1813SUPP

NEYKOV, M., LIN, Q. and LIU, J. S. (2016). Signed support recovery for single index models in high-dimensions. *Ann. Math. Sci. Appl.* **1** 379–426. MR3876481 https://doi.org/10.4310/amsa.2016.v1.n2.a5

RASKUTTI, G., WAINWRIGHT, M. J. and YU, B. (2011). Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Trans. Inf. Theory* **57** 6976–6994. MR2882274 https://doi.org/10.1109/TIT.2011.2165799

SCHOTT, J. R. (1994). Determining the dimensionality in sliced inverse regression. *J. Amer. Statist. Assoc.* **89** 141–148. MR1266291

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

VU, V. Q. and LEI, J. (2012). Minimax rates of estimation for sparse pca in high dimensions. Preprint. Available at arXiv:1202.0786.

YANG, S. S. (1977). General distribution theory of the concomitants of order statistics. *Ann. Statist.* **5** 996–1002. MR0501519

ZHU, L., MIAO, B. and PENG, H. (2006). On sliced inverse regression with high-dimensional covariates. *J. Amer. Statist. Assoc.* **101** 630–643. MR2281245 https://doi.org/10.1198/016214505000001285

ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 301–320. MR2137327 https://doi.org/10.1111/j.1467-9868.2005.00503.x