

ON THE OPTIMALITY OF SLICED INVERSE REGRESSION IN HIGH DIMENSIONS

BY QIAN LIN^{§*} XINRAN LI^{‡*} DONGMING HUANG^{‡*} AND JUN S. LIU^{‡*}

Tsinghua University[§] Harvard University[‡]

The central subspace of a pair of random variables $(y, \mathbf{x}) \in \mathbb{R}^{p+1}$ is the minimal subspace \mathcal{S} such that $y \perp \mathbf{x} | P_{\mathcal{S}} \mathbf{x}$. In this paper, we consider the minimax rate of estimating the central space under the multiple index model $y = f(\beta_1^T \mathbf{x}, \beta_2^T \mathbf{x}, \dots, \beta_d^T \mathbf{x}, \epsilon)$ with at most s active predictors, where $\mathbf{x} \sim N(0, \Sigma)$ for some class of Σ . We first introduce a large class of models depending on the smallest non-zero eigenvalue λ of $\text{var}(\mathbb{E}[\mathbf{x}|y])$, 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), i.e., the ‘*large p , small n problem*’. For these problems, variable selection and dimension reduction are often the indispensable first steps. In 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:

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

and estimates the subspace $\mathcal{S} = \text{span}\{ \beta_1, \dots, \beta_d \}$ via an eigen-analysis of the estimated conditional covariance matrix $\text{var}[\mathbb{E}(\mathbf{x}|y)]$. Note that the individual β_i ’s are not identifiable, but the space \mathcal{S} can be estimated well. Based on the observation that $y \perp \mathbf{x} | P_{\mathcal{S}} \mathbf{x}$, Cook [1998] proposed a more

*Lin’s research is supported in part by the Center for Statistical Science at Tsinghua University. Liu’s research is supported in part by the NSF Grant DMS-1120368 and NIH Grant R01 GM113242-01

Keywords and phrases: sufficient dimension reduction, optimal rates, sliced inverse regression, semi-definite positive programming

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 $y \perp\!\!\!\perp \mathbf{x} \mid P_{\mathcal{S}'\mathbf{x}}$, 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 [Cook, 2000, Li and Wang, 2007, Li, 2000]. The ‘large p , small n ’ characteristic of modern data raises new challenges to these SDR algorithms.

Lin et al. [2018c] 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, i.e., 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 in Lin et al. [2018c] however, 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 et al. [2016] considered the (signed)-support recovery problem of the following class of single index models:

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

where $\mathbf{x} \sim N(0, \mathbf{I}_p)$, $\epsilon \sim N(0, 1)$. 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 et al. [2018c] or Algorithm 1 below) can recover the (signed) support with probability converging to 1 as $n \rightarrow \infty$. That is, the

minimal sample size required to recover the support of β 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, Johnstone and Lu, 2004, Cai et al., 2013, Birnbaum et al., 2013, 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 [Cai et al., 2013, Vu and Lei, 2012, Birnbaum et al., 2013] for SIR when $\Sigma = \mathbf{I}$. However, as was pointed out by Lin et al. [2018c], 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, i.e., the ‘key lemma’ in Lin et al. [2018c]. 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 et al. [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 $\mathbb{E}[\|P_{\mathbf{V}} - P_{\hat{\mathbf{V}}}\|_F^2]$, where \mathbf{V} is an orthogonal matrix formed by an orthonormal basis of \mathcal{S} , and $P_{\hat{\mathbf{V}}}$ is an estimate of $P_{\mathbf{V}}$, the projection matrix associated with the orthogonal matrix \mathbf{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 Subsection 2.7, we show that the computationally efficient algorithm DT-SIR [Lin et al., 2018c] 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.

The rest of the paper is organized as follows. Section 2 presents the main

results of the paper, including the rate of the *oracle risk* in Section 2.4.1 and the rate of the *sparse risk* in Section 2.4.2. Since the lower bound can be obtained by modifying some standard arguments, we defer its related proofs to the online supplementary file [Lin et al., 2018a] and give the proofs of upper bounds in Sections 4.1 and 5. In Section 5 we discuss potential extensions of our results. More auxiliary results and technical lemmas are included in the online supplementary file [Lin et al., 2018a].

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 and Weisberg, 1991, Cook, 1998, Schott, 1994, Ferré, 1998, Li and Wang, 2007, Hsing and Carroll, 1992, Cook et al., 2012]. 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) [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 notations throughout the article. For a matrix \mathbf{V} , we denote its column space by $\text{col}(\mathbf{V})$ and its i -th row and j -th column by $\mathbf{V}_{i,*}$ and $\mathbf{V}_{*,j}$ respectively. For vectors \mathbf{x} and $\boldsymbol{\beta} \in \mathbb{R}^p$, we denote the k -th entry of \mathbf{x} as $\mathbf{x}(k)$ and the inner product $\langle \mathbf{x}, \boldsymbol{\beta} \rangle$ as $\mathbf{x}(\boldsymbol{\beta})$. For two positive number 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 , $\|A\|_F = \text{tr}(AA^\tau)^{1/2}$. For a positive integer p , $[p]$ denotes the index set $\{1, 2, \dots, p\}$. 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 β_i 's in model (1), without loss of generality, we can assume that $\mathbf{V} = (\beta_1, \dots, \beta_d)$ is a $p \times d$ orthogonal matrix (i.e., $\mathbf{V}^\top \mathbf{V} = \mathbf{I}_d$) and the models considered in this paper are

$$(2) \quad y = f(\mathbf{V}^\top \mathbf{x}, \epsilon), \quad \mathbf{V} \in \mathbb{O}(p, d),$$

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

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

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

$$(3) \quad \hat{\mathbf{\Lambda}}_H = \frac{1}{H} \sum_{h=1}^H \bar{\mathbf{x}}_{h,\cdot} \bar{\mathbf{x}}_{h,\cdot}^\top,$$

to estimate $\mathbf{\Lambda} \triangleq \text{var}(\mathbb{E}[\mathbf{x}|y])$, and $\text{col}(\hat{\mathbf{V}}_H)$ to estimate the central space $\text{col}(\mathbf{V})$, where $\hat{\mathbf{V}}_H$ is the matrix formed by the top d eigenvectors of $\hat{\mathbf{\Lambda}}_H$. 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 [Li, 1991, Hsing and Carroll, 1992, Zhu et al., 2006] in addition to the ‘‘linearity condition’’ that is automatically satisfied for Gaussian \mathbf{x} :

(A') Coverage Condition:

$$\text{span}\{\mathbb{E}[\mathbf{x}|y]\} = \text{span}\{\mathbf{V}_{*,1}, \dots, \mathbf{V}_{*,d}\}$$

where $\mathbf{V}_{*,i}$ is the i -th columns of the orthogonal matrix \mathbf{V} .

¹To ease notations and arguments, we assume that $n = cH$.

(B') Smoothness and Tail conditions on the Central Curve $\mathbb{E}[\mathbf{x}|y]$.

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

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

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

$$(4) \quad \begin{aligned} & \tilde{m}^4(y)P(|Y| > y) \rightarrow 0 \text{ as } y \rightarrow \infty \\ & \|\mathbf{m}(y) - \mathbf{m}(y')\|_2 \leq |\tilde{m}(|y|) - \tilde{m}(|y'|)| \text{ for } y, y' \in (-\infty, -B_0) \text{ or } y, y' \in (B_0, \infty). \end{aligned}$$

As in Lin et al. [2018c], where they demonstrated the phase transition phenomenon of SIR in high dimension, we replace Condition (B') by

(B'') Modified Smoothness and Tail conditions.

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

$$(5) \quad \begin{aligned} & \mathbb{E}[\tilde{m}(y)^4 1_{|y| > B_0}] < \infty \\ & \|\mathbf{m}(y) - \mathbf{m}(y')\|_2 \leq |\tilde{m}(|y|) - \tilde{m}(|y'|)| \text{ for } y, y' \in (-\infty, -B_0) \text{ or } y, y' \in (B_0, \infty). \end{aligned}$$

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 et al. [2016].

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

DEFINITION 1 (Sliced Stable Condition). *Let Y be a random variable. For $0 < \gamma_1 < 1 < \gamma_2$, let $\mathcal{A}_H(\gamma_1, \gamma_2)$ denote all partitions $\{-\infty = a_0 \leq a_2 \leq \dots \leq a_H = +\infty\}$ of \mathbb{R} , such that*

$$\frac{\gamma_1}{H} \leq \mathbb{P}(a_h \leq Y \leq a_{h+1}) \leq \frac{\gamma_2}{H}.$$

A curve $\mathbf{m}(y)$ is ϑ -sliced stable with respect to Y , if there exist positive constants $\gamma_1, \gamma_2, \gamma_3$ such that for sufficiently large H , for any partition $\in \mathcal{A}_H(\gamma_1, \gamma_2)$ and any $\beta \in \mathbb{R}^p$, we have

$$(6) \quad \frac{1}{H} \sum_{h=1}^H \text{var}(\beta^\tau \mathbf{m}(Y) | a_{h-1} \leq Y < a_h) \leq \frac{\gamma_3}{H^\vartheta} \text{var}(\beta^\tau \mathbf{m}(Y)).$$

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

Intuitively, $H \rightarrow \infty$ 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 $\hat{\mathbf{\Lambda}}_H$. This is one of the main technical contributions of Lin et al. [2018c]. We henceforth assume that the central curve satisfies the sliced stable condition. As shown by Proposition 1, Condition (\mathbf{B}'') ensures the sliced-stable condition.

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

Inspired by the assumption on the condition number in Cai et al. [2013], we consider the following condition

$$(7) \quad \lambda \leq \lambda_d(\text{var}(\mathbb{E}[\mathbf{x}|y])) \leq \lambda_1(\text{var}(\mathbb{E}[\mathbf{x}|y])) \leq \kappa \lambda \leq 1$$

for some positive constant $\kappa > 1$, which is a refinement of the coverage condition, i.e., $\text{rank}(\text{var}(\mathbb{E}[\mathbf{x}|y])) = d$. Without loss of generality, we assume thereafter $\lambda \leq 1/2$. Since $\mathbf{\Lambda} \triangleq \text{var}(\mathbb{E}[\mathbf{x}|y]) = \mathbf{V} \mathbf{\Lambda}_z \mathbf{V}^\tau$, we know $\lambda_j(\mathbf{\Lambda}) = \lambda_j(\mathbf{\Lambda}_z)$, $j = 1, \dots, d$. In particular, we have $\lambda \leq \lambda_d(\text{var}(\mathbb{E}[\mathbf{z}|y])) \leq \lambda_1(\text{var}(\mathbb{E}[\mathbf{z}|y])) \leq \kappa \lambda \leq 1$, 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 $\mathbf{z} \sim N(0, \mathbf{I}_d)$ and $\epsilon \sim N(0, 1)$. A function $f(\mathbf{z}, \epsilon)$ belongs to the class $\mathcal{F}_d(\lambda, \kappa)$ if the following conditions hold:

- (A) Coverage condition: $0 < \lambda \leq \lambda_d(\mathbf{\Lambda}_z) \leq \dots \leq \lambda_1(\mathbf{\Lambda}_z) \leq \kappa \lambda \leq 1$, where $\mathbf{\Lambda}_z \triangleq \text{var}(\mathbb{E}[\mathbf{z}|f(\mathbf{z}, \epsilon)])$.
- (B) Sliced stable condition: $\mathbf{m}_z(y) = \mathbb{E}[\mathbf{z}|f(\mathbf{z}, \epsilon) = y]$ is sliced stable with respect to $f(\mathbf{z}, \epsilon)$.

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

\mathbf{V} , that is, $y = f(\mathbf{V}^\top \mathbf{x}, \epsilon)$. We are interested in the risk $\mathbb{E}_{\mathcal{M}} \|P_{\hat{\mathbf{V}}} - P_{\mathbf{V}}\|_F^2$ where $P_{\hat{\mathbf{V}}}$ is an estimate of $P_{\mathbf{V}}$ based on these samples. In this subsection, we provide an upper bound on this risk. All detailed proofs are deferred to Section 4 and online supplementary file [Lin et al., 2018a].

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

$$(8) \quad \mathfrak{M}(p, d, \lambda, \kappa) \triangleq \left\{ (\mathbf{V}, f) \mid \mathbf{V} \in \mathbb{O}(p, d), f \in \mathcal{F}_d(\lambda, \kappa) \right\}.$$

We refer to the risk over \mathfrak{M} as the ‘Oracle risk’. The first main result of this article is:

THEOREM 1 (An Upper Bound on the Minimax Oracle Risk.). *Assuming that $\frac{dp}{n\lambda}$ is sufficiently small and $d^2 \leq p$, we have*

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

We will show that the estimate $P_{\hat{\mathbf{V}}_H}$ achieves the rate in Theorem 1 where $\hat{\mathbf{V}}_H$ is the $p \times d$ orthogonal matrix forming by the top- d eigenvectors of $\hat{\mathbf{\Lambda}}_H$ (See Eq. (3)). This appears to contradict a result in Lin et al. [2018c], which states that

$$(10) \quad \|\hat{\mathbf{\Lambda}}_H - \text{var}(\mathbb{E}[\mathbf{x}|y])\|_2 = O_P \left(\frac{1}{H^\vartheta} + \frac{H^2 p}{n} + \sqrt{\frac{H^2 p}{n}} \right).$$

Lin et al. [2018c] indicates that the convergence rate (i) does not depend on d , the dimension of central subspace; (ii) does not depend on λ , the smallest non-zero eigenvalue of $\text{var}(\mathbb{E}[\mathbf{x}|y])$; (iii) depends on H (the number of slices) and seems worse than our upper bound here. The first two differences appear simply because Lin et al. [2018c] have assumed that d is bounded and the non-zero eigenvalues of $\text{var}(\mathbb{E}[\mathbf{x}|y])$ 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 $\text{var}(\mathbb{E}[\mathbf{x}|y])$. As they have pointed out, the convergence rate of $\hat{\mathbf{\Lambda}}_H$ might be different (slower) than the convergence rate of $P_{\hat{\mathbf{V}}_H}$. More precisely, we have

$$(11) \quad \hat{\mathbf{\Lambda}}_H - \mathbf{\Lambda} = \left(\hat{\mathbf{\Lambda}}_H - P_{\mathbf{V}} \hat{\mathbf{\Lambda}}_H P_{\mathbf{V}} \right) + \left(P_{\mathbf{V}} \hat{\mathbf{\Lambda}}_H P_{\mathbf{V}} - \mathbf{\Lambda} \right).$$

From the proof of Theorem 1 of [Lin et al. \[2018c\]](#), 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 $P_{\mathbf{V}}\hat{\mathbf{\Lambda}}_H P_{\mathbf{V}}$ and $\mathbf{\Lambda}$ share the same column space and we are interested in estimating $P_{\mathbf{V}}$, 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 et al. \[2018c\]](#).

REMARK 1. *On the role of H .* Researchers have claimed that the performance of SIR procedure is not sensitive to the choice of H , i.e., 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 $\text{col}(\hat{\mathbf{V}}_H)$ is an optimal estimate of $\text{col}(\mathbf{V})$. In the SIR literature, researchers care about the eigenvectors of $\mathbf{\Lambda}$ 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 $\mathbf{\Lambda}$, and illustrate this phenomenon numerically in Section 3.1.

2.4.2. *Upper bound on the risk of sparse SIR.* [Lin et al. \[2018c\]](#) 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 $\mathbf{V}_{*,1}, \dots, \mathbf{V}_{*,d}$. For a $p \times d$ orthogonal matrix \mathbf{V} (i.e., $\mathbf{V}^T \mathbf{V} = \mathbf{I}_d$), we order the row norms in decreasing order as $\|\mathbf{V}_{(1),*}\|_2 \geq \dots \geq \|\mathbf{V}_{(p),*}\|_2$ and define the weak l_q radius of \mathbf{V} to be

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

Let $\mathbb{O}_{s,q}(p, d) = \{\mathbf{V} \mid \mathbf{V} \in \mathbb{O}(p, d) \text{ such that } \|\mathbf{V}\|_{q,w} \leq s\}$ 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 et al. \[2012\]](#) for sparse co-variance matrix estimation. Furthermore, we need the notion of *effective support*, which was introduced by [Cai et al. \[2013\]](#). The size of *effective support* is defined to be $k_{q,s} \triangleq \lceil x_q(s, d) \rceil$, where

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

and $\lceil a \rceil$ denotes the smallest integer no less than $a \in \mathbb{R}$. See [Cai et al. \[2013\]](#) 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 (\mathbf{V}, f) :

$$(14) \quad \mathfrak{M}_{s,q}(p, d, \lambda, \kappa) \triangleq \left\{ (\mathbf{V}, f) \mid \mathbf{V} \in \mathbb{O}_{s,q}(p, d), f \in \mathcal{F}_d(\lambda, \kappa) \right\}.$$

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

THEOREM 2 (The Upper Bound on Optimal Rates). *Assume that κ is fixed, $d^2 \leq k_{q,s}$ and ϵ_n^2 is sufficiently small. We have*

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

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 $\hat{\mathbf{\Lambda}}_H^{(1)}$ and $\hat{\mathbf{\Lambda}}_H^{(2)}$ be the SIR estimates of $\mathbf{\Lambda} = \text{var}(\mathbb{E}[\mathbf{x}|y])$ based on the first and second sets of samples, respectively. Inspired by the idea in Cai et al. [2013], we introduce the following aggregation estimator $\hat{\mathbf{V}}_E$ of \mathbf{V} .

Aggregation Estimator $\hat{\mathbf{V}}_E$. For each $B \in \mathcal{B}_k$, we let

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

and

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

Our aggregation estimator $\hat{\mathbf{V}}_E$ is defined to be $\hat{\mathbf{V}}_{B^*}$.

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

$$(17) \quad \langle \hat{\mathbf{\Lambda}}_H^{(2)}, \hat{\mathbf{V}}_E \hat{\mathbf{V}}_E^\tau - \hat{\mathbf{V}}_B \hat{\mathbf{V}}_B^\tau \rangle \geq 0$$

for any $\hat{\mathbf{V}}_B$ where $B \in \mathcal{B}$. In Section 5, we will show that the aggregation estimator $\hat{\mathbf{V}}_E$ 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 4-th direction was difficult to discover. We have also observed from extensive numerical studies that the 4-th 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 semi-parametric 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 non-zero eigenvalue of $\text{var}(\mathbb{E}[\mathbf{x}|y])$, is 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 non-zero eigenvalues of $\text{var}(\mathbb{E}[\mathbf{x}|y])$, 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. For $\frac{dp}{n}$ sufficiently small, we have*

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

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 file [Lin et al., 2018a] 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 et al. [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

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

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 \mathbf{B}_j such that the column space of \mathbf{B}_j corresponds to $\theta_j \in \mathbb{G}(p, d)$, we may consider the following finite class of models

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

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 (c.f. Lemma 15). Let $p_{f, \mathbf{B}}$ denote the joint density of (y, \mathbf{x}) . Simple calculation shows (c.f. Lemma 14) that

$$(19) \quad KL(p_{f, \mathbf{B}_1}, p_{f, \mathbf{B}_2}) \leq C \left(\max \|\nabla f\|^2 \right) \|\mathbf{B}_1 - \mathbf{B}_2\|_F^2 \leq C \|\mathbf{B}_1 - \mathbf{B}_2\|_F^2.$$

If we have

$$(20) \quad \|\mathbf{B}_1 - \mathbf{B}_2\|_F^2 \leq \|P_{\mathbf{B}_1} - P_{\mathbf{B}_2}\|_F^2,$$

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

However, (20) is not always true (e.g., it fails if \mathbf{B}_1 and \mathbf{B}_2 are two different orthogonal matrices sharing the same column space). We need to carefully specify \mathbf{B}_j for each $\theta_j \in \Theta \subset \mathbb{G}(p, d)$ such that they satisfy the inequality (20) (c.f. Lemma 22 in the supplementary file). 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 et al. [2013] leads us the following:

THEOREM 4 (Optimal Rates). *Assume that d is bounded. For sufficiently large constant κ and sufficiently small constant λ , we have*

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

PROOF. See the online supplementary file [Lin et al., 2018a].

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 $d = 1$, we have*

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

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:

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

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

$$\text{var}(\mathbb{E}[\boldsymbol{\beta}^\tau \mathbf{x} | y]) = \frac{2\lambda}{1 + 2\lambda} \geq \lambda \text{ and } \nabla f_\lambda \leq C\sqrt{\lambda}.$$

Thus, inequality (19) becomes

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

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 et al. [2013] leads us to the following result:

THEOREM 6 (Optimal Rates : $d = 1$). *Assume that $d = 1$. We have*

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

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 subsection 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 $0 < \lambda \leq 1$, there exists a d -variate function f_λ such that $f_\lambda(x_1, \dots, x_d) + x_{d+1} \in \mathcal{F}_d(\lambda, \kappa)$ and*

$$(25) \quad \|\nabla f_\lambda(x_1, \dots, x_d)\| \leq C\sqrt{\lambda}.$$

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

$$C_1\lambda \leq \lambda_d(\text{var}(\mathbb{E}_\lambda[\mathbf{x}|y])) \leq \lambda_1(\text{var}(\mathbb{E}_\lambda[\mathbf{x}|y])) \leq C_2\kappa\lambda.$$

Note that the density function $p(y)$ of y is the convolution of the density functions of ϵ and $\sqrt{\lambda}f(\mathbf{x})$. Heuristically, if $f(\mathbf{x})$ is (nearly) normal, by the continuity of the convolution operator, we expect that $\lambda_d(\text{var}(\mathbb{E}[\mathbf{x}|y])) \asymp \lambda$. Since we cannot prove it rigorously, we present some supporting numerical evidences in Subsection 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 d is bounded and Conjecture 1 holds, we have*

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

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

THEOREM 8 (Optimal Rates). *Assuming that d is fixed and Conjecture 1 holds, we have*

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

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 $\mathbf{x} \sim N(0, \mathbf{\Sigma})$) 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 $\mathbf{\Sigma}$ with bounded eigenvalues [Raskutti et al., 2011]; and the minimax rate for

sparse PCA is derived only for spiked models where the irrelevant noises are uncorrelated [Cai et al., 2013]. Because the semi-parametric 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 $\mathfrak{M}_s(p, d, \lambda, \kappa, \Sigma)$ consisting of models $y = f(\mathbf{\Gamma}^\tau \mathbf{x}, \epsilon)$, $\mathbf{x} \sim N(0, \Sigma)$ and $\epsilon \sim N(0, 1)$ where the covariance matrix Σ and the $p \times d$ orthogonal matrix $\mathbf{\Gamma}$ satisfying the following condition

$$(28) \quad \|J_K \mathbf{\Gamma}\|_F \leq C \|J_K \Sigma \mathbf{\Gamma}\|_F$$

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

- G1)** Σ has at most k non-zero entries in each row where k is a fixed integer and $C_1 \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq C_2$ for some constants C_1 and C_2 .
- G2)** $\mathbb{E}[\mathbf{x} | y]$ satisfying the sliced stability condition.
- G3)** $0 < \lambda \leq \lambda_d(\text{var}(\mathbb{E}[\mathbf{x} | y])) \leq \dots \leq \lambda_1(\text{var}(\mathbb{E}[\mathbf{x} | y])) \leq \kappa \lambda$ for some constant κ .
- G4)** $|\text{supp}(\mathbf{\Gamma})| \leq s$.

Then, we have the following result:

THEOREM 9 (Optimal Rates). *Assuming that d is fixed, $s = o(p)$, Conjecture 1, and conditions **G1)**- **G4)** hold, we have*

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

PROOF. A sketch of proof is presented in the supplemental file.

In a recent work [Lin et al., 2018b], we proposed the Lasso-SIR algorithm to estimate the central space for general Σ 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 , $1 \leq j \leq p$ and i , $1 \leq i \leq d$, we have

$$(A) \quad \sum_i |\langle \Sigma_j, \mathbf{\Gamma}_i \rangle|^2 \geq C \sum_i |\mathbf{\Gamma}_i(j)|^2 \|\Sigma_j\|_2^2,$$

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

In fact, let K be any integer $\in \{1, 2, \dots, p\}$. We know that Condition (28) implies Condition (A). On the other hand, assuming that condition (A) holds for any $1 \leq j \leq p$. Now, for any $K \subset [p]$, we have

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

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

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 $\frac{\pi}{2}$, i.e., there exists a constant C such that $|\langle \mathbf{\Sigma}_j, \mathbf{\Gamma}_i \rangle| \geq C \|\mathbf{\Gamma}_i\|_2 \|\mathbf{\Sigma}_j\|_2$, then condition (A) holds since $\|\mathbf{\Gamma}_i\|_2 \geq \|\mathbf{\Gamma}_i(j)\|_2$.

2.7. Optimality of DT-SIR. In the previous section, we have proved that the aggregation estimator $\hat{\mathbf{V}}_E$ is rate optimal. In practice, however, it is computationally too expensive. The DT-SIR algorithm proposed in Lin et al. [2018c] is computationally efficient in general, and can be further simplified when $\mathbf{\Sigma}_x = \mathbf{I}$. In this section, we focus on the single index models with the exact sparsity on the loading vector β , i.e., $|\text{supp}(\beta)| = s$.

Algorithm 1 DT-SIR for Single-Index Models

- 1: Let $S = \{i \mid \hat{\Lambda}_H(i, i) > t\}$ for a properly chosen t .
 - 2: Let $\hat{\beta}$ be the principal eigenvector of $\hat{\Lambda}_H(S, S)$.
 - 3: We embed $\hat{\beta}$ into \mathbb{R}^p by filling the entries outside S_t with 0 and denote it by $\hat{\beta}_{DT}$.
-

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 $\hat{\beta}_{DT}$ be the DT-SIR estimate with threshold level $t = C_1 \frac{\log(p)}{n}$ for some constant C_1 , then we have

$$(30) \quad \mathbb{E}_\beta \|P_{\hat{\beta}_{DT}} - P_\beta\|^2 \leq C_2 \frac{s \log(p-s)}{n\lambda}.$$

PROOF. See the online supplementary file [Lin et al., 2018a].

From Theorem 10, it is easy to see that, if $s = O(p^{1-\delta})$, the DT-SIR estimator $P_{\hat{\beta}_{DT}}$ 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 et al., 2018b]. 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 $\hat{\mathbf{\Lambda}}$ can be treated as a generalization of the square of $\mathbb{E}[yx_i]$.

EXAMPLE 1. *Consider the simple model $y = a_1x_1 + a_2x_2 + \epsilon$ where $x_1, x_2 \sim N(0, 1)$ and $\epsilon \sim N(0, 1)$. It is easy to show that*

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

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

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 $\text{var}(\mathbb{E}[\mathbf{x}|y])$. 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 $\text{var}(\mathbb{E}[\mathbf{x}|y])$ 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

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

It is easy to see that the only non-zero eigenvalue of $\text{var}(\mathbb{E}[\mathbf{x}|y])$ is μ . The results are shown in Table 1, where H ranges in $\{2, 5, 10, 50, 100, 200, 500\}$, μ in $\{.5, .3, .1\}$ and n in $\{5000, 10000, 50000, 100000\}$. Each entry is the empirical mean (standard deviation), calculated based on 100 replications, of the SIR estimate of $\hat{\mu}$ for given μ, n and H .

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

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

	$n/1000$	$H = 2$	$H = 5$	$H = 10$	$H = 50$	$H = 100$	$H = 200$	$H = 500$
$\mu = .5$	5	0.319 (0.013)	0.446 (0.017)	0.479 (0.017)	0.503 (0.016)	0.509 (0.017)	0.520 (0.018)	0.551 (0.017)
	10	0.318 (0.009)	0.448 (0.012)	0.480 (0.012)	0.500 (0.012)	0.505 (0.012)	0.510 (0.013)	0.525 (0.012)
	50	0.319 (0.004)	0.448 (0.006)	0.479 (0.005)	0.498 (0.006)	0.500 (0.005)	0.501 (0.006)	0.504 (0.006)
	100	0.319 (0.003)	0.448 (0.004)	0.479 (0.004)	0.498 (0.004)	0.499 (0.004)	0.501 (0.004)	0.503 (0.004)
$\mu = .3$	5	0.190 (0.011)	0.271 (0.012)	0.288 (0.014)	0.307 (0.015)	0.313 (0.015)	0.328 (0.014)	0.371 (0.016)
	10	0.191 (0.008)	0.27 (0.009)	0.288 (0.010)	0.302 (0.010)	0.307 (0.010)	0.312 (0.010)	0.335 (0.012)
	50	0.191 (0.003)	0.269 (0.004)	0.288 (0.005)	0.299 (0.005)	0.3 (0.005)	0.302 (0.005)	0.307 (0.004)
	100	0.191 (0.002)	0.269 (0.003)	0.288 (0.003)	0.299 (0.004)	0.3 (0.003)	0.301 (0.003)	0.303 (0.003)
$\mu = .1$	5	0.064 (0.007)	0.091 (0.008)	0.098 (0.009)	0.109 (0.009)	0.117 (0.009)	0.136 (0.010)	0.190 (0.010)
	10	0.0643 (0.005)	0.0901 (0.006)	0.0973 (0.006)	0.103 (0.006)	0.108 (0.006)	0.117 (0.006)	0.144 (0.007)
	50	0.0638 (0.002)	0.0899 (0.003)	0.0963 (0.003)	0.101 (0.003)	0.101 (0.003)	0.103 (0.003)	0.109 (0.003)
	100	0.0636 (0.001)	0.0898 (0.002)	0.0961 (0.002)	0.100 (0.002)	0.100 (0.002)	0.102 (0.002)	0.104 (0.002)

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. Cautious reader may notice that, in the row with $\mu = .1$ and $n = 5000$, the empirical mean and the standard error are not behaving as we have expected, e.g., 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 $\hat{\lambda}$ depends on the value $\frac{n\mu}{H^2}$, i.e., the larger the $\frac{n\mu}{H^2}$ is, the more concentrated the $\hat{\lambda}$ 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., $d=2$):

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

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

$$(33) \quad g(x) = \begin{cases} x & \text{if } |x| \leq 100 - \delta \\ 0 & \text{if } |x| \geq 100 + \delta \end{cases}$$

and $|g'(x)| \leq C$ for some constant C . Let $\lambda_1(\mu)$ and $\lambda_2(\mu)$ be the two eigenvalues of $\text{var}(\mathbb{E}[\mathbf{x}|y])$. Since we know that the absolute value of the derivative of the link function $\leq C\sqrt{\mu}$, we want to check if $C_1\mu \leq \lambda_2(\mu) \leq \lambda_1(\mu) \leq C_2\mu$ 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 $\{1, .5, .1, .05, .01, .005, .001, 0.0005, 0.0001\}$ and the ratio $n\mu/H^2$ range in $\{2, 4, 10, 20\}$.

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

$n\mu/H^2$	$\mu = 1$	$\mu = .5$	$\mu = .1$	$\mu = .05$	$\mu = .01$	$\mu = .005$	$\mu = .001$	$\mu = .0005$	$\mu = .0001$
2	0.352 (0.002)	0.579 (0.004)	1.438 (0.014)	1.732 (0.013)	1.978 (0.014)	1.999 (0.011)	2.009 (0.010)	1.985 (0.009)	1.994 (0.007)
4	0.341 (0.001)	0.579 (0.002)	1.415 (0.005)	1.697 (0.005)	1.949 (0.006)	1.969 (0.006)	1.991 (0.005)	1.982 (0.005)	1.977 (0.006)
10	0.333 (4.6e-4)	0.571 (1.0e-3)	1.409 (1.8e-3)	1.706 (2.2e-3)	1.944 (2.4e-3)	1.970 (2.6e-3)	1.970 (1.4e-3)	1.964 (1.7e-3)	1.971 (1.8e-3)
20	0.329 (3.2e-4)	0.565 (3.4e-4)	1.412 (1.0e-3)	1.702 (1.5e-3)	1.950 (1.5e-3)	1.971 (1.2e-3)	1.963 (8.9e-4)	1.968 (1.1e-3)	1.969 (1.1e-3)

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

$n\mu/H^2$	$\mu = 1$	$\mu = .5$	$\mu = .1$	$\mu = .05$	$\mu = .01$	$\mu = .005$	$\mu = .001$	$\mu = .0005$	$\mu = .0001$
2	0.110 (4.5e-4)	0.127 (6.4e-4)	0.126 (6.1e-4)	0.099 (3.7e-4)	0.051 (1.9e-4)	0.037 (1.1e-4)	0.027 (6.0e-5)	0.025 (7.1e-5)	0.024 (6.3e-5)
4	0.097 (2.6e-4)	0.114 (2.4e-4)	0.109 (2.6e-4)	0.091 (2.2e-4)	0.039 (7.6e-5)	0.028 (5.2e-5)	0.014 (2.2e-5)	0.013 (2.0e-5)	0.012 (1.6e-5)
10	0.094 (1.0e-4)	0.111 (1.1e-4)	0.103 (1.1e-4)	0.083 (5.9e-5)	0.032 (2.9e-5)	0.021 (1.6e-5)	0.008 (4.9e-6)	0.006 (4.1e-6)	0.005 (2.4e-6)
20	0.092 (4.e-5)	0.108 (6.3e-5)	0.102 (4.9e-5)	0.080 (4.0e-5)	0.003 (1.3e-5)	0.002 (7.6e-6)	0.005 (2.0e-6)	0.004 (1.4e-6)	0.004 (6.9e-7)

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,

$$\text{Model 1 : } y = \mathbf{x}^\tau \boldsymbol{\beta} + \sin(\mathbf{x}^\tau \boldsymbol{\beta}) + \epsilon,$$

$$\text{Model 2 : } y = 2 \arctan(\mathbf{x}^\tau \boldsymbol{\beta}) + \epsilon,$$

$$\text{Model 3 : } y = (\mathbf{x}^\tau \boldsymbol{\beta})^3 + \epsilon,$$

$$\text{Model 4 : } y = \sinh(\mathbf{x}^\tau \boldsymbol{\beta}) + \epsilon,$$

where $\mathbf{x} \sim N(\mathbf{0}, \mathbf{I}_p)$, $\epsilon \sim N(0, 1)$, $\mathbf{x} \perp \epsilon$, 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 $\{3, 5, 7, \dots, 61\}$. 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 $\hat{\boldsymbol{\beta}}_{DT}$ using DT-SIR algorithm, and the results of the average values of $\|P_{\hat{\boldsymbol{\beta}}_{DT}} - P_{\boldsymbol{\beta}}\|^2$ 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 .

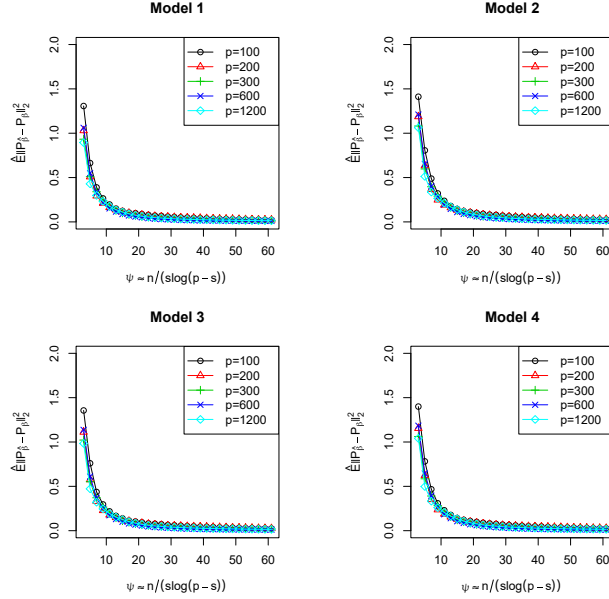


FIG 1. Average values of $\|P_{\hat{\boldsymbol{\beta}}_{DT}} - P_{\boldsymbol{\beta}}\|^2$

According to Theorem 10, $\psi \|P_{\hat{\boldsymbol{\beta}}_{DT}} - P_{\boldsymbol{\beta}}\|^2$ is less than a constant with high

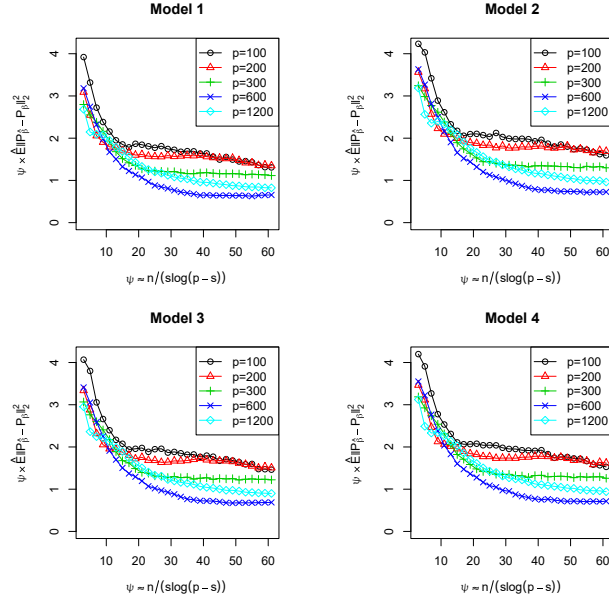


FIG 2. Average values of $\psi \|P_{\hat{\beta}_{DT}} - P_{\beta}\|^2$

probability. Therefore, we also display the average values of $\psi \|P_{\hat{\beta}_{DT}} - P_{\beta}\|^2$ for these models in Figure 2, which demonstrates that $\psi \|P_{\hat{\beta}_{DT}} - P_{\beta}\|^2$ 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 et al. [2018c]:

LEMMA 1. Assume that $f \in \mathcal{F}_d(\lambda, \kappa)$ in the model (2). Let $\hat{\Lambda}_H$ be the SIR estimate (3) of $\text{var}(\mathbb{E}[\mathbf{x}|y])(= \Lambda)$. 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 β that lies in the column space of Λ , we have

$$(34) \quad \left| \beta^\tau (\hat{\Lambda}_H - \Lambda) \beta \right| > \frac{1}{2\nu} \beta^\tau \Lambda \beta$$

with probability at most

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

In particular, 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 lemma’ in Lin et al. [2018c]. \square

Notations: Suppose that we have $n = Hc$ samples (y_i, \mathbf{x}_i) from the distribution defined by the model $\mathcal{M} = (\mathbf{V}, f) \in \mathfrak{M}(p, d, \kappa, \lambda)$. Let $H = H_1 d$ where H_1 is a sufficiently large integer and $\hat{\mathbf{V}} = (\hat{\mathbf{V}}_1, \dots, \hat{\mathbf{V}}_d)$ where $\hat{\mathbf{V}}_i$ is the eigen-vector associated to the i -th largest eigen-value of $\hat{\Lambda}_H$. We introduce the following decomposition

$$\mathbf{x} = P_{\mathcal{S}} \mathbf{x} + P_{\mathcal{S}^\perp} \mathbf{x} \triangleq \mathbf{z} + \mathbf{w},$$

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

$$(35) \quad \begin{aligned} \hat{\Lambda}_H &= \mathcal{Z} \mathcal{Z}^\tau + \mathcal{Z} \mathcal{W}^\tau + \mathcal{W} \mathcal{Z}^\tau + \mathcal{W} \mathcal{W}^\tau \\ &= \Lambda_u + \mathcal{Z} \mathcal{E}^\tau \mathbf{V}^{\perp, \tau} + \mathbf{V}^\perp \mathcal{E} \mathcal{Z}^\tau + \mathbf{V}^\perp \mathcal{E} \mathcal{E}^\tau \mathbf{V}^{\perp, \tau} \end{aligned}$$

where we define $\Lambda_u \triangleq \mathcal{Z} \mathcal{Z}^\tau$ and use the fact $\mathcal{W} = \mathbf{V}^\perp \mathcal{E}$. 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*

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

We will take $t = \max(p, \log(n\lambda/H^2))/n$ in the late argument.

ii) *For any vector $\beta \in \mathbb{R}^p$ and any $\nu > 1$, let $E_\beta(\nu) = \left\{ \left| \beta^\tau (\Lambda_u - \Lambda) \beta \right| > \frac{1}{2\nu} \beta^\tau \Lambda \beta \right\}$. Recall that $H = dH_1$. If we choose H_1 sufficiently large such that $H^\vartheta > C\nu$ for some positive constant C , there exist positive constants C_1, \dots, C_3 and C_4 such that*

$$\mathbb{P}\left(\bigcup_{\beta} E_\beta(\nu)\right) \leq C_1 \exp\left(-C_2 \frac{n\lambda}{H^2 \nu^2} + C_3 \log(H) + C_4 d\right).$$

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

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

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) \square

Let $\mathbf{E} = \mathbf{E}_1 \cap \mathbf{E}_2 \cap \mathbf{E}_3$ where $\mathbf{E}_1 = \left\{ \|\mathcal{W}\mathcal{W}^\tau\| \leq C\rho \right\}$, $\mathbf{E}_2 = \left\{ \|\mathcal{W}\mathcal{Z}^\tau\| \leq 4\sqrt{\kappa \lambda \rho} \right\}$, $\mathbf{E}_3 = \left\{ \|\Lambda_u - \Lambda\| \leq \frac{1}{2\nu} \kappa \lambda \right\}$.

COROLLARY 1. *Lemma 2 implies the following simple results where C stands for some absolute constant which might be varying in different statements.*

- a) $\mathbb{P}(\mathbf{E}^c) \leq \frac{CH^2}{n\lambda}$.
- b) *Conditioning on \mathbf{E}_3 , we have $\lambda_d(\Lambda_u) \geq (1 - \frac{\kappa}{2\nu})\lambda$.*

- c) *Conditioning on \mathbf{E} , if $\frac{p}{n\lambda}$ is sufficiently small, we have $\|\widehat{\mathbf{\Lambda}}_H - \mathbf{\Lambda}_u\| \leq C\sqrt{\frac{\kappa\lambda p}{n}}$.*
- d) *Conditioning on \mathbf{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

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

For II. It is easy to see that

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

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

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

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

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

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

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

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

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

□

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 $\text{var}(\mathbb{E}[\mathbf{x}|y])$, 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 $\text{var}(\mathbb{E}[\mathbf{x}|y])$ rather than that of the central space. Since we have assumed that $\Sigma = \mathbf{I}$ in this paper, the column space of $\text{var}(\mathbb{E}[\mathbf{x}|y])$ coincides with the central space in model (1). When there are correlations between predictors, if we assume that the eigenvectors associated with non-zero eigenvalues of $\text{var}(\mathbb{E}[\mathbf{x}|y])$ are sparse (with sparsity s) instead of assuming that the loading vectors β_i 's are sparse, our argument in this paper implies that $\mathbb{E}[\|P_{\text{col}(\widehat{\text{var}(\mathbb{E}[\mathbf{x}|y])})} - P_{\text{col}(\text{var}(\mathbb{E}[\mathbf{x}|y])})\|_F^2]$ 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 $y = \beta^T \mathbf{x} + \epsilon$ with $\mathbf{x} \sim N(0, \mathbf{I})$ and $s = O(p^{1-\delta})$, the minimax rate [Raskutti et al., 2011] of estimating β is achieved by the simple correlation screening. Analogously, the minimax rate for estimating P_β is achieved by the DT-SIR algorithm of Lin et al. [2018c], 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 Σ might be possible, since it is now well known that Lasso [Tibshirani, 1996] and Dantzig Selector [Candes et al., 2007] achieve the optimal rate of linear regression [Bickel et al., 2009] for general Σ . This speculation warrants further future investigations.

SUPPLEMENTARY MATERIAL

Supplement to “On the optimality of SIR in high dimensions”
(<http://www.e-publications.org/ims/support/download/imsart-ims.zip>).

References.

- Felix Abramovich, Yoav Benjamini, David L Donoho, and Iain M Johnstone. Special invited lecture: adapting to unknown sparsity by controlling the false discovery rate. *The Annals of Statistics*, pages 584–653, 2006.
- A. A. Amini and M. J. Wainwright. High-dimensional analysis of semidefinite relaxations for sparse principal components. In *Information Theory, 2008. ISIT 2008. IEEE International Symposium on*, pages 2454–2458. IEEE, 2008.
- Quentin Berthet and Philippe Rigollet. Computational lower bounds for sparse pca. *arXiv preprint arXiv:1304.0828*, 2013.
- Peter J Bickel, Ya’acov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, pages 1705–1732, 2009.
- Aharon Birnbaum, Iain M Johnstone, Boaz Nadler, and Debashis Paul. Minimax bounds for sparse pca with noisy high-dimensional data. *Annals of statistics*, 41(3):1055, 2013.
- T. Tony Cai, Harrison H. Zhou, et al. Optimal rates of convergence for sparse covariance matrix estimation. *The Annals of Statistics*, 40(5):2389–2420, 2012.
- T. Tony Cai, Zongming Ma, Yihong Wu, et al. Sparse pca: Optimal rates and adaptive estimation. *The Annals of Statistics*, 41(6):3074–3110, 2013.
- Emmanuel Candes, Terence Tao, et al. The dantzig selector: Statistical estimation when p is much larger than n . *The Annals of Statistics*, 35(6):2313–2351, 2007.
- Chun-Houh Chen and Ker-Chau Li. Can sir be as popular as multiple linear regression? *Statistica Sinica*, pages 289–316, 1998.
- R. Dennis Cook. *Regression graphics*. Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons, Inc., New York, 1998. ISBN 0-471-19365-8. . URL <http://dx.doi.org/10.1002/9780470316931>. Ideas for studying regressions through graphics, A Wiley-Interscience Publication.
- R. Dennis Cook. Save: a method for dimension reduction and graphics in regression. *Communications in statistics-Theory and methods*, 29(9-10):2109–2121, 2000.
- R. Dennis Cook and Sanford Weisberg. Comment. *Journal of the American Statistical Association*, 86(414):328–332, 1991.
- R Dennis Cook, Liliana Forzani, Adam J Rothman, et al. Estimating sufficient reductions of the predictors in abundant high-dimensional regressions. *The Annals of Statistics*, 40(1):353–384, 2012.
- Naihua Duan and Ker-Chau Li. Slicing regression: a link-free regression method. *The Annals of Statistics*, pages 505–530, 1991.
- Louis Ferré. Determining the dimension in sliced inverse regression and related methods. *Journal of the American Statistical Association*, 93(441):132–140, 1998.
- Chao Gao, Zongming Ma, Zhao Ren, and Harrison H Zhou. Minimax estimation in sparse canonical correlation analysis. *arXiv preprint arXiv:1405.1595*, 2014.
- Tailen Hsing and Raymond J Carroll. An asymptotic theory for sliced inverse regression. *The Annals of Statistics*, pages 1040–1061, 1992.
- I. M. Johnstone and A. Y. Lu. Sparse principal components analysis. 2004.
- Sungkyu Jung, JS Marron, et al. Pca consistency in high dimension, low sample size context. *The Annals of Statistics*, 37(6B):4104–4130, 2009.
- Bing Li and Shaoli Wang. On directional regression for dimension reduction. *Journal of the American Statistical Association*, 102(479):997–1008, 2007.
- Ker-Chau Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327, 1991.
- Ker-Chau Li. High dimensional data analysis via the sir/phd approach. 2000.
- Lexin Li. Sparse sufficient dimension reduction. *Biometrika*, 94(3):603–613, 2007.
- Lexin Li and Christopher J Nachtsheim. Sparse sliced inverse regression. *Technometrics*, 48(4):503–510, 2006.

- Qian Lin, Xinran Li, Dongming Huang, and Jun S. Liu. Supplementary to 'on optimality of sliced inverse regression in high dimensions'. 2018a.
- Qian Lin, Zhigen Zhao, and Jun S. Liu. Sparse sliced inverse regression via lasso. *Journal of the*, 2018b.
- Qian Lin, Zhigen Zhao, Jun S Liu, et al. On consistency and sparsity for sliced inverse regression in high dimensions. *The Annals of Statistics*, 46(2):580–610, 2018c.
- Zhuang Ma and Xiaodong Li. Subspace perspective on canonical correlation analysis: Dimension reduction and minimax rates. *arXiv preprint arXiv:1605.03662*, 2016.
- Matey Neykov, Qian Lin, and Jun S Liu. Signed support recovery for single index models in high-dimensions. *Annals of Mathematical Sciences and Applications*, 1(2):379–426, 2016.
- Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Minimax rates of estimation for high-dimensional linear regression over-balls. *Information Theory, IEEE Transactions on*, 57(10):6976–6994, 2011.
- James R Schott. Determining the dimensionality in sliced inverse regression. *Journal of the American Statistical Association*, 89(425):141–148, 1994.
- Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.
- Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.
- Vincent Q Vu and Jing Lei. Minimax rates of estimation for sparse pca in high dimensions. *arXiv preprint arXiv:1202.0786*, 2012.
- Shie-Shien Yang. General distribution theory of the concomitants of order statistics. *The Annals of Statistics*, pages 996–1002, 1977.
- Lixing Zhu, Baiqi Miao, and Heng Peng. On sliced inverse regression with high-dimensional covariates. *Journal of the American Statistical Association*, 101(474), 2006.
- Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

SUPPLEMENT TO : “ON OPTIMALITY OF SLICED INVERSE REGRESSION IN HIGH DIMENSIONS”

BY QIAN LIN[§] XINRAN LI[§] DONGMING HUANG[§] AND JUN S. LIU[§]
Harvard University[§]

The following lemmas will be used frequently during the proofs.

LEMMA 3. *Let \mathbf{K} be an $a \times b$ matrix with each entry being i.i.d. standard normal random variables. Then, we have $\mathbb{E}[\|\mathbf{K}\mathbf{K}^\tau\|_F^2] = ab(a + b + 1)$ and $\mathbb{E}[\|\mathbf{K}\|_F^2] = ab$.*

PROOF. It follows from elementary calculations. □

LEMMA 4. *Let A, B be $l \times m$ and $m \times n$ matrices, respectively, we have $\|AB\|_F \leq \|A\|_2 \|B\|_F$, where $\|A\|_2$ denotes the largest singular value of A .*

PROOF. It follows from elementary calculations. □

LEMMA 5. *Let A, B be $m \times l$ orthogonal matrices, i.e., $A^\tau A = I_l = B^\tau B$, and let M be an $l \times l$ positive definite matrix with eigenvalues d_j such as $0 < \lambda \leq d_l \leq d_{l-1} \leq \dots \leq d_1 \leq \kappa\lambda$. If $A^\tau B$ is a diagonal matrix with non-negative entries, then there exists a constant C which only depends on κ such that $\|AMA^\tau - BMB^\tau\|_F \leq C\lambda\|AA^\tau - BB^\tau\|_F$.*

PROOF. Let $\Delta = I_l - B^\tau A$, then $0 \leq \Delta_{ii} \leq 1$ for $1 \leq i \leq l$. If $C > 2\kappa^2 - 1$, we have

$$\begin{aligned} \|AMA^\tau - BMB^\tau\|_F^2 &= 2\text{tr}(M^2\Delta) - \text{tr}(M\Delta M\Delta) \leq 2\kappa^2\lambda^2\text{tr}(\Delta) - \lambda^2\text{tr}(\Delta^2) \\ &\leq C\lambda^2(2\text{tr}(\Delta) - \text{tr}(\Delta^2)) = C\lambda^2\|AA^\tau - BB^\tau\|_F^2. \end{aligned}$$

□

LEMMA 6. *For a positive definite matrix M with eigenvalue $\lambda_1 \geq \dots \geq \lambda_d > 0$ and orthogonal matrices A, B, E, F , i.e., $A^\tau A = B^\tau B = E^\tau E = F^\tau F = I_d$, we have*

$$\frac{\lambda_d}{2}\|AB^\tau - EF^\tau\|_F^2 \leq \langle AMB^\tau, AB^\tau - EF^\tau \rangle \leq \frac{\lambda_1}{2}\|AB^\tau - EF^\tau\|_F^2.$$

PROOF. It is a direct corollary of the Lemma 8 in [Gao et al. \[2014\]](#). \square

LEMMA 7 (Sparse approximation). *Let $\mathbf{V} \in \mathbb{O}_{s,q}(p, d)$ and $k \in [p]$, where $\mathbb{O}_{s,q}(p, d)$ is defined near (12). Let $\|\mathbf{V}_{(i)*}\|$ denote its i -th largest row norm. Then*

$$(36) \quad \sum_{i>k} \|\mathbf{V}_{(i)*}\|^2 \leq \frac{q}{2-q} k(s/k)^{2/q}.$$

In particular, if k is chosen to be $k_{s,q}$ defined near (12), we know that

$$(37) \quad \sum_{i>k} \|\mathbf{V}_{(i)*}\|^2 \leq \frac{q}{2-q} \epsilon_n^2.$$

PROOF. This is a direct corollary of the Lemma 7 in [Cai et al. \[2013\]](#). \square

LEMMA 8. *Let $\Sigma = \mathbf{V}\mathbf{D}\mathbf{V}^\tau$ be a $p \times p$ positive semidefinite matrix where \mathbf{V} is a $p \times d$ orthogonal matrix and \mathbf{D} is a $d \times d$ diagonal matrix with entries $\lambda \leq d_d \leq \dots \leq d_1 \leq \kappa\lambda$. For a subset S of indices with $|S| = k$, let J_S be a diagonal matrix such that $J_S(i, i) = 1$ if $i \in S$ and $J_S(i, i) = 0$ if $i \notin S$. Let $\Sigma_S = J_S \Sigma J_S$ and let $\Sigma_S = \mathbf{V}_1 \mathbf{D}_1 \mathbf{V}_1^\tau$ be the eigen-decomposition of Σ_S . We have*

$$\|\Sigma - \Sigma_S\|_F \leq 2\|\mathbf{D}\|_2 \|J_S \mathbf{V} - \mathbf{V}\|_F.$$

Furthermore, if $\|J_S \mathbf{V} - \mathbf{V}\|_F \leq \frac{1}{8\kappa}$, then $\|\Sigma - \Sigma_S\|_F \leq \lambda_d/4$. By Sin-Theta Lemma (e.g., Lemma 24), we have

$$\|\mathbf{V}\mathbf{V}^\tau - \mathbf{V}_1 \mathbf{V}_1^\tau\|_F \leq 8\kappa \|J_S \mathbf{V} - \mathbf{V}\|_F.$$

PROOF. This comes from a (trivial) elementary calculus. \square

Proof of the Theorem 2. We will adopt the notations introduced in [Lin et al. \[2018c\]](#). We need to introduce some notations and the ‘Oracle estimate’ $\hat{\mathbf{V}}_O$. Since we have randomly divided the samples into two equal sets of samples, we have the corresponding decomposition (35)

$$\hat{\Lambda}_H = \Lambda_u + \mathcal{Z}\mathcal{E}^\tau \mathbf{V}^{\perp, \tau} + \mathbf{V}^\perp \mathcal{E} \mathcal{Z}^\tau + \mathbf{V}^\perp \mathcal{E} \mathcal{E}^\tau \mathbf{V}^{\perp, \tau}.$$

for these two sets of samples. More precisely, for $i = 1, 2$, we can define $\Lambda_H^{(i)}$, $\Lambda_u^{(i)}$, $\mathcal{Z}^{(i)}$, $\mathcal{W}^{(i)}$ and $\mathcal{E}^{(i)}$ for the first and second set of samples respectively according to the decomposition (35). Let $\Lambda = \mathbf{V}\mathbf{D}\mathbf{V}^\tau$ be the spectral decomposition, where \mathbf{V} is $p \times d$ orthogonal matrix and $\mathbf{D} = \text{diag}\{\lambda_1, \dots, \lambda_d\}$

is a diagonal matrix. For $i = 1, 2$, let $\mathbf{\Lambda}_u^{(i)} = \mathbf{V}^{(i)} \mathbf{D}^{(i)} \mathbf{V}^{(i),\tau}$, where $\mathbf{V}^{(i)}$ is $p \times d$ orthogonal matrix and $\mathbf{D}^{(i)} = \text{diag}\{\lambda_1^{(i)}, \dots, \lambda_d^{(i)}\}$ is a diagonal matrix. For any subset S of $[p]$, let J_S be the diagonal matrix defined in Lemma 21. Let $J_S \mathbf{\Lambda}_u J_S = \mathbf{V}_S \mathbf{D}_S \mathbf{V}_S^\tau$ be the spectral decomposition, where \mathbf{V}_S is $p \times d$ orthogonal matrix and $\mathbf{D}_S = \text{diag}\{\lambda_{1,S}, \dots, \lambda_{d,S}\}$ is a diagonal matrix. Let $J_S \mathbf{\Lambda}_u^{(i)} J_S = \mathbf{V}_S^{(i)} \mathbf{D}_S^{(i)} \mathbf{V}_S^{(i),\tau}$ be the spectral decomposition, where $\mathbf{V}_S^{(i)}$ is $p \times d$ orthogonal matrix and $\mathbf{D}_S^{(i)} = \text{diag}\{\lambda_{1,S}^{(i)}, \dots, \lambda_{d,S}^{(i)}\}$ is a diagonal matrix. In the below, we will call \mathbf{V}_S (resp. $\mathbf{V}_S^{(i)}$, $i = 1, 2$) the sparse approximation of \mathbf{V} (resp. $\mathbf{V}^{(i)}$, $i = 1, 2$). From now on, we will choose S to be $[k_{q,s}] \subset [p]$ where $k_{q,s}$ is defined near (12).

Below, we use C to denote an absolute constant, though its exact value may vary from case to case. We also assume that ϵ_n^2 is sufficiently small. For $i = 1, 2$, let $\mathbf{E}_3^{(i)}$ be the event $\left\{ \|\mathbf{\Lambda}_u^{(i)} - \mathbf{\Lambda}\| \leq \frac{1}{2\nu} \kappa \lambda \right\}$. Let $\mathbf{E} := \mathbf{E}_3^{(1)} \cap \mathbf{E}_3^{(2)}$. Lemma 2 shows $\mathbb{P}(\mathbf{E}^c) \leq \frac{CH^2}{n\lambda}$. Conditioning on \mathbf{E} ,

$$(38) \quad \left(1 - \frac{\kappa}{2\nu}\right)\lambda \leq \lambda_d^{(i)} \leq \dots \leq \lambda_1^{(i)} \leq \left(1 + \frac{1}{2\nu}\right)\kappa\lambda \quad \text{for } i = 1, 2.$$

We first prove the following sparse approximation lemma.

LEMMA 9. *Conditioning on \mathbf{E} , we have*

$$(39) \quad \|\mathbf{V}_S^{(i)} \mathbf{V}_S^{(i),\tau} - \mathbf{V}_S \mathbf{V}_S^\tau\|_F^2 \leq C \frac{q}{2-q} \epsilon_n^2, \quad \|J_S \mathbf{\Lambda}_u^{(i)} J_S - \mathbf{\Lambda}_u^{(i)}\|_F \leq C \lambda \epsilon_n$$

and the entries of $\mathbf{D}_S^{(i)} \in (\frac{1}{2}\lambda, 2\kappa\lambda)$ for $i = 1, 2$.

PROOF. Since $\mathbf{V}^{(i)}$ and \mathbf{V} share the same column space, we have $\mathbf{V}^{(i)} = \mathbf{V} \tilde{U}$ for some (stochastic) orthogonal matrix \tilde{U} and $\mathbf{V} \mathbf{V}^\tau = \mathbf{V}^{(i)} \mathbf{V}^{(i),\tau}$. From this we know that

$$\|\mathbf{V}_S^{(i)} \mathbf{V}_S^{(i),\tau} - \mathbf{V}_S \mathbf{V}_S^\tau\|_F \leq \|\mathbf{V} \mathbf{V}^\tau - \mathbf{V}_S \mathbf{V}_S^\tau\|_F + \|\mathbf{V}_S^{(i)} \mathbf{V}_S^{(i),\tau} - \mathbf{V}^{(i)} \mathbf{V}^{(i),\tau}\|_F.$$

Conditioning on \mathbf{E} , Lemma 20, Lemma 21 and (38) imply

$$\|J_S \mathbf{\Lambda}_u^{(i)} J_S - \mathbf{\Lambda}_u^{(i)}\|_F \leq C \kappa \lambda \|\mathbf{V}_S \mathbf{V}^{(i)} - \mathbf{V}^{(i)}\|_F \leq C \kappa \lambda \epsilon_n.$$

Since we have assumed that ϵ_n^2 is sufficiently small, we can assure that the entries of $\mathbf{D}_S^{(i)}$ are in the range $(\frac{1}{2}\lambda, 2\kappa\lambda)$. After applying the Sin-Theta theorem (e.g. Lemma 24), we have

$$\|\mathbf{V}_S^{(i)} \mathbf{V}_S^{(i),\tau} - \mathbf{V}^{(i)} \mathbf{V}^{(i),\tau}\|_F \leq C \kappa \|\mathbf{V}_S \mathbf{V}^{(i)} - \mathbf{V}^{(i)}\|_F \leq C \kappa \sqrt{\frac{q}{2-q}} \epsilon_n.$$

We can apply similar argument to bound $\|\mathbf{V}\mathbf{V}^\tau - \mathbf{V}_S\mathbf{V}_S^\tau\|_F$, which gives us

$$\|\mathbf{V}_S^{(i)}\mathbf{V}_S^{(i),\tau} - \mathbf{V}_S\mathbf{V}_S^\tau\|_F \leq C\kappa\|J_S\mathbf{V} - \mathbf{V}\|_F \leq C\sqrt{\frac{q}{2-q}}\epsilon_n.$$

□

We introduce an ‘Oracle estimator’ $\widehat{\mathbf{V}}_O$ (as if we know the sparse approximation set S) such that

$$(40) \quad \begin{aligned} \widehat{\mathbf{V}}_O &\triangleq \arg \max_{\mathbf{V}} \langle \boldsymbol{\Lambda}_H^{(1)}, \mathbf{V}\mathbf{V}^\tau \rangle = \arg \max_{\mathbf{V}} \text{Tr}(\mathbf{V}^\tau \boldsymbol{\Lambda}_H^{(1)} \mathbf{V}) \\ \text{s.t. } &\mathbf{V}^\tau \mathbf{V} = I_d \text{ and } \text{supp}(\mathbf{V}) = S. \end{aligned}$$

Let $\widehat{\mathbf{V}}_O^\tau \mathbf{V}_S^{(2)} = U_1 \Delta U_2^\tau$ be the singular value decomposition of $\widehat{\mathbf{V}}_O^\tau \mathbf{V}_S^{(2)}$ such that the entries of Δ are non-negative and let $M \triangleq U_2^\tau \mathbf{D}_S^{(2)} U_2$.

Now, we can start our proof of Theorem 2. It is easy to verify that

$$\|\widehat{\mathbf{V}}_E \widehat{\mathbf{V}}_E^\tau - \mathbf{V}\mathbf{V}^\tau\|_F^2 \leq C \left(\|\widehat{\mathbf{V}}_E \widehat{\mathbf{V}}_E^\tau - \widehat{\mathbf{V}}_O \widehat{\mathbf{V}}_O^\tau\|_F^2 + \|\widehat{\mathbf{V}}_O \widehat{\mathbf{V}}_O^\tau - \mathbf{V}_S \mathbf{V}_S^\tau\|_F^2 + \|\mathbf{V}_S \mathbf{V}_S^\tau - \mathbf{V}\mathbf{V}^\tau\|_F^2 \right).$$

For the first term $\|\widehat{\mathbf{V}}_E \widehat{\mathbf{V}}_E^\tau - \widehat{\mathbf{V}}_O \widehat{\mathbf{V}}_O^\tau\|_F^2$, conditioning on \mathbf{E} , we know

$$(41) \quad \|\widehat{\mathbf{V}}_E \widehat{\mathbf{V}}_E^\tau - \widehat{\mathbf{V}}_O \widehat{\mathbf{V}}_O^\tau\|_F^2 \leq \frac{2}{\lambda_d(\mathbf{D}_S^{(2)})} \langle \widehat{\mathbf{V}}_O U_1 M U_1^\tau \widehat{\mathbf{V}}_O^\tau, \widehat{\mathbf{V}}_O \widehat{\mathbf{V}}_O^\tau - \widehat{\mathbf{V}}_E \widehat{\mathbf{V}}_E^\tau \rangle$$

$$(42) \quad \begin{aligned} &\leq \frac{C}{\lambda} \langle \widehat{\mathbf{V}}_O U_1 M U_1^\tau \widehat{\mathbf{V}}_O^\tau - \boldsymbol{\Lambda}_H^{(2)}, \widehat{\mathbf{V}}_O \widehat{\mathbf{V}}_O^\tau - \widehat{\mathbf{V}}_E \widehat{\mathbf{V}}_E^\tau \rangle \\ &\triangleq I + II + III. \end{aligned}$$

where

$$\begin{aligned} I &= \frac{C}{\lambda} \langle \widehat{\mathbf{V}}_O U_1 M U_1^\tau \widehat{\mathbf{V}}_O^\tau - \mathbf{V}_S^{(2)} \mathbf{D}_S^{(2)} \mathbf{V}_S^{(2),\tau}, \widehat{\mathbf{V}}_O \widehat{\mathbf{V}}_O^\tau - \widehat{\mathbf{V}}_E \widehat{\mathbf{V}}_E^\tau \rangle \\ II &= \frac{C}{\lambda} \langle \mathbf{V}_S^{(2)} \mathbf{D}_S^{(2)} \mathbf{V}_S^{(2),\tau} - \boldsymbol{\Lambda}_u^{(2)}, \widehat{\mathbf{V}}_O \widehat{\mathbf{V}}_O^\tau - \widehat{\mathbf{V}}_E \widehat{\mathbf{V}}_E^\tau \rangle \\ III &= \frac{C}{\lambda} \langle \boldsymbol{\Lambda}_u^{(2)} - \boldsymbol{\Lambda}_H^{(2)}, \widehat{\mathbf{V}}_O \widehat{\mathbf{V}}_O^\tau - \widehat{\mathbf{V}}_E \widehat{\mathbf{V}}_E^\tau \rangle. \end{aligned}$$

Inequality (41) follows from applying the Lemma 19 with the positive definite matrix $U_1 M U_1^\tau$. The inequality (42) follows from the definition of $\widehat{\mathbf{V}}_E$ (See (17)) and the fact that the entries of $\mathbf{D}_S^{(2)}$ are in $(\lambda/2, 2\kappa\lambda)$. To simplify the notation, we let

$$R = \|\widehat{\mathbf{V}}_E \widehat{\mathbf{V}}_E^\tau - \mathbf{V}\mathbf{V}^\tau\|_F, \theta^{(i)} = \|\mathbf{V}_S^{(i)} \mathbf{V}_S^{(i),\tau} - \mathbf{V}\mathbf{V}^\tau\|_F, \delta = \|\widehat{\mathbf{V}}_O \widehat{\mathbf{V}}_O^\tau - \mathbf{V}_S^{(1)} \mathbf{V}_S^{(1),\tau}\|_F.$$

For I.: Recall that the entries of $\mathbf{D}_S^{(2)} \in (\frac{1}{2}\lambda, 2\kappa\lambda)$ and that $\widehat{\mathbf{V}}_O U_1$ and $\mathbf{V}_S^{(2)} U_2$ satisfies the condition that $U_1^\tau \widehat{\mathbf{V}}_O^\tau \mathbf{V}_S^{(2)} U_2$ is a diagonal matrix with non-negative entries. Since $M \triangleq U_2^\tau \mathbf{D}_S^{(2)} U_2$, by Lemma 18 and the fact that $\mathbf{V}_S^{(2)} U_2$ and $\mathbf{V}_S^{(2)}$ share the same column space, there exists a constant C such that

$$\|\widehat{\mathbf{V}}_O U_1 M U_1^\tau \widehat{\mathbf{V}}_O^\tau - \mathbf{V}_S^{(2)} \mathbf{D}_S^{(2)} \mathbf{V}_S^{(2),\tau}\|_F \leq C\lambda \|\widehat{\mathbf{V}}_O \widehat{\mathbf{V}}_O^\tau - \mathbf{V}_S^{(2)} \mathbf{V}_S^{(2),\tau}\|_F.$$

Thus, conditioning on \mathbf{E} , we have

$$(43) \quad \begin{aligned} |I| &\leq C \|\widehat{\mathbf{V}}_O \widehat{\mathbf{V}}_O^\tau - \mathbf{V}_S^{(2)} \mathbf{V}_S^{(2),\tau}\|_F \|\widehat{\mathbf{V}}_O \widehat{\mathbf{V}}_O^\tau - \widehat{\mathbf{V}}_E \widehat{\mathbf{V}}_E^\tau\|_F \\ &\leq C(\delta + \theta^{(1)} + \theta^{(2)}) \|\widehat{\mathbf{V}}_O \widehat{\mathbf{V}}_O^\tau - \widehat{\mathbf{V}}_E \widehat{\mathbf{V}}_E^\tau\|_F. \end{aligned}$$

For II.: It is nonzero only if $q \neq 0$. From Lemma 20 and Lemma 21, we know that

$$(44) \quad |II| \leq C \sqrt{\frac{q}{2-q}} \epsilon_n \|\widehat{\mathbf{V}}_O \widehat{\mathbf{V}}_O^\tau - \widehat{\mathbf{V}}_E \widehat{\mathbf{V}}_E^\tau\|_F.$$

For III.: From the equation (35), we have

$$(45) \quad |III| \leq \frac{1}{\lambda} \|\widehat{\mathbf{V}}_O \widehat{\mathbf{V}}_O^\tau - \widehat{\mathbf{V}}_E \widehat{\mathbf{V}}_E^\tau\|_F (2T_2 + T_1)$$

where $T_1 = \max_{B \in \mathcal{B}(k)} \left| \left\langle \mathcal{W}^{(2)} \mathcal{W}^{(2),\tau}, \mathbf{K}_B \right\rangle \right|$, $T_2 = \max_{B \in \mathcal{B}(k)} \left| \left\langle \mathcal{Z}^{(2)} \mathcal{W}^{(2),\tau}, \mathbf{K}_B \right\rangle \right|$ and $\mathbf{K}_B = \|\widehat{\mathbf{V}}_O \widehat{\mathbf{V}}_O^\tau - \widehat{\mathbf{V}}_B \widehat{\mathbf{V}}_B^\tau\|_F^{-1} (\widehat{\mathbf{V}}_O \widehat{\mathbf{V}}_O^\tau - \widehat{\mathbf{V}}_B \widehat{\mathbf{V}}_B^\tau)$. (For any $B \in \mathcal{B}_k$, $\widehat{\mathbf{V}}_B$ is introduced in (16)).

To summarize, conditioning on \mathbf{E} , we have

$$(46) \quad \|\widehat{\mathbf{V}}_E \widehat{\mathbf{V}}_E^\tau - \widehat{\mathbf{V}}_O \widehat{\mathbf{V}}_O^\tau\|_F \leq C \left(\delta + \theta^{(1)} + \theta^{(2)} + \epsilon_n + \frac{1}{\lambda} (2T_2 + T_1) \right).$$

Thus, we have

$$\begin{aligned} R^2 \mathbf{1}_E &\leq C \left(\delta^2 + \left(\theta^{(1)} \right)^2 + \|\widehat{\mathbf{V}}_O \widehat{\mathbf{V}}_O^\tau - \widehat{\mathbf{V}}_E \widehat{\mathbf{V}}_E^\tau\|_F^2 \right) \mathbf{1}_E \\ &\leq C \left(\delta^2 + \left(\theta^{(1)} \right)^2 + C \left(\delta + \theta^{(1)} + \theta^{(2)} + \epsilon_n + \frac{1}{\lambda} (2T_1 + T_2) \right)^2 \right) \mathbf{1}_E \\ &\leq C \left(\delta^2 + \left(\theta^{(1)} \right)^2 + \left(\theta^{(2)} \right)^2 + \epsilon_n^2 \mathbf{1}_E + \left(\frac{1}{\lambda} (2T_1 + T_2) \right)^2 \right) \mathbf{1}_E \end{aligned}$$

If we can prove

$$(47) \quad \mathbb{E} \left(\theta^{(i)} \right)^2 \mathbf{1}_{\mathbf{E}'} \leq C\epsilon_n^2, \quad \mathbb{E} \delta^2 \mathbf{1}_{\mathbf{E}'} \leq C\epsilon_n^2 \quad \text{and} \quad \mathbb{E} (2T_1 + T_2)^2 \mathbf{1}_{\mathbf{E}'} \leq \lambda^2 \epsilon_n^2,$$

for some $\mathbf{E}' \subset \mathbf{E}$ such that $\mathbb{P}((\mathbf{E}')^c) \leq C \frac{H^2}{n\lambda}$, then we have $\mathbb{E} R^2 \mathbf{1}_{\mathbf{E}'} \leq C\epsilon_n^2$. Thus, we have

$$\mathbb{E} R^2 \leq C\epsilon_n^2.$$

□

All we need to prove are the following two Lemmas.

LEMMA 10. *There exists $\mathbf{E}' \subset \mathbf{E}$ such that $\mathbb{P}((\mathbf{E}')^c) \leq C \frac{H^2}{n\lambda}$,*

$$(48) \quad \mathbb{E} \left(\theta^{(i)} \right)^2 \mathbf{1}_{\mathbf{E}'} \leq C\epsilon_n^2 \quad \text{and} \quad \mathbb{E} \delta^2 \mathbf{1}_{\mathbf{E}'} \leq C\epsilon_n^2.$$

PROOF. Since $\mathbf{V}^{(i)}$ and \mathbf{V} share the same column space, conditioning on \mathbf{E} , by Lemma 21 and Lemma 20, we have

$$(49) \quad \theta^{(i)} = \|\mathbf{V}_S^{(i)} \mathbf{V}_S^{(i),\tau} - \mathbf{V} \mathbf{V}^\tau\|_F \leq 4\kappa \|\mathbf{J}_S \mathbf{V}^{(i)} - \mathbf{V}^{(i)}\|_F \leq C \sqrt{\frac{q}{2-q}} \epsilon_n,$$

i.e. $\mathbb{E} (\theta^{(i)})^2 \mathbf{1}_{\mathbf{E}} \leq C\epsilon_n^2$.

Let $\mathbf{Q}_S = \mathbf{J}_S (\mathbf{\Lambda}_H^{(1)} - \mathbf{\Lambda}_u^{(1)}) \mathbf{J}_S$. Let \mathbf{F} consist of the events such that $\|\mathbf{J}_S \mathbf{W}^{(1)} \mathbf{W}^{(1),\tau} \mathbf{J}_S\| \leq C \frac{k}{n}$. Lemma 23 implies that $\mathbb{P}(\mathbf{F}^c) \leq C \frac{H^2}{n\lambda}$. Since we have assumed that ϵ_n^2 is sufficiently small, conditioning on $\mathbf{E} \cap \mathbf{F}$, the decomposition (35) give us

$$(50) \quad \|\mathbf{Q}_S\|_2 \leq C \sqrt{\lambda} \sqrt{\frac{k}{n}} \leq C \frac{\lambda \epsilon_n}{\sqrt{d}}.$$

From (39), we also have $\|\mathbf{J}_S \mathbf{\Lambda}_u^{(1)} \mathbf{J}_S - \mathbf{\Lambda}_u^{(1)}\|_F \leq C \lambda \epsilon_n \leq \frac{\lambda}{8}$. Thus, $\|\mathbf{J}_S \mathbf{\Lambda}_H^{(1)} \mathbf{J}_S - \mathbf{\Lambda}_u^{(1)}\|_F < \frac{\lambda}{4}$, which implies the $(d+1)$ -th largest eigenvalues of $\mathbf{J}_S \mathbf{\Lambda}_H^{(1)} \mathbf{J}_S$ is less than $\frac{\lambda}{4}$. Note that the eigenvalues of $\mathbf{J}_S \mathbf{\Lambda}_u^{(1)} \mathbf{J}_S \in (\frac{1}{2}\lambda, 2\kappa\lambda)$. After applying the Sin-Theta Theorem(Lemma 24) to the pair of symmetric matrices $(\mathbf{J}_S \mathbf{\Lambda}_u^{(1)} \mathbf{J}_S, \mathbf{J}_S \mathbf{\Lambda}_H^{(1)} \mathbf{J}_S)$, we have

$$\delta \leq \frac{8}{\lambda} \|\widehat{\mathbf{V}}_O^{\perp, \tau} \mathbf{Q}_S \mathbf{V}_S^{(1)}\|_F \leq \frac{8}{\lambda} \sqrt{d} \|\mathbf{Q}_S\|_2 \leq C \epsilon_n$$

where the last inequality follows from (50). Thus, we may take $\mathbf{E}' = \mathbf{E} \cap \mathbf{F}$.

LEMMA 11. *There exists positive constant C such that*

$$\mathbb{E}(2T_1 + T_2)^2 \mathbf{1}_E \leq C\lambda^2 \epsilon_n^2$$

PROOF. Since $(2T_1 + T_2)^2 \leq C(T_1^2 + T_2^2)$, we only need to bound $\mathbb{E}T_1^2$ and $\mathbb{E}T_2^2$ separately.

For T_1 . Recall that $\mathcal{W}^{(2)} = \mathbf{V}^\perp \mathcal{E}^{(2)}$ (See notation near (35).) and for each fixed $B \in \mathcal{B}_k$, $\mathbf{K}_B \perp \mathcal{W}^{(2)}$, hence

$$(51) \quad \langle \mathcal{W}^{(2)} \mathcal{W}^{(2),\tau}, \mathbf{K}_B \rangle = \langle \mathcal{E}^{(2)} \mathcal{E}^{(2),\tau}, \mathbf{V}^{\perp,\tau} \mathbf{K}_B \mathbf{V}^\perp \rangle$$

and $\mathbf{V}^{\perp,\tau} \mathbf{K}_B \mathbf{V}^\perp \perp \mathcal{W}^{(2)}$. Note that $\|\mathbf{V}^{\perp,\tau} \mathbf{K}_B \mathbf{V}^\perp\|_F \leq 1$, $\mathcal{E}^{(2)}$ is a $(p-d) \times H$ matrix and $\sqrt{n}\mathcal{E}_{i,j}^{(2)} \sim N(0, 1)$. After applying Lemma 25, we have

$$(52) \quad \mathbb{P} \left(\sqrt{n} |\langle \mathcal{E}^{(2)} \mathcal{E}^{(2),\tau}, \mathbf{V}^{\perp,\tau} \mathbf{K}_B \mathbf{V}^\perp \rangle| \geq 2 \frac{\sqrt{H}}{\sqrt{n}} t + \frac{2}{\sqrt{n}} t^2 \right) \leq 2 \exp(-t^2).$$

After applying Lemma 26 with $N = |\mathcal{B}(k)| \leq \left(\frac{ep}{k}\right)^k$, $a = \frac{2\sqrt{H}}{\sqrt{n}}$, $b = \frac{2}{\sqrt{n}}$ and $c = 2$, we have

$$\begin{aligned} \mathbb{E}T_1^2 &\leq \frac{1}{n} \left(\frac{8H}{n} \log(2eN) + \frac{8}{n} (\log^2(2N) + 4 \log(2eN)) \right) \\ &= \frac{8(H+4) \log(2eN)}{n^2} + \frac{8}{n^2} \log^2(2N) \\ &\leq C\lambda^2 \frac{\log(N)}{n\lambda} \leq C\lambda^2 \epsilon_n^2 \end{aligned}$$

For T_2 . Fix $B \in \mathcal{B}(k_{q,s})$. Since $\mathcal{Z}^{(2)} \perp \mathcal{W}^{(2)}$, $\mathbf{K}_B \perp \mathcal{W}^{(2)}$ and $\mathbf{K}_B \perp \mathcal{Z}^{(2)}$, conditioned on the $\mathcal{Z}^{(2)}$ and \mathbf{K}_B , we know that

$$\sqrt{n} \langle \mathcal{Z}^{(2)} \mathcal{W}^{(2),\tau}, \mathbf{K}_B \rangle = \langle \mathbf{V}^{\perp,\tau} \mathbf{K}_B \mathcal{Z}^{(2)}, \sqrt{n} \mathcal{E}^{(2)} \rangle$$

is distributed according to $N(0, \|\mathbf{V}^{\perp,\tau} \mathbf{K}_B \mathcal{Z}^{(2)}\|_F^2)$. Therefore

$$\sqrt{n} \langle \mathcal{Z}^{(2)} \mathcal{W}^{(2),\tau}, \mathbf{K}_B \rangle \stackrel{d}{=} \|\mathbf{V}^{\perp,\tau} \mathbf{K}_B \mathcal{Z}^{(2)}\|_F W$$

for some $W \sim N(0, 1)$ independent of $\mathcal{Z}^{(2)}$ and \mathbf{K}_B . For simplicity of notation, we denote $\sqrt{n} \langle \mathcal{Z}^{(2)} \mathcal{W}^{(2),\tau}, \mathbf{K}_B \rangle$ by F_B . As a direct corollary, conditioning on \mathbf{E} , we know

$$(53) \quad \mathbb{P}(|F_B| > t) \leq \mathbb{P}(2\kappa\lambda|W| > t) \leq 2 \exp\left(-\frac{t^2}{4\kappa^2\lambda^2}\right).$$

i.e., conditioning on \mathbf{E} , $|F_B|$ is sub-Gaussian and upper exponentially bounded by $4\kappa^2\lambda^2$. From this, we know $\mathbb{E}(T_2^2 \mathbf{1}_E) \leq C\lambda^2 \epsilon_n^2$.

□

Proof of Theorem 10. For a vector $\gamma \in \mathbb{R}^p$, $S \subset [p]$, let $\gamma_S \in \mathbb{R}^p$ such that $\gamma_S(i) = \gamma(i)$ if $i \in S$ and $\gamma_S(i) = 0$ if $i \notin S$. For any non-zero vector γ , let $\tilde{\gamma} = \gamma/\|\gamma\|_2$. For any non-zero vector γ and $t > 0$, let T_t be the indices such that $|\gamma(i)| > t$. We have following elementary Lemmas.

LEMMA 12. *Let $\gamma \in \mathbb{R}^p$ be a unit vector with at most s non-zero entries, then*

$$(54) \quad \|\gamma - \tilde{\gamma}_{T_t}\|_2^2 \leq Cst^2.$$

PROOF. Let $E^2 = \sum_{|\gamma_j| \geq t} \gamma_j^2$, then $\|\gamma - \tau_N(\gamma, t)\|_2^2 = \sum_{\gamma_j \geq t} \gamma_j^2 (1 - \frac{1}{E})^2 + \sum_{\gamma_j < t} \gamma_j^2 = 2(1 - E) \leq 2st^2$. \square

LEMMA 13. *Let $\mathbf{\Lambda} = \lambda \beta \beta^\tau$. For $S \subset [p]$, we have $\lambda(\mathbf{\Lambda}(S, S)) = \lambda \|\beta_S\|^2 \tilde{\beta}_S \tilde{\beta}_S^\tau$.*

PROOF. It follows from (trivial) elementary calculus. \square

Let $T = \{i \mid \mathbf{\Lambda}_H(i, i) > a \frac{\log(p)}{n}\}$, then $\lambda_T = \lambda(\mathbf{\Lambda}(T, T)) \geq \lambda \|\beta_T\|^2 \geq \lambda(1 - \frac{as \log(p)}{n}) \geq a' \lambda$ if $\frac{s \log(p)}{n \lambda}$ is sufficiently small. Since $\hat{\mathbf{\Lambda}}_H(i, i) \sim \frac{1}{n} \chi_H^2$ for $i \notin S$, we have $\mathbb{P}(\hat{\mathbf{\Lambda}}_H(i, i) > a \frac{\log(p)}{n}) \leq \exp(-a \log(p))$. Thus, we have $\mathbb{P}(T \subset S) \geq \mathbb{P}(\max_{i \notin S} \hat{\mathbf{\Lambda}}_H(i, i) \leq a \frac{\log(p)}{n}) \geq 1 - \exp(-(a-1) \log(p))$.

Thus, if $T \subset S$, we have

$$\|\hat{\beta}_T - \beta\|_F^2 \leq \|\hat{\beta}_T - \tilde{\beta}_T\|_F^2 + \|\tilde{\beta}_T - \beta\|_F^2 \leq C \frac{|T|}{n \lambda_T} + Cst^2 \leq C \frac{s \log(p)}{n \lambda}$$

where we have used the Oracle risk Theorem 1 and the Lemma 12. Thus, we know that DT-SIR is rate optimal if $s = O(p^{1-\delta})$.

The lower bound. In this subsection, we provide the proof of the lower bound for Theorem 3, Theorem 4, Theorem 5.

Proof of Theorem 3. Let us consider the Grassmannian $\mathbb{G}(p, d)$ consisting of all the d dimensional subspaces in \mathbb{R}^p and the homogeneous space $\mathbb{O}(p, d)$ consisting of all $p \times d$ orthogonal matrices. There is a tautological map from $\mathbb{O}(p, d)$ to $\mathbb{G}(p, d)$, i.e., $\mathbf{A} \mapsto \mathbf{A} \mathbf{A}^\tau$. For any $\varepsilon \in (0, \sqrt{2d} \wedge (p-d)]$, for any $u \in \mathbb{G}(p, d)$, and any $\alpha \in (0, 1)$, the Lemma 1 in Cai et al. [2013] have constructed a subset $\Theta \subset N(u, \varepsilon)$, an ε neighbourhood of u in $\mathbb{G}(p, d)$, such that

$$|u_i - u_j| \leq 2\varepsilon, \quad |u_i - u_j| \geq \alpha\varepsilon \text{ and } |\Theta| \geq \left(\frac{c_0}{\alpha c_1}\right)^{d(p-d)}$$

where u_i and u_j are two different points $\in \Theta$ and c_0 and c_1 are two absolute constants. Suppose $u = aa^\tau$, Lemma 22 states that for each $u_i \in \Theta \subset \mathbb{G}(p, d)$, there is an $a_i \in \mathbb{O}(p, d)$ such that $a_i a_i^\tau = u_i$ and

$$(55) \quad \|a_i - a\|_F \leq C_2 \|u_i - u\|_F.$$

This implies $\|a_i - a_j\|_F \leq 2\varepsilon$. Let us denote $\tilde{\Theta} = \{a_i\}$ and consider the following models

$$(56) \quad y = f(\mathbf{V}^\tau \mathbf{x}) + \epsilon, \mathbf{V} \in \tilde{\Theta}, \mathbf{x} \sim N(0, \mathbf{I}_p), \epsilon \sim N(0, 1),$$

where f comes from Lemma 15 for the proof of Theorem 3, or for f comes from the Conjecture 1 for the proof of Theorem 7. Simple calculation shows the following:

LEMMA 14. *Let $y = g(\mathbf{B}^\tau \mathbf{x}) + \epsilon, \epsilon \sim N(0, 1)$ where $\mathbf{B} \in \mathbb{O}(p, d)$ and $\mathbf{x} \sim N(0, \mathbf{I}_p)$ and let $p_{B,g}(y, \mathbf{x})$ be the joint density function of (y, \mathbf{x}) , then we have*

$$(57) \quad KL(p_{B,g}, p_{B',g}) \leq \max |\nabla g|^2 \|\mathbf{B} - \mathbf{B}'\|_F^2.$$

PROOF. The density function $p_{\mathbf{B}}(y, \mathbf{x})$ of (y, \mathbf{x}) is

$$(58) \quad p_{\mathbf{B}}(y, \mathbf{x}) = p_{\mathbf{B}}(y|\mathbf{x})p(\mathbf{x}) = \frac{1}{\sqrt{2\pi}} \exp^{-\frac{1}{2}(y-g(\mathbf{B}^\tau \mathbf{x}))^2} p(\mathbf{x})$$

where $p(\mathbf{x})$ is the density function of standard p -dimensional normal distribution. Let $z = y - g(\mathbf{B}^\tau \mathbf{x})$, then we have

$$\begin{aligned} & KL(p_{\mathbf{B}}, p_{\mathbf{B}'}) \\ &= \int \frac{1}{\sqrt{2\pi}} \exp^{-\frac{1}{2}(y-g(\mathbf{B}^\tau \mathbf{x}))^2} p(\mathbf{x}) \left(\frac{1}{2} \left(y - g(\mathbf{B}'^\tau \mathbf{x}) \right)^2 - \frac{1}{2} (y - g(\mathbf{B}^\tau \mathbf{x}))^2 \right) d\mathbf{x} dy \\ &= \int \frac{1}{\sqrt{2\pi}} \exp^{-\frac{1}{2}z^2} p(\mathbf{x}) \left(\frac{1}{2} \left(z + g(\mathbf{B}^\tau \mathbf{x}) - g(\mathbf{B}'^\tau \mathbf{x}) \right)^2 - \frac{1}{2} z^2 \right) d\mathbf{x} dz \\ &= \int \frac{1}{\sqrt{2\pi}} \exp^{-\frac{1}{2}z^2} p(\mathbf{x}) \frac{1}{2} \left(g(\mathbf{B}^\tau \mathbf{x}) - g(\mathbf{B}'^\tau \mathbf{x}) \right)^2 d\mathbf{x} dz \\ &\leq \max |\nabla g|^2 \int p(\mathbf{x}) \mathbf{x}^\tau (\mathbf{B} - \mathbf{B}') (\mathbf{B}^\tau - \mathbf{B}'^\tau) \mathbf{x} d\mathbf{x} \\ &= \max |\nabla g|^2 \text{Tr} \left((\mathbf{B} - \mathbf{B}') (\mathbf{B}^\tau - \mathbf{B}'^\tau) \right) \\ &= \max |\nabla g|^2 \|\mathbf{B} - \mathbf{B}'\|_F^2. \end{aligned}$$

□

The Fano Lemma gives us

$$\begin{aligned}
 & \sup_{\mathbf{V} \in \tilde{\Theta}} \mathbb{E} \|P_{\hat{\mathbf{V}}} - P_{\mathbf{V}}\|_F^2 \\
 (59) \quad & \geq \min \|P_{\mathbf{V}_i} - P_{\mathbf{V}_j}\|_F^2 \left(1 - \frac{\max KL(p_{\mathbf{V}_{i,f}}^n, p_{\mathbf{V}_{j,f}}^n) + \log(2)}{\log(|\Theta|)} \right) \\
 & \geq \alpha^2 \varepsilon^2 \left(1 - \frac{4n\lambda\varepsilon^2 + \log 2}{\log(|\Theta|)} \right).
 \end{aligned}$$

Since $\log |\Theta| > Cd(p-d)$, we know that, if $\frac{\log(|\Theta|)}{n\lambda}$ is sufficiently small, we have

$$(60) \quad \sup_{u \in \Theta} \mathbb{E} \|P_{\hat{\mathbf{V}}} - P_{\mathbf{V}}\|_F^2 \succ \frac{d(p-d)}{n\lambda}$$

by choosing $\varepsilon^2 = \frac{\log(|\Theta|)}{2n\lambda}$. This gives us the desired lower bound for ‘Oracle risk’.

LEMMA 15. *For any $d \in N_+$, one can find a smooth d -variate function f such that for $z \sim N_d(0, I)$, and $Y = f(z)$, then $\text{Var}(\mathbb{E}(z|Y))$ has full rank. Furthermore, for a sufficiently large constant A , the model $Y' = Af(\mathbf{B}^T \mathbf{x}) + \epsilon \in \mathcal{M}(p, d, \lambda, \kappa)$ for some sufficiently small λ and sufficiently large κ where $\mathbf{x} \sim N(0, \mathbf{I}_p)$, $\epsilon \sim N(0, 1)$ and \mathbf{B} is a $p \times d$ orthogonal matrix.*

PROOF. Since the second statement is a direct corollary of the first one, we only prove the first statement. Let $\phi(x)$ be a smooth function which maps $(-\infty, 0]$ to 0 and $[1, \infty)$ to 1 and has positive first derivative over $(0, 1)$. Define $f(z) := \sum_{i \leq d} 2^{i-1} \phi(z_i/\epsilon)$, where ϵ is sufficiently small such that the probability $\mathbb{P}(\exists i, 0 < z_i < \epsilon) \leq d\epsilon/\sqrt{2\pi} < 2^{-d}$.

Any integer $k \in [0, 2^d)$ can be represented as $\sum_{i=1}^d 2^{i-1} a_i$, $a_i \in \{0, 1\}$, and denote the set $\{z : z_i \leq 0, \text{ if } a_i = 0, \text{ or } z_i \geq \epsilon, \text{ if } a_i = 1\}$ by A_k . Since $f(z)$ ’s first derivative is non-zero when at least one $z_i \in (0, \epsilon)$ and $z \sim N(0, \mathbf{I}_d)$, we have

$$(61) \quad \mathbb{P}(f(z) = k) = \mathbb{P}(z \in A_k) \geq 2^{-d} - d\epsilon/\sqrt{2\pi} > 0.$$

Furthermore,

$$(62) \quad \mathbf{m}(k) = \mathbb{E}(z|f(z) = k) \stackrel{a.s.}{=} \mathbb{E}(z|z \in A_k).$$

With some elementary calculus, it follows that the i -th element of $\mathbf{m}(k)$ equals to $-\sqrt{2/\pi}$ if $a_i = 0$ or $c_\epsilon := \mathbb{E}(z_i|z_i > \epsilon)$ if $a_i = 1$. Since $\epsilon d/\sqrt{2\pi} <$

$2^{-d} \leq 1/2$, we have $\sqrt{2/\pi} < c_\epsilon < \sqrt{2/\pi} + 2\epsilon d/\pi$. One can see that the vectors $\{\mathbf{m}(k) : 0 \leq k < 2^d\}$ has rank d . If $\text{Var}(\mathbf{m}(Y))$ is singular, there is a non-zero vector $\alpha \in \mathbb{R}^d$ such that

$$\text{Var}(\alpha^\tau \mathbf{m}(Y)) = 0.$$

Thus $\alpha^\tau \mathbf{m}(Y)$ is a.s. a constant and must equal to its expectation 0. In view of (61), it implies $\alpha^\tau \mathbf{m}(k) = 0$ for all integer $k \in [0, 2^d)$. This leads to $\alpha = 0$, which is a contradiction. \square

Proof of Theorem 4. I. Exact sparsity. With the lower bound on the ‘Oracle risk’, we only need to prove the following to obtain the lower bound in the problem with exact sparsity.

$$(63) \quad \inf_{\hat{\mathbf{V}}} \sup_{\mathcal{M} \in \mathfrak{M}_{s,0}(p,d,\lambda,\kappa)} \mathbb{E}_{\mathcal{M}} \|\hat{\mathbf{V}}\hat{\mathbf{V}}^\tau - \mathbf{V}\mathbf{V}^\tau\|_F^2 \succ d \wedge \frac{s \log \frac{ep}{s}}{n\lambda}.$$

It follows from the arguments in Vu and Lei [2012] and Cai et al. [2013]. More precisely, Vu and Lei [2012] have constructed a set $\Theta' \subset \mathbb{S}^{p-s}$, such that

1. $\delta/\sqrt{2} < \|\beta_1 - \beta_2\|_2 \leq \sqrt{2}\delta$ for all distinct pairs $\beta_1, \beta_2 \in \Theta'$,
2. $\|\beta\|_0 \leq s$ for all $\beta \in \Theta'$,
3. $\log |\Theta'| \geq cs[\log(p-s+1) - \log(s)]$, where $c \geq 0.233$.

Now we consider the following family of models

$$y = f(\mathbf{V}^\tau \mathbf{x}) + \epsilon$$

where $\mathbf{x} \sim N(0, \mathbf{I}_p)$, $\epsilon \sim N(0, 1)$, $\mathbf{V} = \begin{pmatrix} \beta & 0_{(p+1-s) \times (s-1)} \\ 0_{(s-1) \times 1} & I_{s-1} \end{pmatrix}$ and $\beta \in \Theta' \subset S^{p-s}$. The similar Fano type argument near (59) gives us the (63).

II. Weak l_q sparsity. For the lower bound in problems with weak l_q sparsity, we can simply apply the argument of Theorem 2 in Cai et al. [2013]. \square

Assisting Lemmas. The following lemmas have been frequently used during the proofs.

LEMMA 16. *Let \mathbf{K} be an $a \times b$ matrix with each entry being i.i.d. standard normal random variables. Then, we have $\mathbb{E}[\|\mathbf{K}\mathbf{K}^\tau\|_F^2] = ab(a+b+1)$ and $\mathbb{E}[\|\mathbf{K}\|_F^2] = ab$.*

PROOF. It follows from elementary calculations. \square

LEMMA 17. *Let A, B be $l \times m$ and $m \times n$ matrices, respectively, we have $\|AB\|_F \leq \|A\|_2 \|B\|_F$, where $\|A\|_2$ denotes the largest singular value of A .*

PROOF. It follows from elementary calculations. \square

LEMMA 18. *Let A, B be $m \times l$ orthogonal matrices, i.e., $A^T A = I_l = B^T B$, and let M be an $l \times l$ positive definite matrix with eigenvalues d_j such as $0 < \lambda \leq d_l \leq d_{l-1} \leq \dots \leq d_1 \leq \kappa\lambda$. If $A^T B$ is a diagonal matrix with non-negative entries, then there exists a constant C which only depends on κ such that $\|AMA^T - BMB^T\|_F \leq C\lambda\|AA^T - BB^T\|_F$.*

PROOF. Let $\Delta = I_l - B^T A$, then $0 \leq \Delta_{ii} \leq 1$ for $1 \leq i \leq l$. If $C > 2\kappa^2 - 1$, we have

$$\begin{aligned} \|AMA^T - BMB^T\|_F^2 &= 2\text{tr}(M^2\Delta) - \text{tr}(M\Delta M\Delta) \leq 2\kappa^2\lambda^2\text{tr}(\Delta) - \lambda^2\text{tr}(\Delta^2) \\ &\leq C\lambda^2(2\text{tr}(\Delta) - \text{tr}(\Delta^2)) = C\lambda^2\|AA^T - BB^T\|_F^2. \end{aligned}$$

\square

LEMMA 19. *For a positive definite matrix M with eigenvalue $\lambda_1 \geq \dots \geq \lambda_d > 0$ and orthogonal matrices A, B, E, F , i.e., $A^T A = B^T B = E^T E = F^T F = I_d$, we have*

$$\frac{\lambda_d}{2} \|AB^T - EF^T\|_F^2 \leq \langle AMB^T, AB^T - EF^T \rangle \leq \frac{\lambda_1}{2} \|AB^T - EF^T\|_F^2.$$

PROOF. It is a direct corollary of the Lemma 8 in [Gao et al. \[2014\]](#). \square

LEMMA 20 (Sparse approximation). *Let $\mathbf{V} \in \mathbb{O}_{s,q}(p, d)$ and $k \in [p]$, where $\mathbb{O}_{s,q}(p, d)$ is defined near (12). Let $\|\mathbf{V}_{(i)*}\|$ denote its i -th largest row norm. Then*

$$(64) \quad \sum_{i>k} \|\mathbf{V}_{(i)*}\|^2 \leq \frac{q}{2-q} k(s/k)^{2/q}.$$

In particular, if k is chosen to be $k_{s,q}$ defined near (12), we know that

$$(65) \quad \sum_{i>k} \|\mathbf{V}_{(i)*}\|^2 \leq \frac{q}{2-q} \epsilon_n^2.$$

PROOF. This is a direct corollary of the Lemma 7 in [Cai et al. \[2013\]](#). \square

LEMMA 21. Let $\Sigma = \mathbf{V}\mathbf{D}\mathbf{V}^\tau$ be a $p \times p$ positive semidefinite matrix where \mathbf{V} is a $p \times d$ orthogonal matrix and \mathbf{D} is a $d \times d$ diagonal matrix with entries $\lambda \leq d_d \leq \dots \leq d_1 \leq \kappa\lambda$. For a subset S of indexes with $|S| = k$, let J_S be a diagonal matrix such that $J_S(i, i) = 1$ if $i \in S$ and $J_S(i, i) = 0$ if $i \notin S$. Let $\Sigma_S = J_S \Sigma J_S$ and let $\Sigma_S = \mathbf{V}_1 \mathbf{D}_1 \mathbf{V}_1^\tau$ be the eigen-decomposition of Σ_S . We have

$$\|\Sigma - \Sigma_S\|_F \leq 2\|\mathbf{D}\|_2 \|J_S \mathbf{V} - \mathbf{V}\|_F.$$

Furthermore, if $\|J_S \mathbf{V} - \mathbf{V}\|_F \leq \frac{1}{8\kappa}$, then $\|\Sigma - \Sigma_S\|_F \leq \lambda_d/4$. By Sin-Theta Lemma (e.g., Lemma 24), we have

$$\|\mathbf{V}\mathbf{V}^\tau - \mathbf{V}_1 \mathbf{V}_1^\tau\|_F \leq 8\kappa \|J_S \mathbf{V} - \mathbf{V}\|_F.$$

PROOF. This comes from a (trivial) elementary calculus. \square

The following Lemma is adopted from the Lemma 6.5 in Ma and Li [2016].

LEMMA 22. For any matrices $A_1, A_2 \in \mathbb{O}(p, d)$, there exists some $Q \in \mathbb{O}(d, d)$ such that

$$(66) \quad \|A_1 - A_2 Q\|_F \leq \|A_1 A_1^\tau - A_2 A_2^\tau\|_F$$

The following lemmas are borrowed from Vershynin [2010] and Cai et al. [2013].

LEMMA 23. Let $\mathbf{E}_{p \times H}$ be a $p \times H$ matrix, whose entries are independent standard normal random variables. Then for every $t \geq 0$, with probability at least $1 - 2\exp(-t^2/2)$, one has :

$$\lambda_{\text{sing}, \min}^+(\mathbf{E}_{p \times H}) \geq \sqrt{p} - \sqrt{H} - t$$

, and

$$\lambda_{\text{sing}, \max}(\mathbf{E}_{p \times H}) \leq \sqrt{p} + \sqrt{H} + t.$$

LEMMA 24. (Sin-Theta Theorem.) Let \mathbf{A} and $\mathbf{A} + \mathbf{E}$ be symmetric matrices satisfying

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

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

$$\|\mathbf{F}_0 \mathbf{F}_0^\tau - \mathbf{G}_0 \mathbf{G}_0^\tau\| \leq \frac{\min(\|\mathbf{F}_1^\tau \mathbf{E} \mathbf{G}_0\|, \|\mathbf{F}_0^\tau \mathbf{E} \mathbf{G}_1\|)}{\delta},$$

and

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

LEMMA 25. Let $\mathbf{K} \in \mathbb{R}^{p \times p}$ be symmetric such that $\text{Tr}(\mathbf{K}) = 0$ and $\|\mathbf{K}\|_F \leq 1$. Let \mathbf{Z} be an $H \times p$ matrix consisting of independent standard normal entries. Then for any $t > 0$, one has

$$(67) \quad \mathbb{P}\left(\left|\left\langle \mathbf{Z}^\tau \mathbf{Z}, \mathbf{K} \right\rangle\right| \geq 2\sqrt{H}t + 2t^2\right) \leq 2\exp(-t^2).$$

We remind that this lemma is a trivial modification of Lemma 4 in Cai et al. [2013], where they assumed $\|\mathbf{K}\|_F = 1$.

LEMMA 26. Let X_1, \dots, X_N be random variables such that each satisfies

$$(68) \quad \mathbb{P}(|X_i| \geq at + bt^2) \leq c \exp(-t^2)$$

where $a, b, c > 0$. Then

$$(69) \quad \mathbb{E} \max |X_i|^2 \leq (2a^2 + 8b^2) \log(ecN) + 2b^2 \log^2(cN).$$

Beyond the uncorrelated predictors. To avoid trivial but tedious technical arguments, we only sketch the main idea of the proof of Theorem 9 and present the essential difficulties in this subsection.

For simplicity, let us assume that $n = 3Hc$ and divide the samples into three equal sets randomly. We use the first two sets of samples to define the estimate $\widehat{\mathbf{V}}_E$ of \mathbf{V} , an orthogonal matrix satisfying that $\text{col}(\mathbf{V}) = \text{col}(\text{var}(\mathbb{E}[\mathbf{x} | y]))$. To be more precise, we reproduce it below.

Aggregation Estimator $\widehat{\mathbf{V}}_E$. For each $B \in \mathcal{B}_{ks}$, we let

$$(70) \quad \begin{aligned} \widehat{\mathbf{V}}_B &\triangleq \arg \max_{\mathbf{V}} \langle \widehat{\mathbf{\Lambda}}_H^{(1)}, \mathbf{V} \mathbf{V}^\tau \rangle = \arg \max_{\mathbf{V}} \text{Tr}(\mathbf{V}^\tau \widehat{\mathbf{\Lambda}}_H^{(1)} \mathbf{V}) \\ \text{s.t. } &\mathbf{V}^\tau \mathbf{V} = \mathbf{I}_d, \|\mathbf{V}\|_{o,w} = ks \text{ and } \text{supp}(\widehat{\mathbf{V}}_B) \subset B \end{aligned}$$

and

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

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

We first prove that $\|P_{\widehat{\mathbf{V}}_E} - P_{\mathbf{V}}\|_F^2 \leq \frac{s \log(ep/s)}{n\lambda}$ with high probability, then introduce an estimate of $\Sigma^{-1} \mathbf{V}$ based on $\widehat{\mathbf{V}}_E$ and show that it achieves the optimal rate.

In the first step, since $\Sigma^{-1} \text{col}(\mathbf{V}) = \mathcal{S}$ and $\Sigma \neq \mathbf{I}_p$, $\text{col}(\mathbf{V})$ is no longer the central space. Thus, in order to apply the argument for $\Sigma = \mathbf{I}_p$, one has to verify that $y \perp (1 - P_{\mathbf{V}})\mathbf{x}$. We summarized this result into the following lemma.

LEMMA 27. *Suppose that $\mathbf{x} \sim N(0, \Sigma)$, $y = f(\Gamma^\tau \mathbf{x}, \epsilon)$ where Γ is a $p \times d$ orthogonal matrix. Let $\mathbf{x} = P_{\Lambda} \mathbf{x} + (1 - P_{\Lambda}) \mathbf{x} := \mathbf{z} + \mathbf{w}$ be the orthogonal decomposition of \mathbf{x} with respect to $\text{col}(\Lambda)$, then $y \perp \mathbf{w}$.*

PROOF. Let $\tilde{\mathbf{V}}$ be a $p \times d$ orthogonal matrix such that $P_{\tilde{\mathbf{V}}} = P_{\Lambda}$, then we have $P_{\Lambda} \mathbf{x} = P_{\tilde{\mathbf{V}}} \mathbf{x}$, $\mathbf{w} = (1 - P_{\tilde{\mathbf{V}}}) \mathbf{x}$, $P_{\Sigma^{-1} \Lambda} \mathbf{x} = P_{\Sigma^{-1} \tilde{\mathbf{V}}} \mathbf{x}$ and

$$\mathbb{E}[\mathbf{w} (P_{\Sigma^{-1} \tilde{\mathbf{V}}} \mathbf{x})^\tau] = (1 - P_{\tilde{\mathbf{V}}}) \Sigma P_{\Sigma^{-1} \tilde{\mathbf{V}}} = 0$$

In other words, we have $P_{\Sigma^{-1} \Lambda} \mathbf{x} \perp \mathbf{w}$. Since we have $\text{col}(\Gamma) = \Sigma^{-1} \text{col}(\Lambda)$, we know that $y \perp \mathbf{w}$. \square

After carefully checking the argument in the proof for $\Sigma = \mathbf{I}_p$, all we need is to establish some tail bounds of quantity $\langle \mathcal{Z}_H \mathcal{W}_H^\tau, K_B \rangle$. It is clear that we have

$$(71) \quad |\langle \mathcal{Z}_H \mathcal{W}_H^\tau, K_B \rangle| \leq \|\mathcal{Z}_H\|_F \|\mathcal{W}_H^\tau K_B\|_F.$$

There exists a symmetric matrix Σ_1 such that $\sqrt{n} \mathcal{W}_H = \Sigma_1 E$ where E is a $p \times H$ matrix such that entries are i.i.d. $\sim N(0, 1)$. Thus,

$$(72) \quad n \mathcal{W}_H^\tau K_B K_B \mathcal{W}_H = E^\tau \Sigma_1 K_B K_B \Sigma_1 E.$$

Let us consider the spectral decomposition of $\Sigma_1 K_B K_B \Sigma_1 = T \text{diag}\{a_1, \dots, a_p\} T^\tau$ where T is an orthogonal matrix. Since $\|\Sigma_1 K_B\|_F \leq \lambda_{\max}(\Sigma_1) \|K_B\|_F \leq \lambda_{\max}(\Sigma_1)$, we know that

$$(73) \quad \sum_j a_j \leq \lambda_{\max}(\Sigma_1) \leq C$$

for some constant C . Since

$$n \mathcal{W}_H^\tau K_B K_B \mathcal{W}_H \text{ and } E^\tau \text{diag}\{a_1, \dots, a_p\} E$$

share the same distribution, we know that (see e.g., Lin et al. [2018b])

$$(74) \quad \mathbb{P} \left(\left| \|\mathcal{W}_H^\tau K_B\|_F^2 - \mathbb{E} [\|\mathcal{W}_H^\tau K_B\|_F^2] \right| > \alpha \right) \leq 2 \exp^{-C \frac{n^2 \alpha^2}{H \sum_{j=1}^p a_j^2}}$$

for some constant C . Thus,

$$\begin{aligned} \mathbb{P} \left(| \langle \mathcal{Z}_H \mathcal{W}_H^\tau, K_B \rangle | > C\lambda \sqrt{\frac{s \log(p)}{n}} \right) &\leq \mathbb{P} (\|\mathcal{Z}_H\|_F > C\lambda) + \mathbb{P} \left(\|\mathcal{W}_H^\tau K_B\|_F > C\sqrt{\frac{s \log(p)}{n}} \right) \\ &\leq C_1 \left(\exp^{-C_2 n \lambda} + \exp^{-C_3 s \log(p)} \right) \end{aligned}$$

for some constants C_1, C_2 and C_3 . Since we have assumed that $\frac{s \log(p)}{n\lambda}$ is sufficiently small, we have

$$\mathbb{P} \left(\max_B | \langle \mathcal{Z}_H \mathcal{W}_H^\tau, K_B \rangle | > C_1 \lambda \sqrt{\frac{s \log(p)}{n}} \right) \leq C_2 \exp^{-C_3 s \log(p)}$$

for some constants C_1, C_2 and C_3 . Thus, we can prove

$$(75) \quad \inf \sup \mathbb{E} \|P_{\mathbf{V}_E} - P_{\mathbf{V}}\|_F^2 \asymp \frac{s \log(p)}{n\lambda}.$$

For the second step, we need some lemmas.

LEMMA 28. *Let A and B be two $p \times d$ matrices with Frobenius norm bounded below by a constant. If $\|A - B\|_F$ is sufficiently small, there exists constant C such that*

$$(76) \quad \|P_A - P_B\|_F \leq C \|A - B\|_F$$

LEMMA 29. *Let E be a $p \times n$ matrix with i.i.d. standard normal entries. Suppose that $n > p$. We have*

$$(77) \quad \mathbb{P} \left(\left\| \frac{1}{n} E E^\tau - \mathbf{I}_p \right\| > 2\alpha \right) \leq 2 \exp^{-n\alpha^2/2}$$

for any $\alpha > \sqrt{\frac{p}{n}}$.

For the second step, we use the third set of samples to estimate $\Sigma_E = J_E \Sigma J_E$, i.e., $\hat{\Sigma}_E = \frac{1}{Hc} \sum' J_E \mathbf{x}_i \mathbf{x}_i^\tau J_E$ where \sum' stands for summing over the third set of samples. Then we have the desired bound for $\|\hat{\Sigma}_E^- \mathbf{V}_E - \Sigma^{-1} \mathbf{V}\|_F^2$ where A^- stands for the Moore-Penrose inverse of a semi-definite matrix A , i.e., if $A = T \text{diag}\{a_1, \dots, a_k, 0, \dots, 0\} T^\tau$ for some orthogonal matrix and positive number a_1, \dots, a_k , then $A^- = T \text{diag}\{a_1^{-1}, \dots, a_k^{-1}, 0, \dots, 0\} T^\tau$.

With Lemma 22, we could assume that V and V_E satisfying that

$$\|V - V_E\|_F \leq \|P_V - P_{V_E}\|_F$$

Once we have bound on $\|\widehat{\Sigma}_E^- \mathbf{V}_E - \Sigma^{-1} \mathbf{V}\|_F^2$, Lemma 28 yields the bound on

$$\|P_{\widehat{\Sigma}_E^- \mathbf{V}_E} - P_{\Sigma^{-1} \mathbf{V}}\|_F^2$$

. We only need to verify that

$$\begin{aligned} \widehat{\Sigma}_E^- \mathbf{V}_E - \Sigma^{-1} \mathbf{V} &= \widehat{\Sigma}_E^- \mathbf{V}_E - \Sigma_E^{-1} \mathbf{V}_E \\ &\quad + \Sigma_E^{-1} \mathbf{V}_E - \Sigma_E^{-1} \mathbf{V} + \Sigma_E^{-1} \mathbf{V} - \Sigma^{-1} \mathbf{V} \end{aligned}$$

could be bounded by $C \frac{s \log(ep/s)}{n\lambda}$ with probability at least $1 - \exp\left(-C \frac{s \log(p)}{\lambda}\right)$ which follows from the following Lemmas and the estimation of the risk $\|P_{\mathbf{V}_E} - P_{\mathbf{V}}\|_F$

LEMMA 30.

$$(78) \quad \|\widehat{\Sigma}_E^- \mathbf{V}_E - \Sigma_E^{-1} \mathbf{V}_E\|_F \leq \|\widehat{\Sigma}_E^- - \Sigma_E^{-1}\|_2 \leq C_1 \sqrt{\frac{s \log(p)}{n\lambda}}$$

with probability at least $1 - C_2 \exp\left(-C_3 \frac{s \log(p)}{\lambda}\right)$.

PROOF. For any invertible matrices A_1, A_2 , we have $A_1^{-1} - A_2^{-1} = A_1^{-1}(A_1 - A_2)A_2^{-1}$. Fro any deterministic $B \in [p]$ with $|B| = ks$, we have

$$\mathbb{P}\left(\|J_B \widehat{\Sigma} J_B - J_B \Sigma J_B\| > C_1 \sqrt{\frac{s \log(p)}{n\lambda}}\right) \leq C \exp^{-C_1^2 \frac{s \log(p)}{\lambda}}.$$

Thus,

$$\mathbb{P}\left(\|J_E \widehat{\Sigma} J_E - J_E \Sigma J_E\| > C_1 \sqrt{\frac{s \log(p)}{n\lambda}}\right) \leq C \binom{p}{ks} \exp^{-C_1^2 \frac{s \log(p)}{\lambda}} \leq C \exp^{-C' \frac{s \log(p)}{\lambda}}.$$

□

LEMMA 31.

$$(79) \quad \|\Sigma_E^- \mathbf{V}_E - \Sigma_E^- \mathbf{V}\|_F \leq C \|\mathbf{V}_E - \mathbf{V}\|_F \leq C \sqrt{\frac{s \log(p)}{n\lambda}}$$

with probability at least $1 - \exp\left(-C \frac{s \log(p)}{\lambda}\right)$

PROOF. This follows from Lemma 30.

LEMMA 32. *If there exists a constant C such that for any $K \subset [p]$, $\|J_K \Sigma^{-1} \mathbf{V}\|_F \leq C \|J_K \mathbf{V}\|_F$, then for $C' = C + \|\Sigma^{-1}\|_2$,*

$$(80) \quad \|\Sigma_E^- \mathbf{V} - \Sigma^{-1} \mathbf{V}\|_F \leq C' \|J_E \mathbf{V} - \mathbf{V}\|_F$$

PROOF. It is a simple linear algebra exercise.