

# ON THE OPTIMALITY OF SLICED INVERSE REGRESSION IN HIGH DIMENSIONS

BY QIAN LIN<sup>†\*</sup> XINRAN LI<sup>†\*</sup> DONGMING HUANG<sup>†\*</sup> AND JUN S. LIU<sup>‡\*</sup>

*Harvard University<sup>†</sup>*

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  $y \perp\!\!\!\perp \mathbf{x} | P_{\mathcal{S}}\mathbf{x}$ . In this paper, we consider the minimax rate of estimating the central space of the multiple index models  $y = f(\boldsymbol{\beta}_1^\top \mathbf{x}, \boldsymbol{\beta}_2^\top \mathbf{x}, \dots, \boldsymbol{\beta}_d^\top \mathbf{x}, \epsilon)$  with at most  $s$  active predictors where  $\mathbf{x} \sim N(0, \mathbf{I}_p)$ . We first introduce a large class of models depending on the smallest non-zero eigenvalue  $\lambda$  of  $var(\mathbb{E}[\mathbf{x}|y])$ , 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  $\lambda$ . 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. We believe that these (conditional) optimal rate results bring us meaningful insights of general SDR problems in high dimensions.

**1. Introduction.** Because of rapid advances of 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 reductions are often indispensable first steps. In early 1990s, a fascinating supervised dimension reduction method, the sliced inverse regression (SIR) [Li, 1991], was proposed to model univariate response with a low dimensional projection of the predictors. More precisely, SIR postulates the following *multiple index model* for the data:

$$(1) \quad y = f(\boldsymbol{\beta}_1^\top \mathbf{x}, \boldsymbol{\beta}_2^\top \mathbf{x}, \dots, \boldsymbol{\beta}_d^\top \mathbf{x}, \epsilon),$$

and estimates the subspace  $\mathcal{S} = span\{ \boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_d \}$  via an eigen-analysis of the estimated conditional covariance matrix  $var[\mathbb{E}(\mathbf{x}|y)]$ . Note that the individual  $\boldsymbol{\beta}_i$ ’s are not identifiable, but the space  $\mathcal{S}$  can be estimated well.

---

\*Lin’s research is supported by the Center of Mathematical Sciences and Applications at Harvard University. Liu’s research is supported 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

Based on the observation that  $y \perp\!\!\!\perp \mathbf{x} | P_{\mathcal{S}}\mathbf{x}$ , Cook [1998] proposed a more general framework for dimension reduction without loss of information, often referred to as the *Sufficient Dimension Reduction* (SDR). Under this framework, researchers look for the minimal subspace  $\mathcal{S}' \subset \mathbb{R}^p$  such that  $y \perp\!\!\!\perp \mathbf{x} | P_{\mathcal{S}'\mathbf{x}}$  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 among practitioners because of its simplicity and computational efficiency. Asymptotic theories developed for these SDR algorithms have all focused on scenarios where the data dimension  $p$  is either fixed or growing at a much slower rate compared with the sample size  $n$  [Dennis Cook, 2000, Li and Wang, 2007, Li, 2000]. The ‘*large p, small n*’ characteristic of modern data raises new challenges to these SDR algorithms.

Lin et al. [2015] 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 the structural assumption such as sparsity 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 dimension reduction algorithms such as SIR. Motivated by the 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. [2015] and the Sparse-SIR algorithm in Lin et al. [2016b], however, have been shown to provide consistent estimations. We agree with Cook et al. [2012] that a detailed understanding of “the behaviour of these SDR estimators when  $n$  is not large relative to  $p$ ” might be the key to efficient high-dimensional SDR algorithms. 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 is not only interesting in its own right, but will also inform the development and evaluation of other SDR algorithms developed for high-dimensional problems.

Neykov et al. [2015] considered the (signed)-support recovery problem of the following class of single index models

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

where  $\boldsymbol{x} \sim N(0, \mathbf{I}_p)$ ,  $\epsilon \sim N(0, 1)$ . Let  $\xi = \frac{n}{s \log(p)}$ , they proved that **1**) If  $\xi$  is sufficiently small, any algorithm fails to recover the (signed) support of  $\beta$  with probability at least 1/2 and **2**) If  $\xi$  is sufficiently large, the DT-SIR algorithm (see Lin et al. [2015] 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  $\beta$  is of order  $s \log(p)$ . These results shed us 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 models, 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] for SIR when  $\Sigma = \mathbf{I}$ . However, as it was pointed out in Lin et al. [2015], the sample means in corresponding slices 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. [2015]. On the other hand, the observation that the number  $H$  of slices is allowed to be finite when  $d$  is bounded (as 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 dimensional 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 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. 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 construct an estimator (computationally unrealistic) such that the risk of this estimator is of order  $\frac{ds + s \log(ep/s)}{n\lambda} \wedge d$ . Under mild conditions, we further demonstrate that the risk of any estimator is bounded below by  $\frac{s \log(ep/s)}{n\lambda} \wedge 1$  if the dimension of the central space  $d$  is bounded. Thus, the minimax rate of the risk  $\mathbb{E}[\|P_{\mathbf{V}} - P_{\hat{\mathbf{V}}}\|_F^2]$  is  $\frac{ds + s \log(ep/s)}{n\lambda} \wedge d$  if  $d$  is bounded. One of the key components of our analysis

is the linear algebraic Lemma 16, which might be of independent interest and be used in determining lower bounds of minimax estimation rates for other dimension reduction problems. 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 2.6, we show that the computationally efficient algorithm DT-SIR [Lin et al., 2015] 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., 2016a] and give the proofs of upper bounds in Sections 4.1 and 4.2. 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., 2016a].

**2. Main Results.** Since the establishment of the SDR framework about two decades ago, estimating the central space has been investigated under different assumptions [Dennis Cook, 2000, 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) [Dennis Cook, 2000] 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 well understood SDR algorithms, and is of special interests 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  $col(\mathbf{V})$  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 numbers  $a, b$ , we use  $a \vee b$  and  $a \wedge b$  to denote  $\max\{a, b\}$  and  $\min\{a, b\}$ , respectively.

For a matrix  $A$ ,  $\|A\|_F = \text{tr}(AA^\top)^{1/2}$ . For a positive integer  $p$ ,  $[p]$  denotes the index set  $\{1, 2, \dots, p\}$ . 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  $\beta_i$ '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  $\mathbf{x} \sim N(0, \mathbf{I}_p)$ ,  $\epsilon \sim N(0, 1)$ , and  $\mathbb{O}(p, d)$  is the set of all  $p \times d$  orthogonal matrices. Though  $\mathbf{V}$  is not identifiable, the column space  $\text{col}(\mathbf{V})$  is estimable. 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  $y_{(i)}$ .<sup>1</sup> 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  $y_{(k)}$ . Let the sample mean in the  $h$ -th slice be  $\bar{\mathbf{x}}_{h,\cdot}$ , and the overall sample mean be  $\bar{\mathbf{x}}$ . SIR estimates  $\Lambda \triangleq \text{var}(\mathbb{E}[\mathbf{x}|y])$  by

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

and estimates the central space  $\text{col}(\mathbf{V})$  by  $\text{col}(\hat{\mathbf{V}}_H)$  where  $\hat{\mathbf{V}}_H$  is the matrix formed by the top  $d$  eigenvectors of  $\hat{\Lambda}_H$ . Throughout this article, we assume that  $d$ , dimension of the central space, is known.

In order for the SIR to give a consistent estimate of the central space, following sufficient conditions have been suggested (e.g., Li [1991], Hsing and Carroll [1992] and Zhu et al. [2006]):

**A')** Linearity Condition and Coverage Condition:

---

<sup>1</sup>To ease notations and arguments, we assume that  $n = cH$ .

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

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

(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  $(B_0, \infty)$ , such that

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

As in Lin et al. [2015], where they demonstrated the phase transition phenomenon of SIR in high dimension, we replace Condition (B') by

(B'') Modified Smoothness and Tail conditions,

which is all the same as (B') except that eqn (4) is replaced by

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

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. [2015].

**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.** *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  $\vartheta$ -sliced stable with respect to  $Y$ , if there exist positive constants  $\gamma_1, \gamma_2, \gamma_3$  such that 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  $\vartheta$ -sliced stable for some positive constant  $\vartheta$ .

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 placed before  $\text{var}(\beta^\tau \mathbf{x})$  on the RHS of (6). Thus, the sliced stable condition is almost the necessary condition to ensure that the SIR works. A main advantage of the sliced stable condition is that we can easily quantify the deviation properties of the eigenvalues, eigenvectors, and each entries of  $\widehat{\Lambda}_H$ . This is one of the main technical contributions of Lin et al. [2015]. We henceforth assume that the central curve satisfies the sliced stable condition. As shown by Proposition 1, Condition (B'') ensures the sliced-stable condition.

2.3. *The class of functions  $\mathcal{F}_d(\lambda, \kappa)$ .* Let  $\mathbf{z} = \mathbf{V}^\tau \mathbf{x}$ , then  $\mathbf{z} \sim N(0, \mathbf{I}_d)$ . Let  $\boldsymbol{\Lambda}_{\mathbf{z}} = \text{var}(\mathbb{E}[\mathbf{z}|y])$ . 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. The coverage condition, which requires  $\text{rank}(\text{var}(\mathbb{E}[\mathbf{x}|y])) = d$ , can be refined as

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

for some positive constant  $\kappa > 1$ . Since  $\boldsymbol{\Lambda} \triangleq \text{var}(\mathbb{E}[\mathbf{x}|y]) = \mathbf{V} \boldsymbol{\Lambda}_{\mathbf{z}} \mathbf{V}^\tau$ , we know  $\lambda_j(\boldsymbol{\Lambda}) = \lambda_j(\boldsymbol{\Lambda}_{\mathbf{z}})$ ,  $j = 1, \dots, d$ . 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  $\kappa$  is assumed to be a fixed constant. This coverage condition is commonly adopted in the literature (e.g., Cai et al. [2013] and Gao et al. [2014]) when researchers discuss the dimension reduction problems. The class of functions  $f$  satisfying the sliced stable condition and coverage condition is of our main interests in this paper. More precisely, we introduce  $\mathcal{F}_d(\lambda, \kappa)$  as 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 are satisfied.

- (A) Coverage condition:  $0 < \lambda \leq \lambda_d(\boldsymbol{\Lambda}_z) \leq \dots \leq \lambda_1(\boldsymbol{\Lambda}_z) \leq \kappa \lambda \leq 1$ , where  $\boldsymbol{\Lambda}_z \triangleq \text{var}(\mathbb{E}[\mathbf{z}|f(\mathbf{z}, \epsilon)])$ .

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

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

**2.4. Upper bound of the risk.** 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  $\mathbb{E}_{\mathcal{M}} \|P_{\widehat{\mathbf{V}}} - P_{\mathbf{V}}\|_F^2$  where  $P_{\widehat{\mathbf{V}}}$  is an estimate of  $P_{\mathbf{V}}$  based on these samples. In this subsection, we provide an upper bound of this risk. All detailed proofs are deferred to Sections 4.1, 4.2, and online supplementary file [Lin et al., 2016a].

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

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

A main result of this article is:

**THEOREM 1** (An Upper Bound of Oracle Risk). *Assume that  $\frac{dp}{n\lambda}$  is sufficiently small and  $d^2 \leq p$ . We have*

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

In order to establish the upper bound, we consider the estimate  $\widehat{\mathbf{V}}_H$ , which is a  $p \times d$  orthogonal matrix forming by the top- $d$  eigenvectors of  $\widehat{\Lambda}_H$ , and show that  $P_{\widehat{\mathbf{V}}_H}$  achieves the rate in Theorem 1.

A result in Lin et al. [2015], which states that

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

appears to contradict our Theorem 1 here: (i) it does not depend on  $d$ , the dimension of central subspace; (ii) it does not depend on  $\lambda$ , the smallest non-zero eigenvalue of  $\text{var}(\mathbb{E}[\mathbf{x}|y])$ ; (iii) it 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. [2015] 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  $\widehat{\Lambda}_H$  might be different (slower) than the convergence rate of  $P_{\widehat{\mathbf{V}}_H}$ . More precisely, we have

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

From the proof of Theorem 1 of Lin et al. [2015], 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  $P_{\mathbf{V}} \widehat{\Lambda}_H P_{\mathbf{V}}$  and  $\Lambda$  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. [2015].

**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  $\text{col}(\widehat{\mathbf{V}}_H)$  is an optimal estimate of  $\text{col}(\mathbf{V})$ . In the SIR literature, researchers care about the eigenvectors of  $\Lambda$  and ignore the eigenvalue information. In this article, we show that the larger the  $H$ , the more accurate the estimate of the eigenvalues of  $\Lambda$ , and illustrate this phenomenon via numerical simulations in Section 3.1. Taking the eigenvalue information into consideration will bring us more a detailed understanding of SIR.

**2.4.2. Upper bound of the risk of sparse SIR.** Lin et al. [2015] 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 problem. 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.,  $\mathbf{V}^\top \mathbf{V} = \mathbf{I}_d$ ), 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

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

and  $\lceil a \rceil$  denotes the smallest integer no less than  $a \in \mathbb{R}$ . For more detailed discussions of the sparse orthogonal matrices, we refer to Cai et al. [2013].

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 of Optimal Rates). *Assume that  $\kappa$  is fixed,  $d^2 \leq k_{q,s}$ ,  $\epsilon_n^2$  is sufficiently small and  $n\lambda \leq e^p$ . We have*

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

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 randomly into two equal size sets. Let  $\widehat{\boldsymbol{\Lambda}}_H^{(1)}$  and  $\widehat{\boldsymbol{\Lambda}}_H^{(2)}$  be the SIR estimates of  $\boldsymbol{\Lambda} = \text{var}(\mathbb{E}[\mathbf{x}|y])$  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  $\widehat{\mathbf{V}}_E$  of  $\mathbf{V}$ .

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

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

and

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

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

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

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

for any  $\widehat{\mathbf{V}}_B$  where  $B \in \mathcal{B}$ . In Section 4.2, we will show that the aggregation estimator  $\widehat{\mathbf{V}}_E$  achieves the converges rate on the right hand side of (15).

**2.5. Lower Bound and Minimax Risk.** To avoid unnecessary details, we assume that dimension  $d$  of the central space is bounded in this subsection. The semi-parametric characteristic of multiple index models 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)  $\lambda$ , the smallest non-zero eigenvalue of  $\text{var}(\mathbb{E}[\mathbf{x}|y])$ , is bounded below by some positive constant; and (ii) single index models where  $d = 1$ . To the best of our knowledge, even 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. Furthermore, we have observed from extensive numerical studies that the 4-th direction is difficult to detect for  $p = 10$  even with the sample size greater than  $10^6$ . This observation conforms to the existing numerical studies reported in the literature, i.e., most researchers only reported numerical studies for models with  $d \leq 2$  except that Ferré [1998] performed a numerical study for a model with  $d = 4$  and reported that the 4-th direction was hard to discover. Thus, the optimal rate with  $d$  bounded might be a more reasonable target to pursue.

**2.5.1.  $\lambda$  is bounded below by some positive constant.** Assume that  $\lambda$ , the smallest non-zero eigenvalues of  $\text{var}(\mathbb{E}[\mathbf{x}|y])$ , is bounded below by a positive constant. We have the following optimal convergence rate of the *Oracle Risk*.

**THEOREM 3 ( Oracle Risk).** *Assume that  $d, \lambda$  are bounded. We have*

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

**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., 2016a] and briefly sketch its key steps here. One of the key steps in obtaining the lower bound is constructing 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

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

for some absolute constants  $c_0$  and  $c_1$ . For any  $\theta_j \in \Theta$ , if we can choose a  $p \times d$  orthogonal matrix  $\mathbf{B}_j$  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

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

where  $f$  is a  $d$ -variate function with bounded first derivative such that these models belong to  $\mathfrak{M}(p, d, \lambda, \kappa)$ . Let  $p_{f, \mathbf{B}}$  denote the joint density of  $(y, \mathbf{x})$ . Simple calculation shows that

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

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). It seems to be a simple linear algebraic problem, however, its proof requires (slightly) non-trivial work in differential geometry (cf. Lemma 16). <sup>2</sup> Thus we know that the rate in Theorem 1 is optimal if  $d$  and  $\lambda$  are bounded. Once the ‘Oracle risk’ has been established, the standard argument in Cai et al. [2013] leads us the following:

---

<sup>2</sup>Q. Lin appreciates the helpful discussions with Dr. Long Jin

**THEOREM 4 (Optimal Rates ).** *Assume that  $d, \lambda$  are bounded, and  $n\lambda \leq e^p$ . We have*

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

PROOF. See the online supplementary file [Lin et al., 2016a].

**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  $n, \lambda, s$ , and  $p$ .

**THEOREM 5 (Oracle Risk for Single Index Models).** *Assuming that  $d = 1$  and  $n\lambda \leq e^p$ , we have*

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

Since we have proved Theorem 2, all we need to do is to establish a suitable lower bound. Let us consider the following linear model:

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

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

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

Thus, inequality (19) becomes

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

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  $d = 1$  and  $n\lambda \leq e^p$ . We have*

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

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  $f_\lambda$  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 relaxed to that  $\|\nabla f_\lambda(x)\| \leq C\sqrt{\lambda}$  holds with high probability when  $x \sim N(0, \mathbf{I}_d)$ .

The construction in subsection 2.5.2 shows that this conjecture holds for  $d = 1$ . For any  $d > 1$ , 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

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

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

**THEOREM 7 ( Oracle Risk :  $d$  is bounded).** *Assuming that  $d$  is bounded and Conjecture 1 holds, we have*

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

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

**THEOREM 8 (Optimal Rates :  $d$  is bounded ).** *Assuming that  $d$  is fixed,  $n\lambda \leq e^p$  and Conjecture 1 holds, we have*

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

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

2.6. *Optimality of DT-SIR.* In the previous section, we have proved that the aggregation estimator  $\widehat{\mathbf{V}}_E$  is rate optimal. In practice, however, it is computationally too expensive. The DT-SIR algorithm proposed in Lin et al. [2015] is computationally efficient in general, and can be further simplified when  $\Sigma_{\mathbf{x}} = \mathbf{I}$ .

---

**Algorithm 1** DT-SIR

---

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

In this section, we focus on the single index model with the exact sparsity on the loading vector  $\beta$ , i.e.,  $|supp(\beta)| = s$ .

**THEOREM 9.** *Suppose  $s = O(p^{1-\delta})$  for some  $\delta > 0$ ,  $\frac{s \log(p)}{n\lambda}$  is sufficiently small and  $n = O(p^C)$  for some constant  $C$ . Let  $\widehat{\beta}_{DT}$  be the DT-SIR estimate with threshold level  $t = C_1 \frac{\log(p)}{n}$  for some constant  $C_1$ , then we have*

$$(28) \quad \|P_{\widehat{\beta}_{DT}} - P_{\beta}\|^2 \leq C_2 \frac{s \log(p-s)}{n\lambda}$$

*with probability at least  $1 - C_4 \exp(-C_3 \log(p))$  for some positive constants  $C_2, C_3$  and  $C_4$ .*

**PROOF.** See the online supplementary file [Lin et al., 2016a].

From Theorem 9, it is easy to see that, if  $s = O(p^{1-\delta})$ , the DT-SIR estimator  $P_{\widehat{\beta}_{DT}}$  is rate optimal. Since there is a computational barrier for the rate optimal estimate of sparse PCA [Berthet and Rigollet, 2013], the fact that the computationally efficient DT-SIR algorithm achieves the optimal rate suggests that sparse PCA might not be an appropriate prototype of SIR in high dimensions.

**3. Numerical Studies.** In this section, 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  $H$  in SIR: the larger the  $H$ , the more accurate the estimate of eigenvalue of  $var(\mathbb{E}[\mathbf{x}|y])$ . The second experiment aims at providing supporting evidence of 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 of estimating the eigenvalues of  $\text{var}(\mathbb{E}[\mathbf{x}|y])$  depends on the choice of  $H$ . Let us consider the following linear model:<sup>3</sup>

$$(29) \quad \text{Model } \mu : y = \sqrt{\frac{\mu}{1-\mu}} \mathbf{x}_1 + \epsilon, \mathbf{x} \sim N(0, \mathbf{I}_p), \epsilon \sim N(0, 1).$$

It is easy to see that the only non-zero eigenvalue of  $\text{var}(\mathbb{E}[\mathbf{x}|y])$  is  $\mu$ . The results are shown in Table 1, where  $H$  ranges in  $\{2, 5, 10, 50, 100, 200, 500\}$ ,  $\mu$  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  $\mu, n$  and  $H$ .

	$n/1000$	$H = 2$	$H = 5$	$H = 10$	$H = 50$	$H = 100$	$H = 200$	$H = 500$
$\mu = .5$	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)
		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)
		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)
	10	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)
		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)
		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)
		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)
		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)
$\mu = .3$	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)
		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)
		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)
	50	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)
		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)

TABLE 1  
The empirical mean (standard error) of the SIR estimate  $\hat{\lambda}(\mu)$  for  $\mu$

From Table 1, it is clear that the larger the  $H$  is, the more accurate estimation of the eigenvalue is. Cautious reader may notice that, in the row

---

<sup>3</sup>Up to a monotone transform, this is the only case that we can give the explicit value of  $\lambda(\text{var}(\mathbb{E}[\mathbf{x}|y]))$ .

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  $H = 10$  (or 50). 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  $\mu = 0.1$ ,  $n/1000 = 5$  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. Support Evidences of Conjecture 1.** Let us consider the following model with two indexes:

$$(30) \quad \text{Model } \mu : y = \sqrt{\mu}(1 + g(\mathbf{x}_1))(g(\mathbf{x}_1) + g(\mathbf{x}_2)) + \epsilon$$

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

$$(31) \quad g(x) = \begin{cases} x & \text{if } |x| \leq 100 - \delta \\ 0 & \text{if } |x| \geq 100 + \delta \end{cases}$$

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 (30) 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  $\mu$  range in  $\{1, .5, .1, .05, .01, .005, .001\}$  and  $n$  range in  $\{10^3, 10^5, 10^5, 10^6\}$ .

	$n$	$\mu = 1$	$\mu = .5$	$\mu = .1$	$\mu = .05$	$\mu = .01$	$\mu = .005$	$\mu = .001$
$\lambda_1(\mu)/\mu$	$n = 10^3$	0.3358	0.5969	1.6333	2.1297	4.5681	7.1434	30.6206
	$n = 10^4$	0.3276	0.5676	1.4416	1.7627	2.1511	2.3908	4.3137
	$n = 10^5$	0.3272	0.5650	1.4153	1.7092	1.9662	2.0006	2.2554
	$n = 10^6$	0.3268	0.5651	1.4125	1.7052	1.9465	1.9695	1.9780
$\lambda_2(\mu)/\mu$	$n = 10^3$	0.1068	0.1436	0.3227	0.5381	2.3159	4.2023	19.5854
	$n = 10^4$	0.0899	0.1087	0.1248	0.1206	0.2701	0.5061	2.3384
	$n = 10^5$	0.0899	0.1059	0.1014	0.0840	0.0462	0.0620	0.2366
	$n = 10^6$	0.0898	0.1063	0.1001	0.0795	0.0297	0.0190	0.0278

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

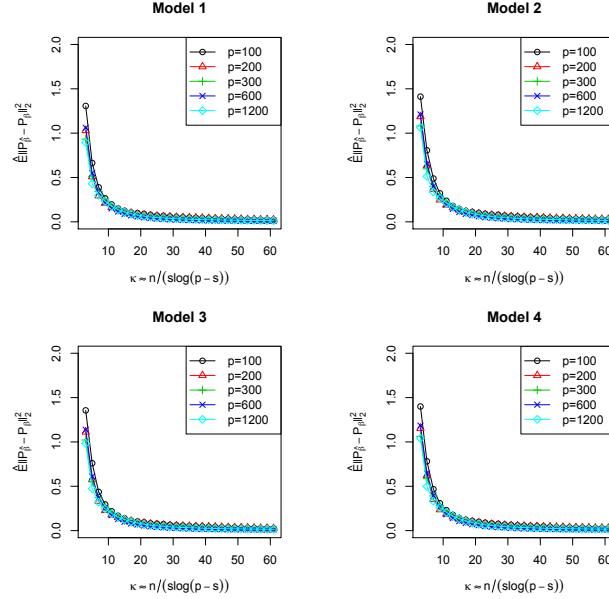
In Table 2, each entry is the average of 100 replications. For fixed  $\mu$ , the larger  $n$ , the more accurate estimation of  $\lambda_i(\mu)/\mu$ ,  $i = 1, 2$ . In particular,

it is easy to see from the row with  $n = 10^6$  that  $\lambda_i(\mu)/\mu, i = 1, 2$ , are bounded. The row with  $n = 10^3$  seems to be contradicting to our conjecture 1, where  $\lambda_i(\mu)/\mu, i = 1, 2$ , might be diverging as  $\mu \rightarrow 0$ . This is actually not a contradiction, since we know that the deviation property of  $\lambda(var(\mathbb{E}[\mathbf{x}|y]))$  depends on the product  $n\lambda$  from Lemma 1. Thus, to get accurate estimate of  $\lambda(var(\mathbb{E}[\mathbf{x}|y]))$ , we require more samples if  $\lambda$  is small.

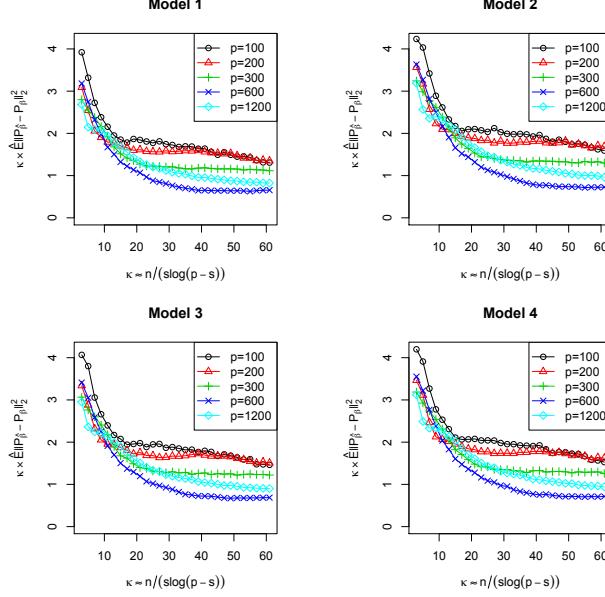
**3.3. Performance of DT-SIR.** In this section, we assume the exact sparsity  $s = O(p^{1-\delta})$  for some  $\delta \in (0, 1)$ , and consider the following data generating models,

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

where  $\mathbf{x} \sim N(\mathbf{0}, \mathbf{I}_p)$ ,  $\epsilon \sim N(0, 1)$ ,  $\mathbf{x} \perp \!\!\! \perp \epsilon$ , and  $\boldsymbol{\beta}$  is a fixed vector with  $s$  nonzero coordinates. Let  $\kappa = \{s \log(p-s)/n\}^{-1}$ . The dimension  $p$  of the predictors takes value in  $\{100, 200, 300, 600, 1200\}$ , the sparsity parameter  $\delta$  is fixed at 0.5, and  $\kappa$  takes values in  $\{3, 5, 7, \dots, 61\}$ . For each  $(p, \kappa)$  combination,  $s = \lfloor p^{1-\delta} \rfloor$ ,  $n = \lfloor \kappa s \log(p-s) \rfloor$ , 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  $(p, \kappa)$  combination are shown in Figure 1, which shows the distance between the estimated projection matrix and the true one becomes smaller as  $\kappa$  increases for all fixed  $p$ .

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

According to Theorem 9,  $\kappa \|P_{\hat{\beta}_{DT}} - P_{\beta}\|^2$  is less than a constant with high probability. Therefore, we also the average values of  $\kappa * \|P_{\hat{\beta}_{DT}} - P_{\beta}\|^2$  for these models in Figure 2, which demonstrates that  $\kappa \|P_{\hat{\beta}_{DT}} - P_{\beta}\|^2$  is a decreasing function of  $\kappa$  and tends to be stable when  $\kappa$  becomes large enough. These empirical results also validate Theorem 9.

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

**4. Proofs.** We need the following technical lemma, which is a direct corollary of the ‘key lemma’ in Lin et al. [2015] :

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  $\text{var}(\mathbb{E}[x|y])(=\Lambda)$ . 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  $\beta$  that lies in the column space of  $\Lambda$ , we have

$$(32) \quad |\beta^\top (\hat{\Lambda}_H - \Lambda) \beta| > \frac{1}{2\nu} \beta^\top \Lambda \beta$$

with probability at most

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

In particular, if  $d$  and  $\nu$  are bounded, we can choose  $H$  to be a large enough finite integer such that (32) holds with high probability.

PROOF. It is a direct corollary of the ‘key lemma’ in Lin et al. [2015].  $\square$

4.1. *Proof of Theorem 1.* 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  $H = H_1d$  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  $\Lambda_H$ . 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  $\mathbf{V}^\perp \mathbf{V}^\perp = 0$ . 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  $\bar{\mathbf{z}}_{h,.}$ ,  $\bar{\mathbf{w}}_{h,.}$ , and  $\bar{\boldsymbol{\epsilon}}_{h,.}$  similar to the definition of  $\bar{\mathbf{x}}_{h,.}$ . Let  $\mathcal{Z} = \frac{1}{\sqrt{H}}(\bar{\mathbf{z}}_{1,.}, \bar{\mathbf{z}}_{2,.}, \dots, \bar{\mathbf{z}}_{H,.})$ ,  $\mathcal{W} = \frac{1}{\sqrt{H}}(\bar{\mathbf{w}}_{1,.}, \bar{\mathbf{w}}_{2,.}, \dots, \bar{\mathbf{w}}_{H,.})$ ,  $\mathcal{E} = \frac{1}{\sqrt{H}}(\bar{\boldsymbol{\epsilon}}_{1,.}, \bar{\boldsymbol{\epsilon}}_{2,.}, \dots, \bar{\boldsymbol{\epsilon}}_{H,.})$  be three  $p \times H$  matrices formed by the vectors  $\frac{1}{\sqrt{H}}\bar{\mathbf{z}}_{h,.}$ ,  $\frac{1}{\sqrt{H}}\bar{\mathbf{w}}_{h,.}$ , and  $\frac{1}{\sqrt{H}}\bar{\boldsymbol{\epsilon}}_{h,.}$ . We have the following decomposition

$$(33) \quad \begin{aligned} \hat{\Lambda}_H &= \mathcal{Z}\mathcal{Z}^\tau + \mathcal{Z}\mathcal{W}^\tau + \mathcal{W}\mathcal{Z}^\tau + \mathcal{W}\mathcal{W}^\tau \\ &= \Lambda_u + \mathcal{Z}\mathcal{E}^\tau \mathbf{V}^{\perp,\tau} + \mathbf{V}^\perp \mathcal{E}\mathcal{Z}^\tau + \mathbf{V}^\perp \mathcal{E}\mathcal{E}^\tau \mathbf{V}^{\perp,\tau} \end{aligned}$$

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  $N(0, \frac{1}{n})$ . 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

$$\mathbb{P}(\|\mathcal{W}\mathcal{W}^\tau\| > C_1\rho) \leq C_2 \exp(-C_3p).$$

ii) For any vector  $\boldsymbol{\beta} \in \mathbb{R}^p$  and any  $\nu > 1$ , let  $E_{\boldsymbol{\beta}}(\nu) = \left\{ \left| \boldsymbol{\beta}^\tau (\Lambda_u - \Lambda) \boldsymbol{\beta} \right| > \frac{1}{2\nu} \boldsymbol{\beta}^\tau \Lambda \boldsymbol{\beta} \right\}$ . 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

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

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

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

PROOF. *i)* is a direct corollary of Lemma 19. *ii)* is a direct corollary of Lemma 1 and the usual  $\epsilon$ -net argument. *iii)* is a direct corollary of *i)* and *ii)*  $\square$

Let  $E = E_1 \cap E_2 \cap E_3$  where  $E_1 = \left\{ \|\mathcal{WW}^\tau\| \leq C\rho \right\}$ ,  $E_2 = \left\{ \|\mathcal{WZ}^\tau\| \leq 4\sqrt{\kappa\lambda\rho} \right\}$ ,  $E_3 = \left\{ \|\Lambda_u - \Lambda\| \leq \frac{1}{2\nu}\kappa\lambda \right\}$ .

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

- a) *If  $n\lambda \leq e^p$ , we have  $\mathbb{P}(E^c) \leq \frac{CH^2}{n\lambda}$ .*
- b) *Conditioning on  $E_3$ , we have  $\lambda_d(\Lambda_u) \geq (1 - \frac{\kappa}{2\nu})\lambda$ .*
- c) *Conditioning on  $E$ , if  $\frac{p}{n\lambda}$  is sufficiently small, we have  $\|\widehat{\Lambda}_H - \Lambda_u\| \leq C\sqrt{\frac{\kappa\lambda p}{n}}$ .*
- d) *Conditioning on  $E$ , If  $\frac{p}{n\lambda}$  is sufficiently small, we have  $\lambda_{d+1}(\widehat{\Lambda}_H) < \frac{1}{4}\lambda$ .*

Now we start the proof of Theorem 1. Note that

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

For *II*. It is easy to see that

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

For *I*. Let  $\Lambda_u = \widetilde{\mathbf{V}}\mathbf{D}_H\widetilde{\mathbf{V}}^\tau$  be the spectral decomposition of  $\Lambda_u$ , where  $\widetilde{\mathbf{V}}$  is a  $p \times d$  orthogonal matrix and  $\mathbf{D}_H$  is a  $d \times d$  diagonal matrix. Conditioning on  $E$ , 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 20) to the pair of symmetric matrices  $(\Lambda_u, \widehat{\Lambda}_H = \Lambda_u + \mathbf{Q})$  where  $\mathbf{Q} \triangleq \widehat{\Lambda}_H - \Lambda_u$ .

Since  $\frac{p}{n\lambda}$  is sufficiently small, conditioning on  $\mathbf{E}$ , we have  $\lambda_{d+1}(\widehat{\Lambda}_H) \leq \frac{1}{4}\lambda$  and  $\lambda_d(\Lambda_u) = \lambda_d(\mathbf{D}_H) \geq \frac{\lambda}{2}$ . Thus, we have

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

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

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

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

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

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

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

□

**4.2. Proof of the Theorem 2.** Before we start proving this Theorem, we need some preparations. First, 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  $\kappa$  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]. □

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

$$(34) \quad \sum_{i>k} \|\mathbf{V}_{(i)*}\|^2 \leq \frac{q}{2-q} k(s/k)^{2/q}.$$

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

$$(35) \quad \sum_{i>k} \|\mathbf{V}_{(i)*}\|^2 \leq \frac{q}{2-q} \epsilon_n^2.$$

PROOF. This is a direct corollary of the Lemma 7 in Cai et al. [2013]. □

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  $\Sigma_S$ . 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 20), 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.  $\square$

Second, we need to introduce some notations and the ‘Oracle estimate’  $\widehat{\mathbf{V}}_O$ . Since we have randomly divided the samples into two equal sets of samples, we have the corresponding decomposition (33)

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

for these two sets of samples. More precisely, for  $i = 1, 2$ , we can define  $\mathbf{\Lambda}_H^{(i)}$ ,  $\mathbf{\Lambda}_u^{(i)}$ ,  $\mathcal{Z}^{(i)}$ ,  $\mathcal{W}^{(i)}$  and  $\mathcal{E}^{(i)}$  for the first and second set of samples respectively according to the decomposition (33). Let  $\mathbf{\Lambda} = \mathbf{V}\mathbf{D}\mathbf{V}^\tau$  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 8. Let  $J_S \mathbf{\Lambda} J_S = \mathbf{V}_S \mathbf{D}_S \mathbf{V}_S^\tau$  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  $\epsilon_n^2$  is sufficiently small. For  $i = 1, 2$ , let  $\mathbf{E}_3^{(i)}$  be the event defined similarly as  $\mathbf{E}_3$  (which is introduced near Corollary 1). Conditioning on  $\mathbf{E} = \mathbf{E}_3^{(1)} \cap \mathbf{E}_3^{(2)}$ , Lemma 2 implies that

$$(36) \quad (1 - \frac{\kappa}{2\nu})\lambda \leq \lambda_d^{(i)} \leq \dots \leq \lambda_1^{(i)} \leq (1 + \frac{1}{2\nu})\kappa\lambda \quad \text{for } i = 1, 2.$$

We first prove the following sparse approximation lemma.

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

$$(37) \quad \|\mathbf{V}_S^{(i)} \mathbf{V}_S^{(i),\tau} - \mathbf{V}_S \mathbf{V}_S^\tau\|_F^2 \leq C \frac{q}{2-q} \epsilon_n^2, \quad \|J_S \mathbf{\Lambda}_u^{(i)} J_S - \mathbf{\Lambda}_u^{(i)}\|_F \leq C \lambda \epsilon_n$$

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  $\mathbf{V}^{(i)} = \mathbf{V}\tilde{\mathbf{U}}$  for some (stochastic) orthogonal matrix  $\tilde{\mathbf{U}}$  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  $\mathbb{E}$ , Lemma 7, Lemma 8 and (36) imply

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

Since we have assumed that  $\epsilon_n^2$  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 20), we have

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

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

$$(38) \quad \begin{aligned} \widehat{\mathbf{V}}_O &\triangleq \arg \max_{\mathbf{V}} \langle \Lambda_H^{(1)}, \mathbf{V}\mathbf{V}^\tau \rangle = \arg \max_{\mathbf{V}} \text{Tr}(\mathbf{V}^\tau \Lambda_H^{(1)} \mathbf{V}) \\ &\text{s.t. } \mathbf{V}^\tau \mathbf{V} = I_d \text{ and } \text{supp}(\mathbf{V}) = S. \end{aligned}$$

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  $\Delta$  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  $\mathbb{E}$ , we know

$$(39) \quad \|\widehat{\mathbf{V}}_E \widehat{\mathbf{V}}_E^\tau - \widehat{\mathbf{V}}_O \widehat{\mathbf{V}}_O^\tau\|_F^2 \leq \frac{2}{\lambda_d(\mathbf{D}_S^{(2)})} \langle \widehat{\mathbf{V}}_O U_1 M U_1^\tau \widehat{\mathbf{V}}_O^\tau, \widehat{\mathbf{V}}_O \widehat{\mathbf{V}}_O^\tau - \widehat{\mathbf{V}}_E \widehat{\mathbf{V}}_E^\tau \rangle$$

$$(40) \quad \begin{aligned} &\leq \frac{C}{\lambda} \langle \widehat{\mathbf{V}}_O U_1 M U_1^\tau \widehat{\mathbf{V}}_O^\tau - \Lambda_H^{(2)}, \widehat{\mathbf{V}}_O \widehat{\mathbf{V}}_O^\tau - \widehat{\mathbf{V}}_E \widehat{\mathbf{V}}_E^\tau \rangle \\ &\triangleq I + II + III. \end{aligned}$$

where

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

Inequality (39) follows from applying the Lemma 6 with the positive definite matrix  $U_1 M U_1^\tau$ . The inequality (40) 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 5 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  $\mathbb{E}$ , we have

$$\begin{aligned} (41) \quad |I| &\leq C \|\widehat{\mathbf{V}}_O \widehat{\mathbf{V}}_O^\tau - \mathbf{V}_S^{(2)} \mathbf{V}_S^{(2),\tau}\|_F \|\widehat{\mathbf{V}}_O \widehat{\mathbf{V}}_O^\tau - \widehat{\mathbf{V}}_E \widehat{\mathbf{V}}_E^\tau\|_F \\ &\leq C(\delta + \theta^{(1)} + \theta^{(2)}) \|\widehat{\mathbf{V}}_O \widehat{\mathbf{V}}_O^\tau - \widehat{\mathbf{V}}_E \widehat{\mathbf{V}}_E^\tau\|_F. \end{aligned}$$

*For II:* It is nonzero only if  $q \neq 0$ . From Lemma 7 and Lemma 8, we know that

$$(42) \quad |II| \leq C \sqrt{\frac{q}{2-q}} \epsilon_n \|\widehat{\mathbf{V}}_O \widehat{\mathbf{V}}_O^\tau - \widehat{\mathbf{V}}_E \widehat{\mathbf{V}}_E^\tau\|_F.$$

*For III:* From the equation (33), we have

$$(43) \quad |III| \leq \frac{1}{\lambda} \|\widehat{\mathbf{V}}_O \widehat{\mathbf{V}}_O^\tau - \widehat{\mathbf{V}}_E \widehat{\mathbf{V}}_E^\tau\|_F (2T_2 + T_1)$$

where  $T_1 = \max_{B \in \mathcal{B}(k)} |\langle \mathcal{W}^{(2)} \mathcal{W}^{(2),\tau}, \mathbf{K}_B \rangle|$ ,  $T_2 = \max_{B \in \mathcal{B}(k)} |\langle \mathcal{Z}^{(2)} \mathcal{W}^{(2),\tau}, \mathbf{K}_B \rangle|$  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  $E$ , we have

$$(44) \quad \|\widehat{\mathbf{V}}_E \widehat{\mathbf{V}}_E^\tau - \widehat{\mathbf{V}}_O \widehat{\mathbf{V}}_O^\tau\|_F \leq C \left( \delta + \theta^{(1)} + \theta^{(2)} + \epsilon_n + \frac{1}{\lambda} (2T_2 + T_1) \right).$$

Thus, we have

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

If we can prove

$$(45) \quad \mathbb{E} (\theta^{(i)})^2 \mathbf{1}_{E'} \leq C \epsilon_n^2, \quad \mathbb{E} \delta^2 \mathbf{1}_{E'} \leq C \epsilon_n^2 \quad \text{and} \quad \mathbb{E} (2T_1 + T_2)^2 \mathbf{1}_{E'} \leq \lambda^2 \epsilon_n^2,$$

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

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

□

All we need to prove are the following two Lemmas.

LEMMA 10. Assume that  $n\lambda \leq e^p$ . There exist  $E' \subset E$  such that  $\mathbb{P}((E')^c) \leq C \frac{H^2}{n\lambda}$  and

$$(46) \quad \mathbb{E} (\theta^{(i)})^2 \mathbf{1}_{E'} \leq C \epsilon_n^2 \quad \text{and} \quad \mathbb{E} \delta^2 \mathbf{1}_{E'} \leq C \epsilon_n^2.$$

PROOF. Since  $\mathbf{V}^{(i)}$  and  $\mathbf{V}$  share the same column space, conditioning on  $E$ , by Lemma 8 and Lemma 7, we have

$$(47) \quad \theta^{(i)} = \|\mathbf{V}_S^{(i)} \mathbf{V}_S^{(i),\tau} - \mathbf{V} \mathbf{V}^\tau\|_F \leq 4\kappa \|J_S \mathbf{V}^{(i)} - \mathbf{V}^{(i)}\|_F \leq C \sqrt{\frac{q}{2-q}} \epsilon_n,$$

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

Let  $\mathbf{Q}_S = J_S (\Lambda_H^{(1)} - \Lambda_u^{(1)}) J_S$ . Let  $F$  consist of the events such that  $\|J_S \mathbf{W}^{(1)} \mathbf{W}^{(1),\tau} J_S\| \leq C \frac{k}{n}$ . Lemma 19 implies that  $\mathbb{P}(F^c) \leq C \frac{H^2}{n\lambda}$ . Since

we have assumed that  $\epsilon_n^2$  is sufficiently small, conditioning on  $E \cap F$ , the decomposition (33) give us

$$(48) \quad \|Q_S\|_2 \leq C\sqrt{\lambda} \sqrt{\frac{k}{n}} \leq C \frac{\lambda \epsilon_n}{\sqrt{d}}.$$

From (37), we also have  $\|J_S \Lambda_u^{(1)} J_S - \Lambda_u^{(1)}\|_F \leq C\lambda \epsilon_n \leq \frac{\lambda}{8}$ . Thus,  $\|J_S \Lambda_H^{(1)} J_S - \Lambda_u^{(1)}\|_F < \frac{\lambda}{4}$ , which implies the  $(d+1)$ -th largest eigenvalues of  $J_S \Lambda_H^{(1)} J_S$  is less than  $\frac{\lambda}{4}$ . Note that the eigenvalues of  $J_S \Lambda_u^{(1)} J_S \in (\frac{1}{2}\lambda, 2\kappa\lambda)$ . After applying the Sin-Theta Theorem( Lemma 20) to the pair of symmetric matrices  $(J_S \Lambda_u^{(1)} J_S, J_S \Lambda_H^{(1)} J_S)$ , we have

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

where the last inequality follows from (48). Thus, we may take  $E' = E \cap F$ .

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 (33).) and for each fixed  $B \in \mathcal{B}_k$ ,  $\mathbf{K}_B \perp \mathcal{W}^{(2)}$ , hence

$$(49) \quad \langle \mathcal{W}^{(2)} \mathcal{W}^{(2),\tau}, \mathbf{K}_B \rangle = \langle \mathcal{E}^{(2)} \mathcal{E}^{(2),\tau}, \mathbf{V}^{\perp,\tau} \mathbf{K}_B \mathbf{V}^\perp \rangle$$

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 21, we have

$$(50) \quad \mathbb{P} \left( \sqrt{n} |\langle \mathcal{E}^{(2)} \mathcal{E}^{(2),\tau}, \mathbf{V}^{\perp,\tau} \mathbf{K}_B \mathbf{V}^\perp \rangle| \geq 2 \frac{\sqrt{H}}{\sqrt{n}} t + \frac{2}{\sqrt{n}} t^2 \right) \leq 2 \exp(-t^2).$$

After applying Lemma 22 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  $\mathcal{Z}^{(2)} \perp\!\!\!\perp \mathcal{W}^{(2)}$ ,  $\mathbf{K}_B \perp\!\!\!\perp \mathcal{W}^{(2)}$  and  $\mathbf{K}_B \perp\!\!\!\perp \mathcal{Z}^{(2)}$ , 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  $N(0, \|\mathbf{V}^{\perp,\tau} \mathbf{K}_B \mathcal{Z}\|_F^2)$ . 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

$$(51) \quad \mathbb{P}(|F_B| > t) \leq \mathbb{P}(2\kappa\lambda|W| > t) \leq 2 \exp\left(-\frac{t^2}{4\kappa^2\lambda^2}\right).$$

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  $\mathbb{E}(T_2^2 \mathbf{1}_{\mathbf{E}}) \leq C\lambda^2 \epsilon_n^2$ .

□

**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  $\lambda$  are bounded. Here  $\lambda$ , 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  $\lambda$ , we will attempt to show that this convergence rate is optimal even for diverging  $d$  and  $\lambda$  in a future research.

The aggregate 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  $\boldsymbol{\Sigma} = \mathbf{I}$  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  $\beta_i'$ s are sparse, our argument in this paper implies that  $\mathbb{E}[\|\widehat{P}_{\text{col}(\text{var}(\mathbb{E}[\mathbf{x}|y]))} - P_{\text{col}(\text{var}(\mathbb{E}[\mathbf{x}|y]))}\|_F^2]$  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  $y = \boldsymbol{\beta}^\top \mathbf{x} + \epsilon$  with  $\mathbf{x} \sim N(0, \mathbf{I})$  and  $s = O(p^{1-\delta})$ , the minimax rate [Raskutti et al., 2011] of estimating  $\boldsymbol{\beta}$  is achieved by the simple correlation screening.

On the other hand, the minimax rate for estimating  $P_{\beta}$  is achieved by the DT-SIR algorithm of Lin et al. [2015], 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  $\Sigma$  might be possible, since it is now well known that Lasso[Tibshirani, 1996] and Dantzig Selector[Candes and Tao, 2007] achieve the optimal rate of linear regression [Bickel et al., 2009] for general  $\Sigma$ . This speculation warrants further future investigations.

## SUPPLEMENTARY MATERIAL

**Supplement to “ On the optimality of SIR in high dimensions”**  
[\(<http://www.e-publications.org/ims/support/download/imsart-ims.zip>\).](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 Debasish 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.
- E. Candes and T. Tao. The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, pages 2313–2351, 2007.
- Chun-Hou 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, 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.

- 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.
- N. Duan and K. C. 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.
- T. Hsing and R. 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.
- L. Li. Sparse sufficient dimension reduction. *Biometrika*, 94(3):603–613, 2007.
- L. Li and C. J. Nachtsheim. Sparse sliced inverse regression. *Technometrics*, 48(4), 2006.
- Qian Lin, Zhigen Zhao, and Jun S. Liu. On the consistency and sparsity of high dimensional sliced inverse regression. *manuscript*, 2015.
- Qian Lin, Xinran Li, Dongming Huang, and S. Jun Liu. Supplementary to 'on optimality of sliced inverse regression in high dimensions'. 2016a.
- Qian Lin, Zhigen Zhao, and S. Jun Liu. Sparse sliced inverse regression for high dimensional data'. 2016b.
- Matey Neykov, Qian Lin, and Jun S. Liu. On support recover of sliced inverse regression in high dimensional. *manuscript*, 2015.
- 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.
- R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B*, 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.
- Lixing Zhu, Baiqi Miao, and Heng Peng. On sliced inverse regression with high-dimensional covariates. *Journal of the American Statistical Association*, 101(474), 2006.
- H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, 67(2):301–320, 2005.

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

BY QIAN LIN<sup>†</sup> XINRAN LI<sup>‡</sup> DONGMING HUANG<sup>‡</sup> AND JUN S. LIU<sup>‡</sup>

*Harvard University<sup>‡</sup>*

**Proof of Theorem 9.** 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  $\gamma$ , let  $\tilde{\gamma} = \gamma/\|\gamma\|_2$ . For any non-zero vector  $\gamma$  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

$$(52) \quad \|\gamma - \tilde{\gamma}_{T_t}\|_2^2 \leq Cst^2.$$

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  $\Lambda = \lambda \beta \beta^\top$ . For  $S \subset [p]$ , we have  $\lambda(\Lambda(S, S)) = \lambda \|\beta_S\|^2 \tilde{\beta}_S \tilde{\beta}_S^\top$ .

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

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

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  $u \in \mathbb{G}(p, d)$ , Cai et al. [2013] have constructed a subset  $\Theta \subset N(u, 2\varepsilon)$ , an  $2\varepsilon$  neighbourhood of  $u$  in  $\mathbb{G}(p, d)$ , such that, for any  $\alpha \in (0, 1)$ , we have

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

where  $u_i$  and  $u_j$  are two different points  $\in \Theta$  and  $c_0$  and  $c_1$  are two absolute constants. Lemma 15 states that if  $\varepsilon$  is sufficiently small, then for each  $u_i \in \Theta \subset \mathbb{G}(p, d)$ , there is an  $u_i \in \mathbb{O}(p, d)$  such that  $a_i a_i^\tau = u_i$

$$(53) \quad C_1 \|u_i - u_j\|_F \leq \|a_i - a_j\|_F \leq C_2 \|u_i - u_j\|_F$$

for some absolute positive constants  $C_1$  and  $C_2$ . Let us denote  $\tilde{\Theta} = \{a_i\}$  and consider the following models

$$(54) \quad y = f(\mathbf{V}^\tau \mathbf{x}) + \epsilon, \quad \mathbf{V} \in \tilde{\Theta}, \quad \mathbf{x} \sim N(0, \mathbf{I}_p), \quad \epsilon \sim N(0, 1).$$

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  $p_{\mathbf{B}, g}(y, \mathbf{x})$  be the joint density function of  $(y, \mathbf{x})$ , then we have

$$(55) \quad KL(p_{\mathbf{B}, g}, p_{\mathbf{B}', g}) \leq |\nabla g|^2 \|B - B'\|_F^2.$$

If  $f$  satisfying the Conjecture 1, the Fano Lemma gives us

$$(56) \quad \begin{aligned} & \sup_{\mathbf{V} \in \tilde{\Theta}} \mathbb{E} \|P_{\hat{\mathbf{V}}} - P_{\mathbf{V}}\|_F^2 \\ & \geq \min \|P_{\mathbf{V}_i} - P_{\mathbf{V}_j}\|_F^2 \left( 1 - \frac{\max KL(p_{\mathbf{V}_i, f}^n, p_{\mathbf{V}_j, f}^n) + \log(2)}{\log(|\Theta|)} \right) \\ & \geq \varepsilon^2 \left( 1 - \frac{n\lambda C_2^2 \varepsilon^2 + 4 \log 2}{\log(|\Theta|)} \right). \end{aligned}$$

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

$$(57) \quad \sup_{u \in \Theta} \mathbb{E} \|P_{\hat{\mathbf{V}}} - P_{\mathbf{V}}\|_F^2 \succ \frac{d(p-d)}{n\lambda}$$

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

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

$$(58) \quad \inf_{\widehat{\mathbf{V}}} \sup_{\mathcal{M} \in \mathfrak{M}_{s,0}(p,d,\lambda,\kappa)} \mathbb{E}_{\mathcal{M}} \|\widehat{\mathbf{V}}\widehat{\mathbf{V}}^T - \mathbf{V}\mathbf{V}^T\|_F^2 \succ d \wedge \frac{s \log \frac{ep}{s}}{n\lambda}.$$

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

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

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 (56) gives us the (58).

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

**A linear algebraic lemma.** In this section, we include a differential geometric argument for the following linear algebraic lemma.

LEMMA 15. *Let  $\pi : A \mapsto AA^\tau$  be the tautological map between the set  $\mathbb{O}(p, d) \subset \mathbb{R}^{p \times d}$  of  $p \times d$  orthogonal matrices and the set  $\mathbb{G}(p, d) \subset \mathbb{R}^{p \times p}$  of all  $d$ -dimensional subspaces in  $\mathbb{R}^p$ .<sup>1</sup> There exists an open set  $U \subset \mathbb{G}(p, d)$  and positive constant  $C$ , such that for any  $n$  and  $u_1, \dots, u_n \in U$ , there exist orthogonal  $p \times d$  matrices  $A_1, \dots, A_n$  such that for any  $i, j$  one has  $A_j A_i^\tau = u_j$  and*

$$(59) \quad \|A_i - A_j\|_F \leq C \|A_i A_i^\tau - A_j A_j^\tau\|_F$$

PROOF. Since we did not find an algebraic proof of this lemma, we resort it to the following geometric lemma. the proof of which is rather involved.

$\square$

---

<sup>1</sup> $\mathbb{G}(p, d)$  is the so-called Grassmannian. Each point in  $\mathbb{G}(p, d)$  can be identified with a projection matrix, thus we can embed it into  $\mathbb{R}^{p^2}$ .

LEMMA 16. *The standard metric on Euclidean space induced the Frobenius distance on  $\mathbb{O}(p, d)$  by and the metric on  $\mathbb{G}(p, d)$ , i.e., for  $A, B \in \mathbb{O}(p, d)$ , one has  $d(A, B) = \|A - B\|_F$  and for  $AA^\tau, BB^\tau \in \mathbb{G}(p, d)$ , one has  $D(AA^\tau, BB^\tau) = \|AA^\tau, BB^\tau\|_F$ . There exist an open set  $U$  and a section  $\sigma : U \mapsto \pi^{-1}(U)$ , i.e.,  $(\pi \circ \sigma = Id_U)$  in the following commutative diagram,*

$$(60) \quad \begin{array}{ccc} \mathbb{O}(p, d) & \xhookrightarrow{\quad} & \pi^{-1}(U) \\ \downarrow \pi & & \downarrow \pi \\ \mathbb{G}(p, d) & \xhookrightarrow{\quad} & U \end{array} \quad = \quad \begin{array}{c} \nearrow \sigma \\ U \end{array}$$

satisfying that, there exist positive constants  $C_1$  and  $C_2$  such that, for any  $A, B \in \sigma(U)$ ,

$$(61) \quad C_1 D(\pi(A), \pi(B)) \leq d(A, B) \leq C_2 D(\pi(A), \pi(B)).$$

PROOF. <sup>2</sup> Note that for any submanifold  $M \subset \mathbb{R}^n$ , there are two distances: the induced distance  $d_M^i$  which induced from the Euclidean distance and the geodesic distance  $d_M^g$  where  $M$  with the induced Riemannian metric.

LEMMA 17. *We embed  $\mathbb{O}(p, d)$  into the set of all  $p \times d$  matrices and  $\mathbb{G}(p, d)$  into the set of all  $p \times p$  matrices. There exist constants  $C_1, C_2$  such that, for any point  $A \in \mathbb{O}(p, d)$ , there exists an open neighbourhood  $U$  of  $A$  such that for any  $u_1, u_2 \in U$ , one has*

$$(62) \quad C_1 d_{\mathbb{O}(p, d)}^i(u_1, u_2) \leq d_{\mathbb{O}(p, d)}^g(u_1, u_2) \leq C_2 d_{\mathbb{O}(p, d)}^i(u_1, u_2).$$

Similarly, there exist constants  $C_1, C_2$  such that, for any point  $S_A \in \mathbb{G}(p, d)$ , there exists an open neighbourhood  $U$  of  $S_A$  such that for any  $u_1, u_2 \in U$ , one has

$$(63) \quad C_1 d_{\mathbb{G}(p, d)}^i(u_1, u_2) \leq d_{\mathbb{G}(p, d)}^g(u_1, u_2) \leq C_2 d_{\mathbb{G}(p, d)}^i(u_1, u_2).$$

PROOF. We only prove the inequality (62). The first part is trivial. Thus we only need to prove the second inequality. Let  $J = (I_d, 0_{d \times (p-d)})^\tau$ . Since  $\mathbb{O}(p, d)$  is homogeneous manifold, we only need to prove that there exists a neighbourhood  $U$  of  $J$  such that for any  $u \in U$ , the (62) holds for  $J$  and  $u$ . Note that any orthogonal matrix  $u$  near  $J$  could be written as

$$(64) \quad u = \begin{pmatrix} I_d \\ B \end{pmatrix} (1 + B^\tau B)^{-1/2} T$$

---

<sup>2</sup>To avoid reproducing standard content in textbook, in this proof, we assume the reader has some familiarities with differential geometry .

where  $B$  is some  $(p-d) \times d$  matrix and  $T$  is an  $d \times d$  orthogonal matrix.

Let us consider the curve

$$(65) \quad u(t) = \begin{pmatrix} I_d \\ tB \end{pmatrix} (1 + t^2 B^\tau B)^{-1/2} T$$

inside  $\mathbb{O}(p, d)$  and denote its length by  $s$ . Since  $d_{\mathbb{O}(p, d)}^g(J, u) \leq s$ , we only need to prove that there exists constant  $C$  such that  $s \leq C \|J - u\|_F$ . Note that

$$(66) \quad \frac{du}{dt} = \begin{pmatrix} 0 \\ B \end{pmatrix} (1 + t^2 B^\tau B)^{-1/2} T - t \begin{pmatrix} I_d \\ tB \end{pmatrix} (1 + t^2 B^\tau B)^{-3/2} B^\tau B T$$

Let  $B^\tau B = VE^2V^\tau$  where  $E^2$  is diagonal matrix with entries  $e_1^2, \dots, e_d^2$ . Then one has

$$(67) \quad \left\| \frac{du}{dt} \right\|^2 = \text{tr}(E^2(1 + t^2 E^2)^{-2})$$

For any orthogonal matrix  $T$  and semi-positive definite matrix  $A$ , one has  $\text{tr}(TA) \leq \text{tr}(A)$ . Thus we have

$$\begin{aligned} \|J - u(t)\|_F^2 &= 2d - 2\text{tr}\left((1 + t^2 B^\tau B)^{-1/2} T\right) \geq 2d - 2\text{tr}\left((1 + t^2 B^\tau B)^{-1/2}\right) \\ &= 2d - 2\text{tr}\left((1 + t^2 E^2)^{-1/2}\right). \end{aligned}$$

Since

$$(68) \quad s^2 = \left( \int_0^t \left\| \frac{du}{da} \right\| da \right)^2 \leq t \int_0^t \left\| \frac{du}{da} \right\|^2 da \leq t \int_0^t \text{tr}(E^2(1 + t^2 E^2)^{-2}) da,$$

we only need to prove that there exists positive constant  $C$ , such that for any  $1 \leq j \leq d$ , one has

$$(69) \quad t \int_0^t \frac{e_j^2}{(1 + a^2 e_j^2)^2} da \leq C \left( 1 - \frac{1}{\sqrt{1 + t^2 e_j^2}} \right)$$

which can be verified directly when both  $e_j$ 's and  $t$  are sufficiently small.  $\square$

Note that  $\mathbb{O}(p, d)$  is a principal bundle over  $\mathbb{G}(p, d)$  with structure group  $\mathbb{O}(d, d)$ . Thus for any point  $A \in \mathbb{O}(p, d)$ , we have a natural decomposition of tangent space at  $A$ :

$$(70) \quad T_A \mathbb{O}(p, d) = V_A \oplus H_A$$

where  $V_A = \{ AX \mid X \text{ is } d \times d \text{ anti-symmetric matrix.} \}$  consists of vector tangent to the fibre and  $H_A = \{ B \mid B \text{ is } p \times d \text{ matrix such that } B^\tau A = 0. \}$  consists of vector perpendicular to  $V_A$ . For any non-zero  $\alpha \in T_A \mathbb{O}(p, d)$ , let  $\alpha = \alpha_V \oplus \alpha_H$  be the decomposition with respect to (70). We introduce

$$(71) \quad \mu(\alpha) \triangleq \frac{\|\alpha_H\|}{\|\alpha\|}.$$

For a subspace  $V \subset T_A \mathbb{O}(p, d)$ , we define  $\mu(V) = \inf_{\alpha \in V \setminus 0} \mu(\alpha)$ .

LEMMA 18. *For any  $A \in \mathbb{O}(p, d)$ , there exists a neighbourhood  $U$  of  $\pi(A) \in \mathbb{G}(p, d)$  such that there exists a smooth map  $\sigma : U \mapsto \pi^{-1}(U)$  satisfying  $\sigma(\pi(A)) = A$ ,  $d\sigma|_{\pi(A)}(T_{\pi(A)} \mathbb{G}(p, d)) = H_A$  and*

$$(72) \quad \forall u \in U, \mu(d\sigma|_u T_u \mathbb{G}(p, d)) \geq \frac{1}{2}.$$

As a direct corollary, we know that there exists two positive constant  $C_1, C_2$  such that for any  $u_1, u_2 \in U$ , one has

$$(73) \quad C_1 d_{\mathbb{G}(p, d)}^g(u_1, u_2) \leq d_{\mathbb{O}(p, d)}^g(\sigma(u_1), \sigma(u_2)) \leq C_2 d_{\mathbb{G}(p, d)}^g(u_1, u_2).$$

PROOF. The existence of  $U$  and the inequality (72) follows from a typical continuity argument. For the second inequality in (73), let  $u(t)$  be a geodesic connect  $u_1$  and  $u_2$ . When  $U$  is sufficient small,  $u(t)$  is unique and and

$$(74) \quad d_{\mathbb{G}(p, d)}^g(u_1, u_2) = \int_0^1 \left\| \frac{du(t)}{dt} \right\|_{\mathbb{G}} dt.$$

Let  $L$  be the length of the curve  $\sigma(u(t))$ , then

$$(75) \quad \begin{aligned} L &= \int_0^1 \left\| \frac{d\sigma(u(t))}{dt} \right\|_{\mathbb{O}} dt = \int_0^1 \left\| d\sigma \left( \frac{du(t)}{dt} \right) \right\|_{\mathbb{O}} dt \\ &\leq 2 \int_0^1 \left\| d\sigma \left( \frac{du(t)}{dt} \right) \right\|_H dt \end{aligned}$$

$$(76) \quad \begin{aligned} &= 2\sqrt{2} \int_0^1 \left\| \frac{du(t)}{dt} \right\|_{\mathbb{G}} dt \\ &= 2\sqrt{2} d_{\mathbb{G}(p, d)}^g(u_1, u_2) \end{aligned}$$

where inequality (75) follows from the inequality (72) and equation (73) follows from the fact that for any  $B \in T_B \mathbb{O}(p, d)$ , ( i.e.,  $B^\tau A + A^\tau B = 0$ ),

one has  $d\pi(B) = AB^\tau + BA^\tau \in T_{AA^\tau}\mathbb{G}(p, d)$  and if  $B \in H_A$ , (i.e.,  $B^\tau A = 0$ ), one has  $\text{tr}d\pi(B)d\pi(B)^\tau = 2\text{tr}BB^\tau$ . In particular, we know that

$$(77) \quad d_{\mathbb{O}(p,d)}^g(\sigma(u_1), \sigma(u_2)) \leq L \leq 2\sqrt{2}d_{\mathbb{G}(p,d)}^g(u_1, u_2)$$

The first inequality in (73) can be proved similar and thus omitted.  $\square$

The Lemma 16 is a direct corollary of the above two Lemmas.  $\square$

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

LEMMA 19. 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

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

COROLLARY 2. One has

$$\frac{1}{2} \left( \sqrt{p} - \sqrt{H} \right) \leq \lambda_{\text{sing},\min}^-(\mathbf{E}_{p \times H}) \leq \lambda_{\text{sing},\max}(\mathbf{E}_{p \times H}) \leq \frac{3}{2} \left( \sqrt{p} + \sqrt{H} \right).$$

with probability converging to one, as  $n \rightarrow \infty$ .

LEMMA 20. (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 21. 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

$$(78) \quad \mathbb{P} \left( \left| \langle \mathbf{Z}^\tau \mathbf{Z}, \mathbf{K} \rangle \right| \geq 2\sqrt{Ht} + 2t^2 \right) \leq 2 \exp(-t^2).$$

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 22. Let  $X_1, \dots, X_N$  be i.i.d such that

$$(79) \quad \mathbb{P}(|X_1| \geq at + bt^2) \leq c \exp(-t^2)$$

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

$$(80) \quad \mathbb{E} \max |X_i|^2 \leq (2a^2 + 8b^2) \log(ecN) + 2b^2 \log^2(cN).$$