Submitted to the Annals of Statistics
arXiv: arXiv:0000.0000

ON CONSISTENCY AND SPARSITY FOR SLICED INVERSE REGRESSION IN HIGH DIMENSIONS

By Qian Lin[†*] Zhigen Zhao[‡*] and Jun S. Liu[†*]

Harvard University[†]
Temple University[‡]

We provide here a framework to analyze the phase transition phenomenon of slice inverse regression (SIR), a supervised dimension reduction technique introduced by Li [1991]. Under mild conditions, the asymptotic ratio $\rho = \lim p/n$ is the phase transition parameter and the SIR estimator is consistent if and only if $\rho = 0$. When dimension p is greater than n, we propose a diagonal thresholding screening SIR (DT-SIR) algorithm. This method provides us with an estimate of the eigen-space of the covariance matrix of the conditional expectation $var(\mathbb{E}[\boldsymbol{x}|y])$. The desired dimension reduction space is then obtained by multiplying the inverse of the covariance matrix on the eigen-space. Under certain sparsity assumptions on both the covariance matrix of predictors and the loadings of the directions, we prove the consistency of DT-SIR in estimating the dimension reduction space in high dimensional data analysis. Extensive numerical experiments demonstrate superior performances of the proposed method in comparison to its competitors.

1. Introduction. For a continuous multivariate random variable $(y, \boldsymbol{x})$ where $\boldsymbol{x} \in \mathbb{R}^p$ and $y \in \mathbb{R}$, a subspace $\mathcal{S}' \subset \mathbb{R}^p$ is called the effective dimension reduction (EDR) space if $y \perp\!\!\!\perp \boldsymbol{x}|P_{\mathcal{S}'}(\boldsymbol{x})$ where $\perp\!\!\!\perp$ stands for independence. Under mild conditions (Cook [1996]), the intersection of all the EDR spaces is again an EDR space, which is denoted as $\mathcal{S}$ and called the central space. Many algorithms were proposed to find such subspace $\mathcal{S}$ under the assumption $d = dim\mathcal{S} \ll p$. This line of research is commonly known as sufficient dimension reduction. The Sliced Inverse Regression (SIR, Li [1991]) is the first, yet the most widely used method in sufficient dimension reduction, due to its simplicity, computational efficiency and generality. The asymptotic properties of SIR are of particular interest in the last two decades. The consistency of SIR has been proved for fixed p in Li [1991], Hsing and Carroll

*Lin's research is supported by the Center of Mathematical Sciences and Applications at Harvard University. Zhao's research is supported by the NSF Grant DMS-1208735. Liu's research is supported by the NSF Grant DMS-1120368 and NIH Grant R01 GM113242-01

MSC 2010 subject classifications: Primary 62J02; secondary 62H25

Keywords and phrases: dimension reduction, random matrix theory, sliced inverse regression

[1992], Zhu and Ng [1995] and Zhu and Fang [1996]. Later, Zhu et al. [2006] have obtained the consistency if $p = o(\sqrt{n})$. A similar restriction also appears in two recent work (see Zhong et al. [2012] and Jiang and Liu [2014]). When $p > n$, a common strategy pursued by many recent researchers is to make sparsity assumptions that only a few predictors play a role in explaining and predicting y and apply various regularization methods. For instance, Li and Nachtsheim [2006], Li [2007] and Yu et al. [2013] applied LASSO (Tibshirani [1996]), Dantzig selector (Candes and Tao [2007]) and elastic net (Zou and Hastie [2005]) respectively to solve the generalized eigenvalue problems raised by a variety of SDR algorithms.

However, a piece of jigsaw is missing in the understanding of SIR. If the dimension p diverges as n increases, when will the SIR break down? A similar question has been asked for a variety of SDR estimates in Cook et al. [2012]. In this paper, we prove that, under certain technical assumptions, the SIR estimator is consistent if and only if $\rho = \lim \frac{p}{n} = 0$. Such a result on inconsistency, on the other hand, provides theoretical justifications for imposing certain structural assumption, such as sparsity, in high dimensional setting. This behaviour of SIR in high dimension, which will be called the phase transition phenomenon, is similar to that of the principal component analysis (PCA), an unsupervised counterpart of SIR. This extension is, however, by no means trivial. After all the samples $(y_i, \boldsymbol{x}_i)$ are sliced into H bins according to the order statistics of y_i , the sliced samples are neither independent nor identically distributed. This difference increases the difficulty significantly. In this paper, we provide a new framework to study the phase transition behaviour of SIR. The technical tools developed here can potentially be extended to study the phase transition behaviour of other SDR estimators.

The second part of the article aims at extending the original SIR to the scenario with ultra-high dimension ($p = o(\exp(n^{\xi}))$). Based on equation (2), once we obtain consistent estimates $\widehat{\boldsymbol{\eta}}_i$'s of the top eigenvectors $\boldsymbol{\eta}_i$'s of $var\,(\mathbb{E}[\boldsymbol{x}|y])$, the central space can be estimated by multiplying any consistent estimate $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}$ of $\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}$, the precision matrix of $\boldsymbol{x}$, on the space spanned by the $\widehat{\boldsymbol{\eta}}_i$'s . In other words, we may focus our study on the estimation of the top eigen-space of $var\,(\mathbb{E}[\boldsymbol{x}|y])$. Appropriate sparsity assumptions on the $\boldsymbol{\beta}_i$'s and $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ guarantee us the sparsity of the $\boldsymbol{\eta}_i$'s. Motivated by recent work in sparse PCA (Johnstone and Lu [2004]), we propose a diagonal screening procedure based on new statistics $var_{H,c}(\boldsymbol{x}(k))$, which are the diagonal elements of $var\,(\mathbb{E}[\boldsymbol{x}|y])$. After ranking the predictors according to the magnitude of $var_{H,c}(\boldsymbol{x}(k))$ decreasingly, we choose the set $\mathcal{I}$ consisting of the first R predictors for further analysis. The SIR is further applied on these

predictors to estimate the top d eigenvectors $\boldsymbol{\eta}_1^{\mathcal{I}}, \cdots, \boldsymbol{\eta}_d^{\mathcal{I}}$ of $var(\mathbb{E}[\boldsymbol{x}^{\mathcal{I}}|y])$, denoted by $\widehat{\boldsymbol{\eta}}_1^{\mathcal{I}}, \cdots, \widehat{\boldsymbol{\eta}}_d^{\mathcal{I}}$. We embed these vectors into $\mathbb{R}^p$ by filling 0's for entries outside the chosen set $\mathcal{I}$, and denote these new vectors as $\widehat{\boldsymbol{\eta}}_1, \cdots, \widehat{\boldsymbol{\eta}}_d$. The final directions are spanned by $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}\widehat{\boldsymbol{\eta}}_1, \cdots \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}\widehat{\boldsymbol{\eta}}_d$. We call this two-stage algorithm as **D**iagonal **T**hresholding **SIR** (DT-SIR). We prove that DT-SIR is consistent in estimating the central space under certain regularity conditions. Extensive simulation studies show that DT-SIR performs significantly better than its competitors.

The rest of the paper is organized as follows. In Section 2, we briefly describe the SIR procedure and introduce the notations. In Section 3, after a brief review of existing asymptotic results of SIR procedure, we state Theorems 2 and 3 to discuss the phase transition phenomenon of SIR. In Section 4, we propose the DT-SIR method and show that DT-SIR is consistent in high dimensional data analysis. In Section 5, we provide simulation studies to compare DT-SIR with its competitors. Concluding remarks and discussions are put in Section 6. All the proofs are presented in appendices.

2. Preliminaries and notations.

Let us consider the multi-index model

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

where $x \in \mathbb{R}^p$, ϵ is the noise and f is an unknown link function. Without loss of generality, we assume that $\mathbb{E}[\boldsymbol{x}] = 0 \in \mathbb{R}^p$. Though the $p \times d$ matrix $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_d)$ is not identifiable, $< \boldsymbol{\beta} >$, the space spanned by the $\boldsymbol{\beta}_i$'s, might be identified. Li [1991] proposed the *Sliced Inverse Regression* (SIR) procedure to estimate the space $< \boldsymbol{\beta} >$ without knowing $f(\cdot)$. More precisely, under the following two conditions (**A1**)-(**A2**), SIR provides the estimate $\widehat{\boldsymbol{\Lambda}}_p$ and $\widehat{\boldsymbol{\eta}}_1, \cdots, \widehat{\boldsymbol{\eta}}_d$ of $\boldsymbol{\Lambda}_p = var\left(\mathbb{E}[\boldsymbol{x}|y]\right)$ and its top d eigenvectors $\boldsymbol{\eta}_1, \cdots, \boldsymbol{\eta}_d$ respectively.

- (**A1**). **Linearity condition:** For any $\boldsymbol{\xi} \in \mathbb{R}^p$, $\mathbb{E}[\boldsymbol{\xi}^\tau \boldsymbol{x}|\boldsymbol{\beta}_1^\tau \boldsymbol{x}, \cdots, \boldsymbol{\beta}_d^\tau \boldsymbol{x}]$ is a linear combination of $\boldsymbol{\beta}_1^\tau \boldsymbol{x}, \cdots, \boldsymbol{\beta}_d^\tau \boldsymbol{x}$.
- (**A2**). **Coverage condition:** The dimension of the space spanned by the central curve equals the dimension of the central space, i.e., $d' = d$.

We remind that condition (**A2**) is an implicit condition on f. e.g., If f is symmetric, then the condition (**A2**) fails. Based on the observation that the space spanned by the $\boldsymbol{\eta}_i$'s is the same as the space spanned by the $\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta}_i$'s, i.e.,

$$< \left\{ \boldsymbol{\eta}_1, \cdots, \boldsymbol{\eta}_{d'} \right\} >= \boldsymbol{\Sigma}_{\boldsymbol{x}} < \left\{ \boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_d \right\} >, \tag{2}$$

one knows that $< \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}\widehat{\boldsymbol{\eta}}_1, \cdots, \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}\widehat{\boldsymbol{\eta}}_d >$ is an estimate of $< \{\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_d\} >$. Throughout this paper, we assume that d is fixed and the d-th largest eigenvalue λ_d of $\boldsymbol{\Lambda}_p$ is bounded away from 0 when $n, p \to \infty$.

We adopt the following widely used notations (See e.g., Zhu and Ng [1995]). Given n i.i.d. samples $(y_i, \boldsymbol{x}_i)$, $i = 1, \cdots, n$, we divide them into H slices according to the order statistics $y_{(i)}$. To ease notations and arguments, we assume that $n = cH$ and $H = o(\log(n) \wedge \log(p))$ throughout this paper. Express the data as $y_{h,j}$ and $\boldsymbol{x}_{h,j}$ where (h, j) is the double subscript in which h refers to the slice number and j refers to the order number of an observation in the h-th slice. In other words,

$$y_{h,j} = y_{(c(h-1)+j)}, \qquad \boldsymbol{x}_{h,j} = \boldsymbol{x}_{(c(h-1)+j)}. \tag{3}$$

Here $\boldsymbol{x}_{(k)}$ is the concomitant of $y_{(k)}$. We denote the sample mean in the h-th slice by $\overline{\boldsymbol{x}}_{h,\cdot}$ and the mean of all the samples by $\overline{\overline{\boldsymbol{x}}}$. Then the SIR estimator $\widehat{\boldsymbol{\Lambda}}_p$ of the matrix $\boldsymbol{\Lambda}_p$ is

$$\widehat{\boldsymbol{\Lambda}}_p = \frac{1}{H-1}\sum_{h=1}^{H}(\bar{\boldsymbol{x}}_{h,\cdot} - \overline{\overline{\boldsymbol{x}}})(\bar{\boldsymbol{x}}_{h,\cdot} - \overline{\overline{\boldsymbol{x}}})^{\tau}. \tag{4}$$

Let S_h be the h-th interval $(y_{h-1,c}, y_{h,c}]$ for $2 \le h \le H-1$, $S_1 = (-\infty, y_{1,c}]$ and $S_H = (y_{H-1,c}, \infty)$. Note that these intervals depend on the order statistics $y_{(i)}$ and are thus random. For any ω in the product sample space, we define a random variable $\delta_h = \delta_h(\omega) = \int_{y \in S_h(\omega)} f(y)dy$ where $f(y)$ is the density function of y.

In addition, we adopt the following notations throughout this paper.

- For $\mathcal{I} \subset \{1, \cdots, n\}, \mathcal{J} \subset \{1, \cdots, p\}$ and a $n \times p$ matrix $\boldsymbol{A}$, $\boldsymbol{A}^{\mathcal{I},\mathcal{J}}$ denotes the $|\mathcal{I}| \times |\mathcal{J}|$ sub-matrix formed by restricting the rows of $\boldsymbol{A}$ to $\mathcal{I}$ and columns to $\mathcal{J}$. In articular, $\boldsymbol{A}^{-,\mathcal{J}}$ denotes the sub-matrix formed by restricting the columns to $\mathcal{J}$;
- For a matrix $\boldsymbol{B} = \boldsymbol{A}^{\mathcal{I},\mathcal{J}} \in \mathbb{R}^{|\mathcal{I}| \times |\mathcal{J}|}$, we embed it into $\mathbb{R}^{p \times p}$ by putting 0 on entries outside $\mathcal{I} \times \mathcal{J}$ and denote the new matrix as $e(\boldsymbol{B})$. Similar notations apply to vectors;
- For two positive numbers a and b, we use $a \vee b$ to denote $\max\{a, b\}$ and $a \wedge b$ to denote $\min\{a, b\}$;
- Let $\tau(x, t) = x1(|x| > t)$ be the hard thresholding function;
- Throughout the paper, C, C_1 and C_2 are used to denote generic absolute constants, though the actual value may vary from case to case;
- For a vector $\boldsymbol{x}$, we denote the k-th entry of $\boldsymbol{x}$ as $\boldsymbol{x}(k)$;
- Let $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ be two vectors with the same dimension, the angle between these two vectors is denoted as $\angle(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$.

- For two sequences a_n,b_n, we let $a_n \ll b_n$ stand for $a_n = O(b_n^{\epsilon})$ for some positive $\epsilon < 1$ and let $a_n \succ b_n$ stand for $\lim \frac{b_n}{a_n} = 0$.

3. Consistency of SIR. First, we impose the following condition on the covariance matrix.

- **(A3) Boundedness Condition:** There exist positive constants C_1, C_2 such that
$$C_1 \leq \lambda_{min}(\boldsymbol{\Sigma}_{\boldsymbol{x}}) \leq \lambda_{max}(\boldsymbol{\Sigma}_{\boldsymbol{x}}) \leq C_2$$
where $\lambda_{min}(\boldsymbol{\Sigma}_{\boldsymbol{x}})$ and $\lambda_{max}(\boldsymbol{\Sigma}_{\boldsymbol{x}})$ are the minimal and maximal eigenvalues of $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ respectively.

Second, we assume that the central curve satisfies the following condition:

- (**T1**) The central curve $\boldsymbol{m}(y) \triangleq \mathbb{E}[\boldsymbol{x}|y]$ has finite fourth moment and is κ-sliced stable (defined below) with respect to y and $\boldsymbol{m}(y)$

DEFINITION 1. *The central curve $\boldsymbol{m}(y) \in \mathbb{R}^p$ is called κ-sliced stable with respect to y for some $\kappa > 0$, if for given constants $\mathfrak{a}_1 < 1 < \mathfrak{a}_2$, there exists a positive constant $\mathfrak{a}$ such that for any unit vector $\boldsymbol{\beta}$ and for any partition*

$$-\infty < a_1 < a_2 < \cdots < a_{H-1} < \infty$$

of $\mathbb{R}$ satisfying $\frac{\mathfrak{a}_1}{H} \leq P(a_i \leq y < a_{i+1}) \leq \frac{\mathfrak{a}_2}{H}$, one has

$$\frac{1}{H}\Big|\sum_{h=0}^{H} var(\boldsymbol{\beta}^{\tau}\boldsymbol{m}(y)|a_h \leq y \leq a_{h+1})\Big| \leq \frac{\mathfrak{a}}{H^{\kappa}} var(\boldsymbol{\beta}^{\tau}\boldsymbol{m}(y)) \tag{5}$$

where we denote $a_0 = -\infty, a_H = \infty$. The central curve is sliced stable if it is κ-sliced stable for some positive constant κ.

REMARK 1. *Intuition behind the sliced stable condition.* Suppose there are n samples $\boldsymbol{m}_i \triangleq \boldsymbol{m}(y_i)$, let $\boldsymbol{m}_{h,i}$ and $\overline{\boldsymbol{m}}_{h,\cdot}$ be defined similar to $\boldsymbol{x}_{h,i}$ and $\overline{\boldsymbol{x}}_{h,\cdot}$ respectively. On the one hand, one has the classical consistent estimate $\frac{1}{n}\sum_i \boldsymbol{m}_i\boldsymbol{m}_i^{\tau}$ of $var(\boldsymbol{m}(y))$. On the other hand , if one expects that the slice estimate $\frac{1}{H}\sum_h \overline{\boldsymbol{m}}_{h,\cdot}\overline{\boldsymbol{m}}_{h,\cdot}^{\tau}$ of $var(\boldsymbol{m}(y))$ is consistent, one must require that the average loss of variance in each slice (i.e., $\frac{1}{c}\sum \boldsymbol{m}_{h,i}\boldsymbol{m}_{h,i}^{\tau} - \overline{\boldsymbol{m}}_{h,\cdot}\overline{\boldsymbol{m}}_{h,\cdot}^{\tau}$) to be decreasing as H is increasing. In Definition 1, we simply choose the decreasing rate to be a power of H.

Note that the sliced stable condition could be viewed as a property of the pair of the function $\widetilde{\boldsymbol{m}}(y) = \boldsymbol{\beta}^{\tau}\boldsymbol{m}(y)$ and the random variable y.

i) If y is exponential or Gaussian, then y is sliced stable. Numerical experiments also show that Pareto distribution is sliced stable if its 4-th moment exists.

ii) If y is sliced stable and the function $\widetilde{\boldsymbol{m}}$ has a bounded first derivative, then $\widetilde{\boldsymbol{m}}(y)$ is sliced stable.

iii) If y is bounded and $\widetilde{\boldsymbol{m}}$ is Hölder continuous, then $\widetilde{\boldsymbol{m}}(y)$ is sliced stable.

From the above list, we know that the sliced stable condition holds for a large class of functions and random variables. We would like to point out that the sliced stable property is also an intrinsic (geometric) property of $\boldsymbol{m}(y)$. i.e., It only depends on the curve $\boldsymbol{m}(y)$ itself and does not depend on its embeding into the ambient space.

Hsing and Carroll [1992] (later used in Zhu et al. [2006], Zhu and Ng [1995]) have introduced the following conditions on the central curve to prove the consistency of SIR.

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

$$\lim_{n\to\infty} \sup_{y\in\Pi_n(B)} n^{-1/4} \sum_{i=2}^{n} \|\boldsymbol{m}(y_i) - \boldsymbol{m}(y_{i-1})\|_2 = 0, \forall B > 0. \tag{6}$$

Second, they assumed that for $B_0 > 0$, there exists a non-decreasing function $\widetilde{m}(y)$ on (B_0, ∞), such that

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

We conjecture that the sliced stable condition can be derived from this condition.

REMARK 2. *Two special forms of sliced stable condition.*

i) Choosing $\boldsymbol{\beta}^\tau = (0, 0, \cdots, 0, 1, 0, \cdots, 0)$ where 1 is at the k-th position, one has

$$|\sum_{h=0}^{H} var(\boldsymbol{m}(y,k)|a_h \leq y \leq a_{h+1})| \leq \mathfrak{a} H^{1-\kappa} var(\boldsymbol{m}(y,k)) \tag{7}$$

where $\boldsymbol{m}(y,k)$ is the k-th coordinate of the central curve $\boldsymbol{m}(y)$.

ii) Since the equation (5) holds for any unit vector $\boldsymbol{\beta}$, one has

$$\|\sum_{h=0}^{H} var(\boldsymbol{m}(y)|a_h \le y \le a_{h+1})\|_2 \le \mathfrak{a} H^{1-\kappa} \|var(\boldsymbol{m}(y))\|_2 \tag{8}$$

Now, we are ready to state our main results.

THEOREM 1. *Assuming the conditions* (**T1**) *and* (**A1**) − (**A3**)*, for sufficiently large* H *and* n*, one has:*

$$\|\widehat{\boldsymbol{\Lambda}}_p - \boldsymbol{\Lambda}_p\|_2 \le O_P\left(\frac{1}{H^{\kappa \wedge 1}} + \frac{H^2 p}{n} + \sqrt{\frac{H^2 p}{n}}\right). \tag{9}$$

As a direct corollary of Theorem 1, if $\rho = \lim_{n\to\infty} \frac{p}{n} = 0$, one may choose $H = \log\left(\frac{n}{p}\right)$ such that the right hand side of equation (9) converges to 0. In other words, we have proved that $\widehat{\boldsymbol{\Lambda}}_p$ is a consistent estimate of $\boldsymbol{\Lambda}_p$ if $\rho = 0$.

THEOREM 2. *Assuming the conditions* (**T1**), (**A1**) − (**A3**), $\boldsymbol{x}$ *is sub-Gaussian and* $\rho = \lim \frac{p}{n} = 0$ *, then*

$$\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1} \widehat{\boldsymbol{\Lambda}}_p - \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1} \boldsymbol{\Lambda}_p\|_2 \to 0 \quad \text{as } n \to \infty$$

with probability converging to one where $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^{\tau}$.

We define the distance $\mathcal{D}(\boldsymbol{V}, \boldsymbol{V}')$ of two d-dimensional subspaces $\boldsymbol{V}$ and $\boldsymbol{V}'$ as the operator norm (or Frobenius norm) of $P_{\boldsymbol{V}} - P_{\boldsymbol{V}'}$ where $P_{\boldsymbol{V}}$ and $P_{\boldsymbol{V}'}$ are the projection matrices associated with these two spaces. Simple linear algebra shows that if $\widetilde{\boldsymbol{\beta}}_i$'s are the eigenvectors of the generalized eigen-vector problem

$$\boldsymbol{\Sigma}_{\boldsymbol{x}} \widetilde{\boldsymbol{\beta}}_i = \lambda_i \boldsymbol{\Lambda}_p \widetilde{\boldsymbol{\beta}}_i, \tag{10}$$

then

$$span\Big\{\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_d\Big\} = span\Big\{\widetilde{\boldsymbol{\beta}}_1, \cdots, \widetilde{\boldsymbol{\beta}}_d\Big\}.$$

Let $\widehat{\boldsymbol{\beta}_1}$,...,$\widehat{\boldsymbol{\beta}_d}$ be the top generalized eigenvectors of $(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}, \widehat{\boldsymbol{\Lambda}}_p)$. Recall that the d-th eigenvalue of $\boldsymbol{\Lambda}_p$ is assumed to be bounded away from 0. Therefore Theorem 2 implies that $\mathcal{D}(< \widehat{\boldsymbol{\beta}} >, < \boldsymbol{\beta} >) \to 0$ when $\rho = 0$.

REMARK 3. *Discussion on the sub-Gaussian assumption.* In Theorem 2, the sub-Gaussian assumption assures the consistency of $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}$ when $\rho =$

0. It can be replaced by sub-exponential assumption (See e.g., Adamczak et al. [2008]). In general, it is widely believed that $\widehat{\mathbf{\Sigma}}_{\boldsymbol{x}}$ converges to $\mathbf{\Sigma}_{\boldsymbol{x}}$ if $\rho = \lim_{n\to\infty} \frac{p}{n} = 0$. However, without further moment assumption, the best result so far requires $\lim \frac{p\log(p)}{n} = 0$, (See e.g., Vershynin [2010]).

We have already shown that, under mild conditions, the SIR procedure provides us with a consistent estimate of the sufficient dimension reduction space when $\rho = 0$. It is then natural to ask: is this condition necessary? Our next theorem gives the answer.

THEOREM 3. *Assuming the conditions* (**T1**), (**A1**) − (**A3**) *and* $\boldsymbol{x} \sim N(0, \mathbf{I}_p)$ *for the single index model*

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

one has:

i) When $\rho = \lim \frac{p}{n} \in (0, \infty)$, $\|\widehat{\mathbf{\Lambda}}_p - \mathbf{\Lambda}_p\|_2$ *is (as a function of* ρ *) dominated by* $\sqrt{\rho} \vee \rho$ *if* $H, n \to \infty$*;*

ii) Let $\widehat{\boldsymbol{\beta}}$ *be the principal eigenvector of the SIR estimator* $\widehat{\mathbf{\Lambda}}_p$*. If* $\rho = \lim \frac{p}{n} > 0$*, then there exists a positive constant* $c(\rho) > 0$ *such that*

$$\liminf_{n\to\infty} \mathbb{E}\angle(\boldsymbol{\beta}, \widehat{\boldsymbol{\beta}}) > c(\rho)$$

with probability converges to one.

We illustrated this result via the numerical studies of the linear model

$$y = \boldsymbol{x}^\tau \boldsymbol{\beta} + \epsilon \text{ where } \boldsymbol{\beta}^\tau = (1, 0, \cdots, 0), \boldsymbol{x} \sim N(0, \mathbf{I}_p), \epsilon \sim N(0, 1).$$

In Figure 1, for fixed ratio $\rho = \frac{p}{n}$ which varies among $\{\ .1, .3, .7, 1, 2, 4\ \}$ across all the panels, we have plotted the $\mathbb{E}\angle(\boldsymbol{\beta}, \widehat{\boldsymbol{\beta}})$ against the dimension p where the $\boldsymbol{\beta}$ is estimated by the SIR with the slice number $H = 10$. For each p, the $\mathbb{E}\angle(\boldsymbol{\beta}, \widehat{\boldsymbol{\beta}})$ is calculated based on 100 iterations. It is seen that this expected angle converges to a positive number when the ratio ρ is non-zero. In Figure 2, we have plotted the $\mathbb{E}\angle(\boldsymbol{\beta}, \widehat{\boldsymbol{\beta}})$ against the ratio $\rho = \frac{p}{n}$, varying between 0.01 and 4 with an increment of 0.01. The sample size n is 200 and the slice number H is 10. It is seen that the expected angle decreases to zero as the ratio approaches zero. When the ratio increases, the expected angle increases, preventing the SIR from producing consistent estimation.

Results in this section have shown that there is a phase transition phenomenon of the SIR procedure. That is, the estimate of the dimension reduction space is consistent if and only if the ratio $\rho = \lim \frac{p}{n} = 0$. This provides a theoretical justification of imposing additional structure assumption such as sparsity in high dimension.

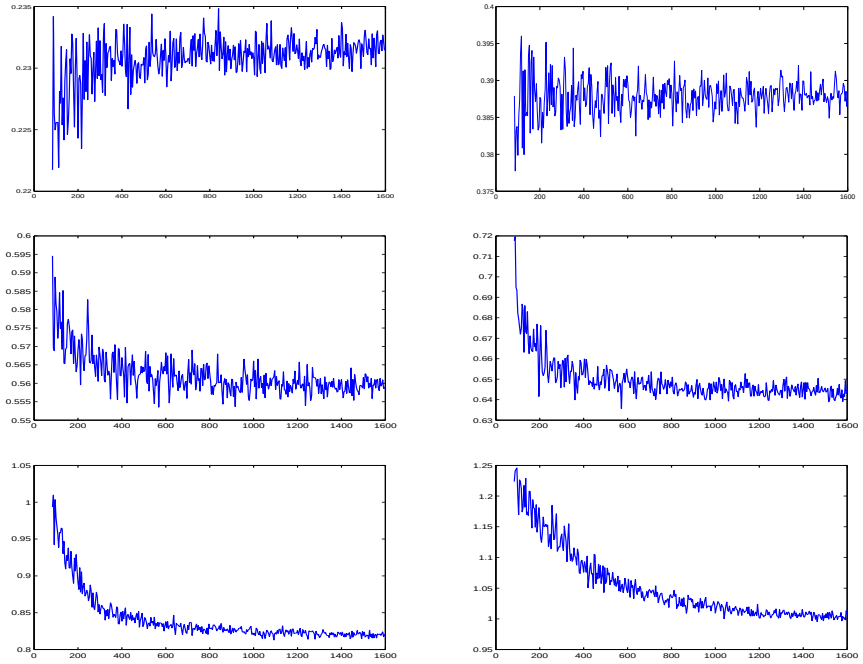

Fig 1: Simulated value of $\mathbb{E}\angle(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$ as function of dimension p for $\rho = .1, .3, .7, 1, 2, 4$ (up left, up right,middle left, middle right, lower left, lower right resp.) where $\widehat{\boldsymbol{\beta}}$ is estimated by SIR.

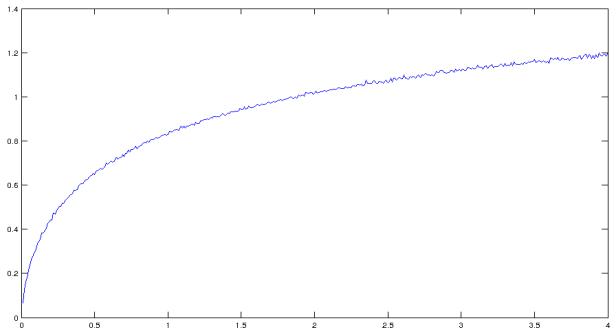

Fig 2: The relationship of $\mathbb{E}\angle(\boldsymbol{\beta}, \widehat{\boldsymbol{\beta}})$ and the ratio p/n where $\widehat{\boldsymbol{\beta}}$ is estimated by SIR.

4. SIR in ultra-high dimension. As we have shown in Section 3, the SIR estimator fails to be consistent if $\rho = \lim \frac{p}{n} \neq 0$. Hence, when $p \gg n$, some structural assumptions are necessary for getting a consistent estimate of the central space. In this paper, we assume that both the loadings of all the directions $\boldsymbol{\beta}_j$'s and the covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ are sparse. Other structural assumptions will be studied in our future work. For $\boldsymbol{\beta}_i$'s, we impose the following prevalent sparsity condition.

- **(A4)** $s = |\mathcal{S}| \ll p$ where $\mathcal{S} = \left\{ i \mid \boldsymbol{\beta}_j(i) \neq 0 \text{ for some } j,\ 1 \leq j \leq d \right\}$ and $|\mathcal{S}|$ is the number of elements in the set $\mathcal{S}$.

For $\boldsymbol{\Sigma}_{\boldsymbol{x}}$, the following class of covariance matrices has been introduced in Bickel and Levina [2008]. (See also Cai et al. [2010].)

$$\mathcal{U}(\epsilon_0, \alpha, C) = \Big\{ \boldsymbol{\Sigma}_{\boldsymbol{x}} : \max_j \sum_i \{|\sigma_{i,j}| : |i-j| > l\} \leq C l^{-\alpha} \text{ for all } l > 0, \\ \text{and } 0 < \epsilon_0 \leq \lambda_{min}(\boldsymbol{\Sigma}_{\boldsymbol{x}}) \leq \lambda_{max}(\boldsymbol{\Sigma}_{\boldsymbol{x}}) \leq \frac{1}{\epsilon_0} \Big\}.$$

In this paper, to simplify the notations and arguments, we choose a slightly stronger condition.

- **(A5)** $\boldsymbol{\Sigma}_{\boldsymbol{x}} \in \mathcal{U}(\epsilon_0, \alpha, C)$ and $\max_{1 \leq i \leq p} r_i$ is bounded where r_i is the number of non-zero elements in the i-th row of $\boldsymbol{\Sigma}_{\boldsymbol{x}}$.

Let $\mathcal{T} = \left\{ k \mid var\left(\mathbb{E}[\boldsymbol{x}(k)|y]\right) \neq 0 \right\}$. Note that $var\left(\mathbb{E}[\boldsymbol{x}(k)|y]\right) = 0$ if and only if the k-th coordinate of $\boldsymbol{\eta}_j$ is zero for all $j = 1, \cdots, d$ where $\boldsymbol{\eta}_i$'s are the eigenvectors of $\boldsymbol{\Lambda}_p$. With the above sparsity assumptions **(A4)** and **(A5)** , it is easy to see $|\mathcal{T}| = O(s)$ from equation (2) . On the population level, $var(\mathbb{E}(\boldsymbol{x}(k))|y)$ can separate $\mathcal{T}$ from $\mathcal{T}^c$. When there are only finite samples, we use

$$var_{H,c}(\boldsymbol{x}(k)) = \frac{1}{H-1} \sum_{h=1}^{H} (\bar{\boldsymbol{x}}_{h,\cdot}(k) - \overline{\overline{\boldsymbol{x}}}(k))^2. \tag{11}$$

as an estimate of $var(\mathbb{E}(\boldsymbol{x}(k))|y)$. These are the diagonal elements of the matrix $\widehat{\boldsymbol{\Lambda}}_p$. Note that these quantities depend on the sliced sample means, which are neither independent nor identically distributed, the usual concentration inequalities for χ^2 are no longer applicable. One needs extra efforts to get the concentration inequalities; this concentration result is one of the main technical contributions of this paper, which will be extended in our future research.

REMARK 4. *Connection to other screening statistics.* The link function f is not involved explicitly in the definition of $var_{H,c}(\boldsymbol{x}(k))$; only the order statistics of response is required. This nonparametric characteristic of the method is of particular interest of us in future research. Screening statistics inspired by the sliced inverse regression idea have been proposed in various formats. (See, e.g., Jiang and Liu [2014], Zhu et al. [2011b] and Cui et al. [2014]).

With the quantities $var_{H,c}(\mathbb{E}[\boldsymbol{x}(k)|y])$, we define the inclusion set $\mathcal{I}_p(t)$ and the exclusion set $\mathcal{E}_p(t)$,which depend on a thresholding value t, as following:

(12)
$$\mathcal{I}_p(t) = \left\{ k \mid var_{H,c}(\boldsymbol{x}(k)) > t \right\} \text{ and } \mathcal{E}_p(t) = \left\{ k \mid var_{H,c}(\boldsymbol{x}(k)) \leq t \right\}.$$

Note that $\mathcal{I}_p(t)$ can be viewed as an estimate of $\mathcal{T}$ and is thus also denoted by $\widehat{\mathcal{T}}$. After reducing the dimension to a level that can be handled (i.e., $|\widehat{\mathcal{T}}| \prec n$ or $\lim \frac{|\widehat{\mathcal{T}}|}{n} = 0$), the SIR estimator $\widehat{\boldsymbol{\Lambda}}^{\widehat{\mathcal{T}},\widehat{\mathcal{T}}}$ is a consistent estimate of $\boldsymbol{\Lambda}^{\mathcal{T},\mathcal{T}}$, which is guaranteed by Theorem 1. Let $\widehat{\boldsymbol{\eta}}_i^{\widehat{\mathcal{T}}}$ $i = 1, ..., d$ be the top d eigenvectors of $\widehat{\boldsymbol{\Lambda}}^{\widehat{\mathcal{T}},\widehat{\mathcal{T}}}$. We embed them into $\mathbb{R}^p$ by filling 0 for entries outside $\widehat{\mathcal{T}}$, which we denote by $\widehat{\boldsymbol{\eta}}_i$. Finally, we use $< \{ \widehat{\boldsymbol{\beta}}_i \} >$ to estimate space $< \{ \boldsymbol{\beta}_i \} >$ where $\widehat{\boldsymbol{\beta}}_i = \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1} \widehat{\boldsymbol{\eta}}_i$ and $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}$ is a consistent estimate of $\boldsymbol{\Sigma}_{\boldsymbol{x}}$. Estimating the covariance matrix and precision matrix in high dimensional data is a challenging problem alone and not the main focus of this paper. We estimate them using the existing methods from Bickel and Levina [2008].

We summarize the above procedures into the following two-stage algorithm: **D**iagonal **T**hresholding screening **SIR** (DT-SIR).

ALGORITHM (DT-SIR).

1. *Calculate* $var_{H,c}(\boldsymbol{x}(k))$ *according* (11) *for* $k = 1, 2, \cdots, p$;
2. *Let* $\widehat{\mathcal{T}} = \left\{ k \mid var_{H,c}(\boldsymbol{x}(k)) > t \right\}$ *for an appropriate* t ;
3. *Let* $\widehat{\boldsymbol{\Lambda}}_p^{\widehat{\mathcal{T}},\widehat{\mathcal{T}}}$ *be the SIR estimator of the conditional covariance matrix for the data* $(y, \boldsymbol{x}^{-,\widehat{\mathcal{T}}})$ *according to equation* (4);
4. *Calculate* $\widehat{\boldsymbol{\eta}}_i = e(\widehat{\boldsymbol{\eta}}_i^{\widehat{\mathcal{T}}})$ *where* $\widehat{\boldsymbol{\eta}}_i^{\widehat{\mathcal{T}}} (1 \leq i \leq d)$*'s are the top eigenvectors of* $\widehat{\boldsymbol{\Lambda}}^{\widehat{\mathcal{T}},\widehat{\mathcal{T}}}$;

5. *Calculate* $\widehat{\boldsymbol{\beta}}_i = \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}\widehat{\boldsymbol{\eta}}_i$ *where* $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}$ *is a consistent estimate of* $\boldsymbol{\Sigma}_{\boldsymbol{x}}$*;*

6. *The central space is estimated by* $< \{ \widehat{\boldsymbol{\beta}}_i \} >$.

A practical way to choose the appropriate t in step 2 will be presented in Section 5 . To ensure the theoretical properties, we need the assumption on the signal strength.

- (**S1**) There exist positive constants C and ω such that $var(\mathbb{E}[\boldsymbol{x}(k)|y]) > \frac{C}{s^\omega}$ when $\mathbb{E}[\boldsymbol{x}(k)|y]$ is not constant.

We aslo need the assumption on the tail distribution of $\boldsymbol{x}$.

- (**T2**) There exists a constant K such that every coordinate $\boldsymbol{x}(k)$ is sub-Gaussian and upper-exponentially bounded by K. (For the definition, see e.g., Definition 3 .)

With these conditions, we now have :

THEOREM 4. *Assuming model (1), conditions* **(T1)** *and* **(A1-A5)**, *the signal strength condition* **(S1)**, *and sub-Gaussian assumption* **(T2)**, *let* $t = \frac{a}{s^\omega}$ *where a is a sufficiently small positive constant such that* $t < \frac{1}{2}var(m(y,k))$ *for any* $k \in \mathcal{T}$, *one has*

i) $\mathcal{T}^c \subset \mathcal{E}_p$ *holds with probability at least*

$$1 - C_1 \exp\left(-C_2 \frac{n}{H^2 s^\omega} + C_3 \log(H) + \log(p-s)\right); \tag{13}$$

ii) $\mathcal{T} \subset \mathcal{I}_p$ *holds with probability at least*

$$1 - C_4 \exp\left(-C_5 \frac{n}{H^2 s^\omega} + C_6 \log(H) + \log(s)\right), \tag{14}$$

for some positive constants $C_1, \cdots, C_6$.

Theorem 4 has a simple implications. If $\frac{n}{s^\omega} \succ \log(p) + \log(s)$, one may choose $H = \log(\frac{n}{s^\omega \log(p)})$, so that

$$\frac{n}{H^2 s^\omega} \succ \log(p) + \log(H) + \log(s),$$

from which , we know $\mathcal{T} = \mathcal{I}_p$ with probability convergences to one.

Next, we have the following theorems for the consistency of DT-SIR.

THEOREM 5. *Under the same assumptions of Theorem 4, we choose t as described in Theorem 4,* $\widehat{\mathcal{T}} = \mathcal{I}(t)$ *and* $H = \log(\frac{n}{s^\omega \log(p)})$, *then*

$$\|e(\widehat{\mathbf{\Lambda}}_p^{\widehat{T},\widehat{T}}) - \mathbf{\Lambda}_p\|_2 \to 0 \text{ as } n \to \infty$$

with probability converges to one.

As a direct corollary, we know that

THEOREM 6. *Let $\widehat{\mathbf{\Sigma}}_{\boldsymbol{x}}$ be the estimator of co-variance matrix from Bickel and Levina [2008]. Under the same assumptions of Theorem 5, one has*

$$\|\widehat{\mathbf{\Sigma}}_X^{-1} e(\widehat{\mathbf{\Lambda}}_p^{\widehat{T},\widehat{T}}) - \mathbf{\Sigma}_X^{-1}\mathbf{\Lambda}_p\|_2 \to 0 \text{ as } n \to \infty$$

with probability converges to one.

5. Simulation. We consider the following settings in generating the design matrix $\boldsymbol{x}$ and the response y. In Settings I-III, each row of $\boldsymbol{x}$ is independently sampled from $N(\mathbf{0}, \mathbf{I})$.

- **Setting I.** $y_i = \sin(x_{i1} + x_{i2}) + \exp(x_{i3} + x_{i4}) + 0.5 * \epsilon_i$, where $\epsilon_i \overset{\text{iid}}{\sim} N(0, 1)$;
- **Setting II.** $y_i = \sum_{j=1}^{7} x_{ij} * \exp(x_{i8} + x_{i9}) + 0.5 * \epsilon_i$ where $\epsilon_i \overset{\text{iid}}{\sim} N(0, 1)$;
- **Setting III.** $y_i = \sum_{j=1}^{10} x_{ij} * \exp(\sum_{i=11}^{20} x_{ij}) + \epsilon_i$ where $\epsilon_i \overset{\text{iid}}{\sim} N(0, 1)$;

In Settings IV to VI, each row of $\boldsymbol{x}$ is independently sampled from $N(\mathbf{0}, \mathbf{\Sigma})$.

- **Setting IV.** $y_i = (x_{i1} + x_{i2} + x_{i3})^3/2 + 0.5 * \epsilon_i$, where $\epsilon_i \overset{\text{iid}}{\sim} N(0, 1)$ and $\mathbf{\Sigma} = (\sigma_{ij})$ is tridiagonal with $\sigma_{ii} = 1$, $\sigma_{i,i+1} = \sigma_{i+1,i} = \rho$ and $\sigma_{i,i+2} = \sigma_{i+2,i} = \rho^2$;
- **Setting V.** $y_i = \sum_{j=1}^{7} x_{ij} * \exp(x_{i8} + x_{i9}) + \epsilon_i$, where $\epsilon_i \overset{\text{iid}}{\sim} N(0, 1)$, and $\mathbf{\Sigma} = \boldsymbol{B} \otimes \mathbf{I}_{p/10}$ with $\boldsymbol{B} = (b_{ij})_{1 \le i \le 10, 1 \le j \le 10}$ given as $b_{ij} = \rho^{|i-j|}$;
- **Setting VI.** Assume the same setting as in Setting V except that $\mathbf{\Sigma} = (\sigma_{ij})$ is tridiagonal with $\sigma_{ii} = 1, \sigma_{i,i+1} = \sigma_{i+1,i} = \rho$ and $\sigma_{i,i+2} = \sigma_{i+2,i} = \rho^2$.

The following methods are applied to the sample $(\boldsymbol{x}_i, y_i)$. In DT-SIR, we first screen all the predictors according to the statistic $var_{H,c}(\boldsymbol{x}(k))$. The second step requires a tuning parameter t which is chosen by using the auxiliary variable, an idea first proposed by Luo et al. [2006] and extended by Wu et al. [2007] and Zhu et al. [2011a]. In our setting, for a given sample $(y_i, \boldsymbol{x}_i)$, we generate $\boldsymbol{z}_i \sim N(\mathbf{0}, \mathbf{I}_{p'})$ where p' is sufficiently large and chosen as p in our simulations. It is known that $\boldsymbol{y}$ and $\boldsymbol{z}$ are independent. Choose the threshold t as

$$\hat{t} = \max_{1 \le k \le p'} \{var_{H,c}(\boldsymbol{z}(k))\}$$

to obtain the inclusion set $\mathcal{I}_p(\hat{t})$. We then continue the calculation with steps 3-5 as described in the algorithm.

We also consider alternative methods in the screening step, such as Sure Independent Ranking and Screening (SIRS) in Zhu et al. [2011a] and SIR for variable selection via Inverse modeling (SIRI) in Jiang and Liu [2014]. For SIRS, the threshold is chosen according to the auxiliary statistic (2.9) of Zhu et al. [2011a]. For SIRI, the predictors are chosen according to 10-fold cross validation. After the screening step, similar to DT-SIR, we then apply steps 3-5 in the algorithm to estimate $\boldsymbol{\beta}$. These two methods are denoted as SIRS-SIR and SIRI-SIR in the following discussions. Another method that we compare with is the sparse SIR, abbreviated as SpSIR, proposed in Li [2007].

After obtaining an estimator $\hat{\boldsymbol{\beta}}$, we calculate $\mathcal{D}(< \hat{\boldsymbol{\beta}} >, < \boldsymbol{\beta} >)$, the distance between estimated space $< \hat{\boldsymbol{\beta}} >$ and the true space $< \boldsymbol{\beta} >$, as a measure of the estimation error. We replicate this step 100 times, and calculate the average distance based on these four methods and report these numbers in Table 1. For each setting, the average distance of the optimal method is highlighted using bold fonts. We further run a two-sample T-test to test if the actual estimation error based on each method is significantly different from that based on the optimal method.

Under all the settings, the DT-SIR is much smaller than SpSIR. The p-values for comparing DT-SIR and SpSIR are all smaller than 0.01. When $p \geq n$, the sparse SIR completely fails because the average distance of the estimated space to the true space is $\sqrt{2d}$, indicating that the space estimated by sparse SIR is perpendicular to the true space spanned by β.

Under settings III-VI, the DT-SIR is the best among all the four methods except for the case when $n = 500, p = 1000$. The small p-values further indicate that the differences are significant. When $n = 500$ and $p = 1000$ under settings IV-VI, the average distance of SIRI-SIR is the smallest. However, there is no or weak evidence showing that the estimation error based on DT-SIR is significant different from that based on SIRI-SIR. Under settings I-II, the average distance of DT-SIR is not always the smallest. However, for most cases, there is no significant difference between DT-SIR and the optimal method. There are only two exceptions that we would like to point out. When $n = 500, p = 1000$ under setting I, the DT-SIR is worse than SIRS-SIR; when $n = 2000, p = 2000$ under setting II, DT-SIR is worse than SIRI-SIR.

To graphically show the performance of various methods, we consider the setting IV with $d = 1$. Consider two cases when $(n, p) = (2000, 1000)$ and $(n, p) = (500, 100)$. We calculate the estimated directions $\hat{\boldsymbol{\beta}}$ using various

TABLE 1

The average distance of the space estimated by the various methods to the true space spanned by $< \beta >$ under various settings. The "" in cells represent the level of significance when running the two-sample T-test comparing actual estimation error based on DT-SIR and its competitor: (***)–p-value < 0.01 ; (**)– $0.01 <$ p-value ≤ 0.05; (*)– $0.05 <$ p-value ≤ 0.1.*

		DT-SIR	SIRI-SIR	SIRS-SIR	SpSIR		DT-SIR	SIRI-SIR	SIRS-SIR	SpSIR
	n	p=1000				p	n=2000			
I	500	0.655(***)	0.751(***)	**0.492**	2(***)	500	0.213	0.312(***)	**0.206**	1.44(***)
	1000	**0.3**	0.431(***)	0.309	2(***)	1000	**0.221**	0.341(***)	0.226	1.58(***)
	2000	**0.221**	0.341(***)	0.226	1.58(***)	2000	0.241	0.29(**)	**0.214**	2(***)
	3000	0.167	0.245(***)	**0.149**	1.48(***)	3000	0.23	0.278(**)	**0.218**	2(***)
II	500	0.383	0.396	**0.371**	2(***)	500	0.163	**0.16**	0.19(***)	0.83(***)
	1000	0.235	**0.227**	0.256(**)	2(***)	1000	0.161	**0.157**	0.189(***)	1.25(***)
	2000	0.161	**0.157**	0.189(***)	1.25(***)	2000	0.172(**)	**0.159**	0.196(***)	2(***)
	3000	0.134	**0.129**	0.153(***)	0.975(***)	3000	0.164	**0.158**	0.199(***)	2(***)
III	500	**1.15**	1.48(***)	1.38(***)	2(***)	500	**0.272**	0.353(**)	0.29(***)	0.916(***)
	1000	**0.426**	0.974(***)	0.596(***)	2(***)	1000	**0.263**	0.403(***)	0.29(***)	1.33(***)
	2000	**0.263**	0.403(***)	0.29(***)	1.33(***)	2000	**0.262**	0.368(**)	0.285(***)	2(***)
	3000	**0.214**	0.297(**)	0.238(***)	1.06(***)	3000	**0.269**	0.344(*)	0.291(***)	2(***)
IV	500	0.263	**0.257**	0.333	1.41(***)	500	**0.145**	0.409(***)	0.182(***)	0.248(***)
	1000	**0.219**	0.447(***)	0.25(*)	1.41(***)	1000	**0.161**	0.4(***)	0.196(***)	0.42(***)
	2000	**0.161**	0.4(***)	0.196(***)	0.42(***)	2000	**0.16**	0.395(***)	0.198(***)	1.41(***)
	3000	**0.134**	0.377(***)	0.177(***)	0.297(***)	3000	**0.15**	0.395(***)	0.216(***)	1.41(***)
V	500	0.546	**0.529**	0.562(**)	2(***)	500	**0.272**	0.434(***)	0.353(***)	1.09(***)
	1000	**0.401**	0.463(***)	0.514(***)	2(***)	1000	**0.288**	0.418(***)	0.341(***)	1.51(***)
	2000	**0.288**	0.418(***)	0.341(***)	1.51(***)	2000	**0.289**	0.418(***)	0.351(***)	2(***)
	3000	**0.249**	0.399(***)	0.284(***)	1.24(***)	3000	**0.3**	0.417(***)	0.372(***)	2(***)
VI	500	0.568(*)	**0.535**	0.566	2(***)	500	**0.307**	0.479(***)	0.368(***)	1.1(***)
	1000	**0.427**	0.524(***)	0.548(***)	2(***)	1000	**0.311**	0.469(***)	0.351(***)	1.51(***)
	2000	**0.311**	0.469(***)	0.351(***)	1.51(***)	2000	**0.309**	0.461(***)	0.399(***)	2(***)
	3000	**0.265**	0.456(***)	0.307(***)	1.25(***)	3000	**0.31**	0.46(***)	0.408(***)	2(***)

methods and compute the angle between $< \hat{\boldsymbol{\beta}} >$ and $< \boldsymbol{\beta} >$. We replicate this step 100 times to calculate the average angles based on all the methods. The result are displayed in Figure 3. The DT-SIR is better than SIRI-SIR and SpSIR in the left panel and is better than DC-SIR and SpSIR in the right panel.

TABLE 2

Comparison of computing time under setting II.

		DT-SIR	SIRI-SIR	SIRS-SIR	SpSIR		DT-SIR	SIRI-SIR	SIRS-SIR	SpSIR
	n	p=1000				p	n=2000			
II	500	1"	1'47"	10"	5"	500	1"	4'15"	1'5"	1"
	1000	1'	2'58"	34"	5"	1000	2"	5'16"	2'11"	6"
	2000	2"	5'16"	2'11"	6"	2000	8"	7'25"	4'41"	41"
	3000	3"	7'39"	4'56"	7"	3000	19"	9'13"	7'40"	2'9"

The DT-SIR is computationally efficient. To show this, the computing time for one replication under Setting II for various pairs of (n, p) is reported in Table 2. The comparison is done on a computer with Intel i5-3330 CPU@3.00GHz and 8GB memory. It is clearly seen that DT-SIR is much faster than all its competitors. Consider the case when $p = 3000, n = 2000$. The computation time of DT-SIR is only 19 seconds; while the time for SIRI-SIR is 9 minutes 13 seconds and the time for SIRS-SIR is 7 minutes

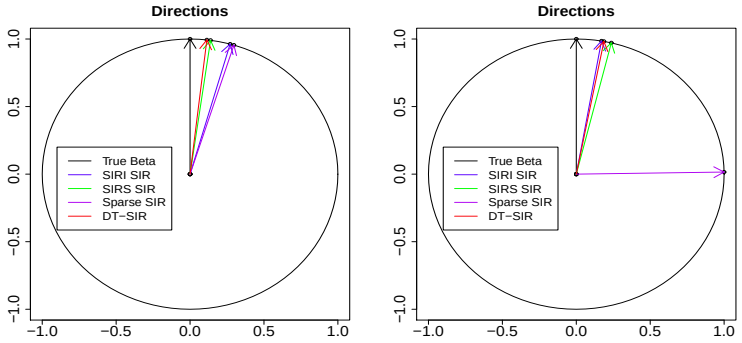

Fig 3: Simulated value of $E\angle(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$ for the various methods. Left panel: $(n, p) = (2000, 1000)$; Right panel: $(n, p) = (500, 1000)$.

40 seconds. Here, the SIRI-SIR needs significant time mainly due to cross-validations. This comparison clearly demonstrate the advantage of DT-SIR in the high dimensional data analysis.

6. Conclusion. When the dimension p diverges to infinity, classical statistical procedure often fails unless additional structures such as sparsity conditions were imposed. Understanding boundary conditions of a statistical procedure provides theoretical justification and practical guidance. In this paper, we provide a new framework to show that $\rho = \lim \frac{p}{n}$ is the phase transition parameter of the SIR procedure. Under certain conditions, it is shown that the SIR estimator is consistent if and only if $\rho = 0$. When $\rho > 0$, the original SIR fails to be consistent. We thus propose the two-stage method, DT-SIR for ultra-high dimension reduction which is shown to be consistent. We have used simulated examples to demonstrate the advantages of DT-SIR compared to its competitors. This method is computationally fast and can be easily implemented for large data sets.

7. Appendix A: Proof of theorems in Section 3. The proofs of the main theorems depend on many assisting lemmas which are put in Section 9.

7.1. *Outline of the Proof of Theorem 1.* Let $\mathcal{S}$ be the central subspace of dimension $d \ll p$. i.e., $y \perp\!\!\!\perp \boldsymbol{x} | \boldsymbol{P}_{\mathcal{S}}\boldsymbol{x}$ and $dim(\mathcal{S}) = d$. One has the decom-

position

$$\begin{aligned} \boldsymbol{x} &= \boldsymbol{P}_{\mathcal{S}}\boldsymbol{x} + \boldsymbol{P}_{\mathcal{S}^\perp}\boldsymbol{x} \triangleq \boldsymbol{z} + \boldsymbol{w} \\ &= \mathbb{E}[\boldsymbol{z}|y] + \boldsymbol{z} - \mathbb{E}[\boldsymbol{z}|y] + \boldsymbol{w} \triangleq \boldsymbol{m} + \boldsymbol{v} + \boldsymbol{w} \end{aligned} \tag{15}$$

where $\boldsymbol{z} = \boldsymbol{P}_{\mathcal{S}}\boldsymbol{x}$, $\boldsymbol{m} = \mathbb{E}[\boldsymbol{z}|y]$, $\boldsymbol{v} = \boldsymbol{z} - \mathbb{E}[\boldsymbol{z}|y]$ and $\boldsymbol{w} = \boldsymbol{P}_{\mathcal{S}^\perp}\boldsymbol{x}$. Note that $\boldsymbol{m}$ lies in the central curve, $\boldsymbol{v}$ lies in the central space and $\boldsymbol{w}$ lies in the space perpendicular to $\mathcal{S}$.

For a given data set $(y, \boldsymbol{x})$, the SIR procedure sorts and divides $n = Hc$ samples into H slices of equal size according to the order statistics $y_{(i)}$. In this subsection, instead of working directly with the estimator $\widehat{\boldsymbol{\Lambda}}_p$ in (4), we consider a simpler estimator

$$\widetilde{\boldsymbol{\Lambda}}_p \triangleq \frac{1}{H}\sum_{h=1}^{H} \overline{\boldsymbol{x}}_{h,\cdot}\overline{\boldsymbol{x}}_{h,\cdot}^{\tau}$$

of $\boldsymbol{\Lambda}_p$ as an intermedium where $\overline{\boldsymbol{x}}_{h,\cdot}$ is the sample mean of the h-th slice.

For $i = 1, 2, \cdots, n$, we can decompose each sample $\boldsymbol{x}_i$ as $\boldsymbol{z}_i + \boldsymbol{w}_i (= \boldsymbol{m}_i + \boldsymbol{v}_i + \boldsymbol{w}_i)$. Similar to the definition of $\boldsymbol{x}_{h,j}$, $\overline{\boldsymbol{x}}_{h,\cdot}$ and $\overline{\overline{\boldsymbol{x}}}$ (See e.g., equation (3)), we can define $\boldsymbol{m}_{h,j}$, $\overline{\boldsymbol{m}}_{h,\cdot}$, $\overline{\overline{\boldsymbol{m}}}$, $\boldsymbol{z}_{h,j}$, $\overline{\boldsymbol{z}}_{h,\cdot}$, $\overline{\overline{\boldsymbol{z}}}$, $\boldsymbol{v}_{h,j}$, $\overline{\boldsymbol{v}}_{h,\cdot}$, $\overline{\overline{\boldsymbol{v}}}$ and $\boldsymbol{w}_{h,j}$, $\overline{\boldsymbol{w}}_{h,\cdot}$, $\overline{\overline{\boldsymbol{w}}}$, according to the order statistics $y_{(i)}$ respectively. Consequently, we can define $\widetilde{\boldsymbol{\Lambda}}_{\boldsymbol{m}}$ and $\widetilde{\boldsymbol{\Lambda}}_{\boldsymbol{z}}$. We will prove $\|\widetilde{\boldsymbol{\Lambda}}_{\boldsymbol{m}} - \boldsymbol{\Lambda}_{\boldsymbol{p}}\|_2 \to 0$, $\|\widetilde{\boldsymbol{\Lambda}}_{\boldsymbol{z}} - \boldsymbol{\Lambda}_{\boldsymbol{p}}\|_2 \to 0$, $\|\widetilde{\boldsymbol{\Lambda}}_{\boldsymbol{p}} - \boldsymbol{\Lambda}_{\boldsymbol{p}}\|_2 \to 0$ and $\|\widetilde{\boldsymbol{\Lambda}}_{\boldsymbol{p}} - \widehat{\boldsymbol{\Lambda}}_{\boldsymbol{p}}\|_2 \to 0$ sequentially.

LEMMA 1. *Assuming the conditions in Theorem 1, one has*

$$\|\widetilde{\boldsymbol{\Lambda}}_{\boldsymbol{m}} - \boldsymbol{\Lambda}_p\|_2 \leq O_P\left(\frac{\sqrt{d}H^2}{\sqrt{n}}\right) + O_P\left(\frac{1}{H^\kappa}\right),$$

and

$$\|\widetilde{\boldsymbol{\Lambda}}_{\boldsymbol{m}}\|_2 = \|\boldsymbol{\Lambda}_p\|_2 + O_P\left(\frac{\sqrt{d}H^2}{\sqrt{n}}\right) + O_P\left(\frac{1}{H^\kappa}\right),$$

where the right hand side is bounded when H and n are sufficiently large.

LEMMA 2. *Assuming the conditions in Theorem 1, one has*

$$\|\widetilde{\boldsymbol{\Lambda}}_{\boldsymbol{z}} - \boldsymbol{\Lambda}_p\|_2 \leq O_P\left(\frac{\sqrt{d}H^2}{\sqrt{n}}\right) + O_P\left(\frac{1}{H^\kappa}\right),$$

and

$$\|\widetilde{\boldsymbol{\Lambda}}_{\boldsymbol{z}}\|_2 = \|\boldsymbol{\Lambda}_p\|_2 + O_P\left(\frac{\sqrt{d}H^2}{\sqrt{n}}\right) + O_P\left(\frac{1}{H^\kappa}\right),$$

where the right hand side is bounded.

The slight difference between Lemma 1 and Lemma 2 is that there is an extra randomness $\boldsymbol{v}$ in $\boldsymbol{z}$, so one needs additional efforts to bound it.

LEMMA 3. *Assuming the conditions in Theorem 1, one has*

$$\|\widetilde{\boldsymbol{\Lambda}}_{\boldsymbol{p}} - \boldsymbol{\Lambda}_p\|_2 \leq O_P\left(\frac{H^2 p}{n} + \frac{1}{H^\kappa} + \sqrt{\frac{H^2 p}{n}}\right),$$

and

$$\|\widetilde{\boldsymbol{\Lambda}}_{\boldsymbol{p}}\|_2 = \|\boldsymbol{\Lambda}_p\|_2 + O_P\left(\frac{H^2 p}{n} + \frac{1}{H^\kappa} + \sqrt{\frac{H^2 p}{n}}\right).$$

where the right hand side is bounded if one chooses $H = \log\left(\frac{n}{p}\right)$ *when* $\rho = \lim_{n\to\infty} \frac{p}{n} = 0$.

Theorem 1 follows from Lemma 3. Note that

$$\widehat{\boldsymbol{\Lambda}}_{\boldsymbol{p}} - \widetilde{\boldsymbol{\Lambda}}_{\boldsymbol{p}} = \frac{1}{H-1}\widetilde{\boldsymbol{\Lambda}}_{\boldsymbol{p}} - \frac{H}{H-1}\overline{\overline{\boldsymbol{x}}}\overline{\overline{\boldsymbol{x}}}^\tau. \tag{16}$$

Since $\|\overline{\overline{\boldsymbol{x}}}\|_2^2 = O_P(\frac{p}{n})$ and $\|\boldsymbol{\Lambda}_p\|_2$ is bounded, the above difference between $\widetilde{\boldsymbol{\Lambda}}_{\boldsymbol{p}}$ and $\widehat{\boldsymbol{\Lambda}}_{\boldsymbol{p}}$ is bounded by $O_P(\frac{1}{H} + \frac{Hp}{n} + \sqrt{\frac{p}{n}})$.

$$\begin{aligned}\|\widehat{\boldsymbol{\Lambda}}_{\boldsymbol{p}} - \boldsymbol{\Lambda}_p\|_2 &\leq \|\widehat{\boldsymbol{\Lambda}}_{\boldsymbol{p}} - \widetilde{\boldsymbol{\Lambda}}_{\boldsymbol{p}}\|_2 + \|\widetilde{\boldsymbol{\Lambda}}_{\boldsymbol{p}} - \boldsymbol{\Lambda}_p\|_2 \\ &\leq O_P\left(\frac{H^2 p}{n} + \frac{1}{H^{\kappa\wedge 1}} + \sqrt{\frac{H^2 p}{n}}\right).\end{aligned} \tag{17}$$

7.2. *Proofs of Lemma 1 , Lemma 2 and Lemma 3.*

7.2.1. *Proof of Lemma 1.* In order to prove Lemma 1, one only needs to prove that for any ϵ, there exists a constant C, such that for any unit vector $\boldsymbol{\beta}$, one has

$$\mathbb{P}\left(\left|\boldsymbol{\beta}^\tau\left(\widetilde{\boldsymbol{\Lambda}}_{\boldsymbol{m}} - \boldsymbol{\Lambda}_p\right)\boldsymbol{\beta}\right| > C\left(\frac{\sqrt{d}H^2}{\sqrt{n}} + \frac{1}{H^\kappa}\right)\right) \leq \epsilon. \tag{18}$$

Below we will simply state it as : for any unit vector $\boldsymbol{\beta}$,

$$\left|\boldsymbol{\beta}^\tau\left(\widetilde{\boldsymbol{\Lambda}}_{\boldsymbol{m}} - \boldsymbol{\Lambda}_p\right)\boldsymbol{\beta}\right| \leq O_P\left(\frac{\sqrt{d}H^2}{\sqrt{n}}\right) + O_P\left(\frac{1}{H^\kappa}\right). \tag{19}$$

Since in all the proof below, we can choose the constant terms are invariant with respect to $\boldsymbol{\beta}$, this abuse of notation will not bring us troubles. Note that

$$\widetilde{\boldsymbol{\Lambda}}_{\boldsymbol{m}} - \boldsymbol{\Lambda}_p = \frac{1}{H}\sum \overline{\boldsymbol{m}}_{h,\cdot} \cdot \overline{\boldsymbol{m}}_{h,\cdot}^{\tau} - \boldsymbol{\Lambda}_p. \tag{20}$$

Let $\mu_h = \mathbb{E}[\boldsymbol{m}(y)\big|y \in S_h]$. For any unit vector $\boldsymbol{\beta}$, one has

$$\Big|\frac{1}{H}\sum_h (\boldsymbol{\beta}^{\tau}\overline{\boldsymbol{m}}_{h,\cdot})^2 - var(\boldsymbol{\beta}^{\tau}\boldsymbol{m}(y))\Big| \leq A_1 + A_2 \tag{21}$$

where

$$A_1 = \Big|\frac{1}{H}\sum_h (\boldsymbol{\beta}^{\tau}\mu_h)^2 - var(\boldsymbol{\beta}^{\tau}\boldsymbol{m}(y))\Big|, \tag{22}$$

$$A_2 = \Big|\frac{1}{H}\sum_h (\boldsymbol{\beta}^{\tau}\overline{\boldsymbol{m}}_{h,\cdot})^2 - \frac{1}{H}\sum_h (\boldsymbol{\beta}^{\tau}\mu_h)^2\Big|. \tag{23}$$

One only needs to prove $A_1 \leq O_P(\frac{1}{H^{\kappa}})$ and $A_2 \leq O_P(\frac{\sqrt{d}H^2}{\sqrt{n}})$.

For A_1, one has

LEMMA 4. *Let $\epsilon = \frac{1}{Hn_0+1}$ for a sufficiently large n_0 such that $\frac{\mathfrak{a}_1}{H} < \frac{1}{H} - \epsilon < \frac{1}{H} + \epsilon < \frac{\mathfrak{a}_2}{H}$, there exist positive constants C and C' such that $A_1 \leq \frac{C'}{H^{\kappa}} var(\boldsymbol{\beta}^{\tau}\boldsymbol{m}(y))$ with probability at least*

$$1 - CH^2\sqrt{Hc+1}\exp\left(-(Hc+1)\frac{\epsilon^2}{32}\right). \tag{24}$$

In particular,

$$A_1 \leq O_P(\frac{1}{H^{\kappa}}).$$

PROOF. For any unit vector $\boldsymbol{\beta}$, one has

$$\Big|\frac{1}{H}\sum_h (\boldsymbol{\beta}^{\tau}\mu_h)^2 - var(\boldsymbol{\beta}^{\tau}\boldsymbol{m}(y))\Big| \leq B_1 + B_2$$

where

$$B_1 = \Big|var(\boldsymbol{\beta}^{\tau}\boldsymbol{m}(y)) - \sum_h \delta_h \left(\boldsymbol{\beta}^{\tau}\mu_h\right)^2\Big| \tag{25}$$

$$B_2 = \Big|\frac{1}{H}\sum_h (\boldsymbol{\beta}^{\tau}\mu_h)^2 - \sum_h \delta_h \left(\boldsymbol{\beta}^{\tau}\mu_h\right)^2\Big|. \tag{26}$$

Recall the definition of the random intervals $S_h, h = 1, 2, \cdots, H$ and random variable $\delta_h = \delta_h(\omega) = \int_{y \in S_h(\omega)} f(y)dy$. Define the event $E(\epsilon) = \left\{ \omega \;\middle|\; |\delta_h - \frac{1}{H}| > \epsilon, \forall h \right\}$. For any $\omega \in E(\epsilon)^c$, one has

$$
\begin{aligned}
B_1 =& \sum_h \delta_h(\omega) var(\boldsymbol{\beta}^\tau \boldsymbol{m}(y) | y \in S_h(\omega)) \\
\leq& (\frac{1}{H} + \epsilon) \sum_h var(\boldsymbol{\beta}^\tau \boldsymbol{m}(y) | y \in S_h(\omega)) \qquad (27) \\
\leq& (1 + H\epsilon) \frac{\mathfrak{a}}{H^\kappa} var(\boldsymbol{\beta}^\tau \boldsymbol{m}(y)) \qquad (28)
\end{aligned}
$$

where inequality (27) follows from $\delta_h(\omega) \leq \frac{1}{H} + \epsilon$ and the inequality (28) follows from sliced stable condition (5) , and

$$
\begin{aligned}
B_2 \leq& \epsilon \sum_h (\boldsymbol{\beta}^\tau \mu_h)^2 = \sum_h \frac{\epsilon}{\delta_h} \delta_h (\boldsymbol{\beta}^\tau \mu_h)^2 \\
\leq& \frac{H\epsilon}{1 - H\epsilon} \sum_h \delta_h (\boldsymbol{\beta}^\tau \mu_h)^2 \qquad (29)
\end{aligned}
$$

where inequality (29) follows from $\delta_h \geq \frac{1}{H} - \epsilon$.

From (28), one then has

$$
\sum_h \delta_h (\boldsymbol{\beta}^\tau \mu_h)^2 \leq (1 + (1 + H\epsilon) \frac{\mathfrak{a}}{H^\kappa}) var(\boldsymbol{\beta}^\tau \boldsymbol{m}(y)) \qquad (30)
$$

and from (29), one then has

$$
B_2 \leq \frac{H\epsilon}{1 - H\epsilon} (1 + (1 + H\epsilon)) \frac{\mathfrak{a}}{H^\kappa} var(\boldsymbol{\beta}^\tau \boldsymbol{m}(y)). \qquad (31)
$$

So when $E(\epsilon)^c$ occurs, one has

$$
\begin{aligned}
&\left| \frac{1}{H} \sum_h (\boldsymbol{\beta}^\tau \mu_h)^2 - var(\boldsymbol{\beta}^\tau \boldsymbol{m}(y)) \right| \\
&\leq (1 + H\epsilon) \frac{\mathfrak{a}}{H^\kappa} var(\boldsymbol{\beta}^\tau \boldsymbol{m}(y)) + \frac{H\epsilon}{1 - H\epsilon} (1 + (1 + H\epsilon)) \frac{\mathfrak{a}}{H^\kappa} var(\boldsymbol{\beta}^\tau \boldsymbol{m}(y)). \qquad (32)
\end{aligned}
$$

Consequently, for some positive constants C' and C, one has

$$
\left| \frac{1}{H} \sum_h (\boldsymbol{\beta}^\tau \mu_h)^2 - var(\boldsymbol{\beta}^\tau \boldsymbol{m}(y)) \right| \leq \frac{C'}{H^\kappa} var(\boldsymbol{\beta}^\tau \boldsymbol{m}(y)) \qquad (33)
$$

with probability at least

$$1 - CH^2\sqrt{Hc+1}\exp\left(-(Hc+1)\frac{\epsilon^2}{32}\right) \tag{34}$$

by Lemma 14. In particular, since $var(\boldsymbol{\beta}^\tau \boldsymbol{m}(y))$ is bounded, one has

$$A_1 = O_P(\frac{1}{H^\kappa})$$

□

REMARK 5. From (33) , one has the following two inequalities

$$\frac{1}{H}\sum_h (\boldsymbol{\beta}^\tau \mu_h)^2 \leq \left(1 + \frac{C'}{H^\kappa}\right) var\left(\boldsymbol{\beta}^\tau \boldsymbol{m}(y)\right) \tag{35}$$

and

$$\frac{1}{H}\sum_h |\left(\boldsymbol{\beta}^\tau \mu_h\right)| \leq \left(\left(1 + \frac{C'}{H^\kappa}\right) var\left(\boldsymbol{\beta}^\tau \boldsymbol{m}(y)\right)\right)^{1/2}. \tag{36}$$

hold with probability at least

$$1 - CH^2\sqrt{Hc+1}\exp\left(-(Hc+1)\frac{\epsilon^2}{32}\right). \tag{37}$$

In particular, $\frac{1}{H}\sum_h (\boldsymbol{\beta}^\tau \mu_h)^2$ and $\frac{1}{H}\sum_h \left|\left(\boldsymbol{\beta}^\tau \mu_h\right)\right|$ are bounded by $O_P(1)$.

LEMMA 5.

$$A_2 \leq O_P\left(\frac{\sqrt{d}H^2}{\sqrt{n}}\right)$$

PROOF. From Corollary 1 in Section 9 , one needs to treat the H-th slice separately. Note that

$$A_2 \leq A_2' + \frac{1}{H}\left|\left(\boldsymbol{\beta}^\tau \overline{\boldsymbol{m}}_{H,\cdot}\right)^2 - \left(\boldsymbol{\beta}^\tau \mu_H\right)^2\right|.$$

where

$$A_2' \triangleq \frac{1}{H}\sum_{h=1}^{H-1}\left|\left(\boldsymbol{\beta}^\tau \overline{\boldsymbol{m}}_{h,\cdot}\right)^2 - \left(\boldsymbol{\beta}^\tau \mu_h\right)^2\right|.$$

Let $\overline{\boldsymbol{m}}_{h,1:(c-1)} = \frac{1}{c-1}\sum_{i=1}^{c-1}\boldsymbol{m}_{h,i}$, for $h = 1, ..., H-1$, then

$$\begin{aligned} A_2' \leq & \frac{1}{H}\sum_{h=1}^{H-1}\left|\boldsymbol{\beta}^\tau \overline{\boldsymbol{m}}^2_{h,1:(c-1)} - (\boldsymbol{\beta}^\tau\mu_h)^2\right| + \frac{2}{Hc}\sum_{h=1}^{H-1}(\boldsymbol{\beta}^\tau\mu_h)^2 \\ & + \frac{2(c-1)}{c}\sqrt{\frac{1}{H}\sum_{h=1}^{H-1}\boldsymbol{\beta}^\tau\overline{\boldsymbol{m}}^2_{h,1:(c-1)}}\sqrt{\frac{1}{Hc^2}\sum_{h=1}^{H-1}(\boldsymbol{\beta}^\tau\boldsymbol{m}_{h,c})^2} \\ & + \frac{1}{Hc^2}\sum_{h=1}^{H-1}(\boldsymbol{\beta}^\tau\boldsymbol{m}_{h,c})^2 \end{aligned} \tag{38}$$

so

$$A_2 \leq I + II + III + IV$$

where

$$I = \frac{1}{H}\sum_{h=1}^{H-1}\left|\left(\boldsymbol{\beta}^\tau\overline{\boldsymbol{m}}_{h,1:(c-1)}\right)^2 - (\boldsymbol{\beta}^\tau\mu_h)^2\right| + \frac{1}{H}\left|\left(\boldsymbol{\beta}^\tau\overline{\boldsymbol{m}}_{H,\cdot}\right)^2 - (\boldsymbol{\beta}^\tau\mu_H)^2\right|, \tag{39}$$

$$II = \frac{2}{Hc}\sum_{h=1}^{H}(\boldsymbol{\beta}^\tau\mu_h)^2, \tag{40}$$

$$III = \frac{2(c-1)}{c}\sqrt{\frac{1}{H}\sum_{h=1}^{H-1}\boldsymbol{\beta}^\tau\overline{\boldsymbol{m}}^2_{h,1:(c-1)} + \frac{1}{H}\left(\boldsymbol{\beta}^\tau\overline{\boldsymbol{m}}_{H,\cdot}\right)^2}\sqrt{IV}, \tag{41}$$

$$IV = \frac{1}{Hc^2}\sum_{h=1}^{H}(\boldsymbol{\beta}^\tau\boldsymbol{m}_{h,c})^2. \tag{42}$$

Let η be the maximal value in $\left|\boldsymbol{\beta}^\tau\boldsymbol{m}_{h,1:(c-1)} - \boldsymbol{\beta}^\tau\mu_h\right|, h = 1, \cdots, H-1$ and $\left|\boldsymbol{\beta}^\tau\boldsymbol{m}_{h,\cdot} - \boldsymbol{\beta}^\tau\mu_h\right|$. According to Lemma 16, $\boldsymbol{m}_{h,i}, i = 1, .., c-1$ can be treated as i.i.d. random samples of $\boldsymbol{m}|_{y\in S_h}$ for $h = 1, \cdots, H-1$, $\boldsymbol{m}_{H,i}, i = 1, \cdots, c$ can be treated as i.i.d. random samples of $\boldsymbol{m}|_{y\in S_H}$ and $var((\boldsymbol{\beta}^\tau\boldsymbol{m})\,|y \in S_h) \leq CHvar(\boldsymbol{\beta}^\tau\boldsymbol{m}),\ \forall h = 1, .., H$ for some positive constant C, one has

that for any constant $a > 0$,

$$\mathbb{P}(\eta > \frac{a\sqrt{d}H}{\sqrt{c}}) \leq \sum_{h=1}^{H-1} \mathbb{P}\left(\left|\boldsymbol{\beta}^\tau \boldsymbol{m}_{h,1:(c-1)} - \boldsymbol{\beta}^\tau \mu_h\right| > \frac{a\sqrt{d}H}{\sqrt{c}}\right) + \mathbb{P}\left(\left|\boldsymbol{\beta}^\tau \boldsymbol{m}_{H,\cdot} - \boldsymbol{\beta}^\tau \mu_H\right| > \frac{a\sqrt{d}H}{\sqrt{c}}\right) \tag{43}$$

$$\leq \sum_{h=1}^{H-1} \frac{\mathbb{E}[\left(\frac{1}{c-1}\sum_{i=1}^{c-1}(\boldsymbol{\beta}^\tau \boldsymbol{m}_{h,i} - \boldsymbol{\beta}^\tau \mu_h)\right)^2]}{a^2 H^2 d/c} + \frac{\mathbb{E}[\left(\frac{1}{c-1}\sum_{i=1}^{c}(\boldsymbol{\beta}^\tau \boldsymbol{m}_{H,i} - \boldsymbol{\beta}^\tau \mu_H)\right)^2]}{a^2 H^2 d/c} \tag{44}$$

$$\leq \frac{2Cvar((\boldsymbol{\beta}^\tau \boldsymbol{m}))}{a^2}.$$

In particular, $\eta \leq O_P(\frac{\sqrt{d}H}{\sqrt{c}})$.

For I. From (36), one has

$$I \leq 2\eta^2 + \frac{\eta}{H}\sum_h \left|\boldsymbol{\beta}^\tau \mu_h\right| \leq O_P(\frac{\sqrt{d}H}{\sqrt{c}}). \tag{45}$$

For II. From (35), one has

$$II \leq O_P(\frac{1}{c}).$$

For VI. One has

$$VI \leq \frac{1}{Hc^2}\sum_{i=1}^{n}(\boldsymbol{\beta}^\tau \boldsymbol{m}_i)^2 \leq O_P\left(\frac{d}{c}\right) \tag{46}$$

since $\boldsymbol{\beta}^\tau \boldsymbol{m}_i$ are i.i.d samples of $\boldsymbol{\beta}^\tau \boldsymbol{m}(y)$.

For III. From (35) and the estimation of I, one has

$$\frac{1}{H}\sum_{h=1}^{H-1} \boldsymbol{\beta}^\tau \overline{\boldsymbol{m}}^2_{h,1:(c-1)} + \frac{1}{H}(\boldsymbol{\beta}^\tau \overline{\boldsymbol{m}}_{H,\cdot})^2 \leq I + \frac{1}{H}\sum_{h=1}^{H}(\boldsymbol{\beta}^\tau \mu_h)^2.$$

is bounded. So, $III = O_P\left(\frac{\sqrt{d}}{\sqrt{c}}\right)$. To sum up, one has

$$A_2 \leq O_P(\frac{\sqrt{d}H}{\sqrt{c}}) \leq O_P(\frac{\sqrt{d}H^2}{\sqrt{n}})$$

□

7.2.2. *Proof of Lemma 2* Note that

$$\widetilde{\mathbf{\Lambda}}_z - \mathbf{\Lambda}_p = \widetilde{\mathbf{\Lambda}}_m - \mathbf{\Lambda}_p + \mathcal{W}_2 + \mathcal{C}_2 + \mathcal{C}_2^\tau,$$

where

$$\mathcal{W}_2 = \frac{1}{H}\sum_h \overline{\boldsymbol{v}}_{h,\cdot}\overline{\boldsymbol{v}}_{h,\cdot}^\tau, \quad \text{and} \quad \mathcal{C}_2 = \frac{1}{H}\sum_h \overline{\boldsymbol{m}}_{h,\cdot}\overline{\boldsymbol{v}}_{h,\cdot}^\tau.$$

To prove Lemma 2, it suffices to show that

$$\|\mathcal{W}_2\|_2 \leq O_p(\frac{H^2 d}{n}) \text{ and } \|\mathcal{C}_2\|_2 \leq O_P(\sqrt{\frac{H^2 d}{n}}).$$

For $\mathcal{W}_2$, since for any unit vector $\boldsymbol{\beta}$, one has

$$\begin{aligned}
\boldsymbol{\beta}^\tau \mathcal{W}_2 \boldsymbol{\beta} &= \frac{1}{H}\sum_h \boldsymbol{\beta}^\tau \overline{\boldsymbol{v}}_{h,\cdot}\overline{\boldsymbol{v}}_{h,\cdot}^\tau \boldsymbol{\beta} \\
&= \frac{1}{H}\sum_h \left(\frac{c-1}{c}\frac{1}{c-1}\sum_{i=1}^{c-1}\boldsymbol{\beta}^\tau \boldsymbol{v}_{h,i} + \frac{1}{c}\boldsymbol{\beta}^\tau \boldsymbol{v}_{h,c}\right)^2 \\
&\leq \frac{2}{H}\sum_h \left(\frac{1}{c-1}\sum_{i=1}^{c-1}\boldsymbol{\beta}^\tau \boldsymbol{v}_{h,i}\right)^2 + \frac{2}{Hc^2}\sum_{i=1}^{n}(\boldsymbol{\beta}^\tau \boldsymbol{v}_i)^2
\end{aligned}$$

We apply the simple fact $(a+b)^2 \leq 2(a^2+b^2)$ in the inequality.

By definition, we know that $\mathbf{\Sigma}_{\boldsymbol{v}}$, the covariance matrix of $\boldsymbol{v}$, has bounded largest eigenvalue and $\boldsymbol{v}|_{y\in S_h}$ has mean 0. From Lemma 12, we know that $\boldsymbol{\beta}^\tau \boldsymbol{v}_{h,i}$ $1 \leq i \leq c-1$ can be treated as i.i.d. sample from $\boldsymbol{\beta}^\tau \boldsymbol{v}|_{y\in S_h}$ and from Lemma 13 we know the largest eigenvalue of the covariance matrix $\mathbf{\Sigma}_{\boldsymbol{v}|_{y\in S_h}}$ of $\boldsymbol{v}|_{y\in S_h}$ is bounded by $CH\lambda_{max}(\mathbf{\Sigma}_{\boldsymbol{v}})$ for some positive constant C. Write $\boldsymbol{v}_{h,i}$ as $A_{\boldsymbol{v}|_{y\in S_h}}\boldsymbol{\alpha}_{h,i}$, where $A_{\boldsymbol{v}|_{y\in S_h}}$ is a $p \times d$ matrix such that $A_{\boldsymbol{v}|_{y\in S_h}}A_{\boldsymbol{v}|_{y\in S_h}}^\tau = \mathbf{\Sigma}_{\boldsymbol{v}|_{y\in S_h}}$ and $\boldsymbol{\alpha}_{h,i}$ are i.i.d. random samples of a random variable $\boldsymbol{\alpha}$ with an identity covariance matrix. It is easy to verify that $\|\boldsymbol{v}|_{y\in S_h}\|_2^2 \leq CH\|\boldsymbol{\alpha}\|_2^2 = O_P(dH)$.

For any unit vector $\boldsymbol{\beta}$,

$$\boldsymbol{\beta}^\tau \mathcal{W}_2 \boldsymbol{\beta} \leq \frac{2}{H} \sum_h O_P\left(\frac{dH}{c-1}\right) + O_P(\frac{d}{c}) = O_P\left(\frac{dH}{c}\right). \tag{47}$$

In particular, we have

$$\|\mathcal{W}_2\|_2 \leq O_P(\frac{dH}{c}) = O_P(\frac{dH^2}{n}). \tag{48}$$

For $\mathcal{C}_2$, by Cauchy inequality, we have

$$\begin{aligned} \|\mathcal{C}_2\|_2^2 \leq & \|\frac{1}{H}\sum \overline{\boldsymbol{m}}_{h\cdot}\overline{\boldsymbol{m}}_{h\cdot}^\tau\|_2 \cdot \|\frac{1}{H}\sum \overline{\boldsymbol{v}}_{h\cdot}\overline{\boldsymbol{v}}_{h\cdot}^\tau\|_2 \\ \leq & (\|\boldsymbol{\Lambda}_p\|_2 + O_P(\frac{dH^2}{\sqrt{n}}) + O_P(\frac{1}{H^\kappa}))O_p(\frac{H^2 d}{n}) \qquad (49) \\ = & O_p(\frac{H^2 d}{n}). \end{aligned}$$

where inequality (49) follows from Lemma 1 and (48).

Consequently,

$$\begin{aligned} \|\frac{1}{H}\sum_h \overline{\boldsymbol{z}}_{h,\cdot}\overline{\boldsymbol{z}}_{h,\cdot}^\tau - \boldsymbol{\Lambda}_p\|_2 \leq & \|\mathcal{L}\|_2 + \|\mathcal{W}_2\|_2 + \|\mathcal{C}_2\|_2 + \|\mathcal{C}_2^\tau\|_2 \\ \leq & O_P(\frac{\sqrt{d}H^2}{\sqrt{n}} + \frac{1}{H^\kappa} + \frac{H^2 d}{n} + \sqrt{\frac{H^2 d}{n}}) \qquad (50) \\ \leq & O_P(\frac{\sqrt{d}H^2}{\sqrt{n}} + \frac{1}{H^\kappa}). \end{aligned}$$

□

7.2.3. *Proof of Lemma 3* Note that $\widetilde{\boldsymbol{\Lambda}}_{\boldsymbol{p}} - \boldsymbol{\Lambda}_p = \widetilde{\boldsymbol{\Lambda}}_{\boldsymbol{z}} - \boldsymbol{\Lambda}_p + \mathcal{W}_1 + \mathcal{C}_1 + \mathcal{C}_1^\tau$ where

$$\mathcal{W}_1 = \frac{1}{H}\sum_h \overline{\boldsymbol{w}}_{h,\cdot}\overline{\boldsymbol{w}}_{h,\cdot}^\tau, \quad \text{and} \quad \mathcal{C}_1 = \frac{1}{H}\sum_h \overline{\boldsymbol{z}}_{h,\cdot}\overline{\boldsymbol{w}}_{h,\cdot}^\tau.$$

To prove Lemma 3, it suffices to show that

$$\|\mathcal{W}_1\|_2 \leq O_p(\frac{H^2 p}{n}) \text{ and } \|\mathcal{C}_1\|_2 \leq O_P(\sqrt{\frac{H^2 p}{n}})$$

For $\mathcal{W}_1$, similar to the proof of Lemma 2, we have

$$\boldsymbol{\beta}^\tau \mathcal{W}_1 \boldsymbol{\beta} \leq \frac{2}{H}\sum_h \left(\frac{1}{c-1}\sum_{i=1}^{c-1} \boldsymbol{\beta}^\tau \boldsymbol{w}_{h,i}\right)^2 + \frac{2}{Hc^2}\sum_{i=1}^{n} (\boldsymbol{\beta}^\tau \boldsymbol{w}_i)^2 .$$

Similar argument leads to

$$\boldsymbol{\beta}^\tau \mathcal{W}_2 \boldsymbol{\beta} \leq \frac{2}{H}\sum_h O_P\left(\frac{Hp}{c-1}\right) + O_P(\frac{p}{c}) = O_P\left(\frac{pH^2}{n}\right). \tag{51}$$

For $\mathcal{C}_1$, according to Lemma 2 , one has

$$\|\frac{1}{H}\sum_h \overline{\boldsymbol{z}}_{h,\cdot}\overline{\boldsymbol{z}}_{h,\cdot}^\tau\|_2 \leq \|\boldsymbol{\Lambda}_p\|_2 + O_P\left(\frac{\sqrt{d}H^2}{\sqrt{n}} + \frac{1}{H^\kappa}\right)$$

and

$$\begin{aligned}\|\mathcal{C}_1\|_2^2 &\leq \|\frac{1}{H}\sum_h \overline{\boldsymbol{z}}_{h,\cdot}\overline{\boldsymbol{z}}_{h,\cdot}^\tau\|_2 \|\frac{1}{H}\sum_h \overline{\boldsymbol{w}}_{h,\cdot}\overline{\boldsymbol{w}}_{h,\cdot}^\tau\|_2 \\ &\leq \left(\|\boldsymbol{\Lambda}_p\|_2 + O_P\left(\frac{\sqrt{d}H^2}{\sqrt{n}} + \frac{1}{H^\kappa}\right)\right) O_P(\frac{H^2 p}{n}).\end{aligned}$$

Since $\|\boldsymbol{\Lambda}_p\|_2$ is bounded, we know $\|\mathcal{C}_1\|_2 \leq O_P(\sqrt{\frac{H^2 p}{n}})$.

Consequently,

$$\begin{aligned}&\|\widetilde{\boldsymbol{\Lambda}}_{\boldsymbol{p}} - \boldsymbol{\Lambda}_p\|_2 \leq \|\widetilde{\boldsymbol{\Lambda}}_{\boldsymbol{z}} - \boldsymbol{\Lambda}_p\|_2 + \|\mathcal{W}_1\|_2 + \|\mathcal{C}_1\|_2 + \|\mathcal{C}_1^\tau\|_2 \\ \leq& O_P\left(\frac{\sqrt{d}H^2}{\sqrt{n}} + \frac{1}{H^\kappa} + \frac{H^2 p}{n} + \sqrt{\frac{H^2 p}{n}}\right) \\ \leq& O_P\left(\frac{1}{H^\kappa} + \frac{H^2 p}{n} + \sqrt{\frac{H^2 p}{n}}\right).\end{aligned} \tag{52}$$

□

7.3. *Proof of Theorem 2.* Theorem 2 is a direct corollary of Theorem 1 and Lemma 16. In fact, we have:

$$\begin{aligned}&\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{X}}^{-1}\widehat{\boldsymbol{\Lambda}}_p - \boldsymbol{\Sigma}_X^{-1}\boldsymbol{\Lambda}_p\|_2 \\ &\leq \|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{X}}^{-1} - \boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1}\|_2 \|\widehat{\boldsymbol{\Lambda}}_p\|_2 + \|\boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1}\|_2 \|\widehat{\boldsymbol{\Lambda}}_{\boldsymbol{p}} - \boldsymbol{\Lambda}_p\|_2,\end{aligned}$$

which $\to 0$ if $\rho = \lim_{n\to\infty}\frac{p}{n} = 0$. □

7.4. *Proof of Theorem 3*

7.4.1. *Proof of Theorem 3 (i)* This part is almost the same as the proof of Theorem 1, except that the standard normal assumption on $\boldsymbol{x}$ will provide us with a sharper bound of $\mathcal{W}_1$ and $\mathcal{C}_1$. Since $\boldsymbol{x}$ is normal and $y \perp\!\!\!\perp \boldsymbol{x}|\boldsymbol{P}_{\mathcal{S}}\boldsymbol{x}$, one knows that $\boldsymbol{P}_{\mathcal{S}^\perp}\boldsymbol{x} \perp\!\!\!\perp \boldsymbol{P}_{\mathcal{S}}\boldsymbol{x}$ and $y \perp\!\!\!\perp \boldsymbol{P}_{\mathcal{S}^\perp}\boldsymbol{x}$.

For $\mathcal{W}_1$, since $\boldsymbol{w} = \boldsymbol{P}_{\mathcal{S}^\perp}\boldsymbol{x}$ is normal and independent of y, there exists a normal random variable $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{I})$ such that $\boldsymbol{w} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\epsilon}$ where $\Sigma = cov(\boldsymbol{w})$. In particular, one may write $\boldsymbol{w}_{h,i} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\epsilon}_{h,i}$ and $\overline{\boldsymbol{w}}_{h,\cdot} = \frac{1}{c}\sum_i \boldsymbol{w}_{h,i} = \frac{1}{\sqrt{c}}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\epsilon}_h$, $h = 1, \cdots, H$ where $\boldsymbol{\epsilon}_{h,i}$ are i.i.d random samples of standard normal distribution and $\boldsymbol{\epsilon}_h = \frac{1}{\sqrt{c}}\sum_i \boldsymbol{\epsilon}_{h,i}$, $h = 1, \cdots, H$ are i.i.d random sample of standard normal distributions. So,

$$\mathcal{W}_1 = \frac{1}{Hc}\boldsymbol{\Sigma}^{1/2}\boldsymbol{E}_{p\times H}\boldsymbol{E}_{p\times H}^\tau\boldsymbol{\Sigma}^{1/2},$$

where $\boldsymbol{E}_{p\times H} = (\boldsymbol{\epsilon}_h)_{h=1,\cdots,H}$ is a $p \times H$ matrix with all the entries being an i.i.d. random sample generated from standard normal distributions. Combining with Corollary 4, we know

$$\|\mathcal{W}_1\|_2 \le C\left(\sqrt{\frac{p}{n}} + \sqrt{\frac{H}{n}}\right)^2 \le O_P\left(\frac{p}{n}\right).$$

For $\mathcal{C}_1$, from Lemma 2, one knows

$$\|\frac{1}{H}\sum_h \overline{\boldsymbol{z}}_{h,\cdot}\overline{\boldsymbol{z}}_{h,\cdot}^\tau\|_2 \le \|\boldsymbol{\Lambda}_p\|_2 + O_P\left(\frac{\sqrt{d}H^2}{\sqrt{n}} + \frac{1}{H^\kappa}\right).$$

By Cauchy inequality, one has

$$\begin{aligned}\|\mathcal{C}_1\|_2^2 &\le \|\frac{1}{H}\sum_h \overline{\boldsymbol{z}}_{h,\cdot}\overline{\boldsymbol{z}}_{h,\cdot}^\tau\|_2\|\frac{1}{H}\sum_h \overline{\boldsymbol{w}}_{h,\cdot}\overline{\boldsymbol{w}}_{h,\cdot}^\tau\|_2 \\ &\le \left(\|\boldsymbol{\Lambda}_p\|_2 + O_P\left(\frac{\sqrt{d}H^2}{\sqrt{n}} + \frac{1}{H^\kappa}\right)\right)O_P(\frac{p}{n}) \\ &\le O_P\left(\frac{p}{n}\right)\end{aligned}$$

Now the (52) in Lemma 3 can be replaced by

$$\begin{aligned}&\|\widetilde{\boldsymbol{\Lambda}}_{\boldsymbol{p}} - \boldsymbol{\Lambda}_p\|_2 \le \|\widetilde{\boldsymbol{\Lambda}}_{\boldsymbol{z}} - \boldsymbol{\Lambda}_p\|_2 + \|\mathcal{W}_1\|_2 + \|\mathcal{C}_1\|_2 + \|\mathcal{C}_1^\tau\|_2 \\ \le& O_P(\frac{\sqrt{d}H^2}{\sqrt{n}} + \frac{1}{H^\kappa} + \frac{p}{n} + \sqrt{\frac{p}{n}}) \\ \le& O_P(\frac{1}{H^\kappa} + \frac{p}{n} + \sqrt{\frac{p}{n}}).\end{aligned} \tag{53}$$

Note that , the difference between $\widetilde{\mathbf{\Lambda}}_{\boldsymbol{p}}$ and $\widehat{\mathbf{\Lambda}}_{\boldsymbol{p}}$ is bounded by $O_P(\frac{1}{H}+\frac{p}{n})$, so

$$\|\widehat{\mathbf{\Lambda}}_{\boldsymbol{p}}-\mathbf{\Lambda}_p\|_2 \le O_P(\frac{1}{H^{1\wedge\kappa}}+\frac{p}{n}+\sqrt{\frac{p}{n}})$$

In particular, if $H,n\to\infty$ and $\rho=\lim\frac{p}{n}\in(0,\infty)$, one knows that $\|\widehat{\mathbf{\Lambda}}_p-\mathbf{\Lambda}_p\|_2$ is dominated by $\rho\vee\sqrt{\rho}$ as a function of ρ. □

7.4.2. *Proof of Theorem 3 (ii)* The proof is similar to the proof of Theorem 2 in Johnstone and Lu [2009] but more challenge. Since we are working on single index model with $\boldsymbol{x}$ is standard normal, the decomposition (15) becomes

$$\boldsymbol{x}=\boldsymbol{z}+\boldsymbol{w}$$

where $\boldsymbol{z}=P_{\boldsymbol{\beta}}\boldsymbol{x}=\boldsymbol{\beta}z(y)$ for some scalar function $z(y)$ and $\boldsymbol{w}=P_{\boldsymbol{\beta}^\perp}\boldsymbol{x}$ are independent normal random variables. Now $\overline{\boldsymbol{x}}_{h,\cdot}$, the sample mean in h-th slice, has the following decomposition

$$\overline{\boldsymbol{x}}_{h,\cdot}=\overline{\boldsymbol{z}}_{h,\cdot}+\overline{\boldsymbol{w}}_{h,\cdot}.$$

Let $\mathbf{X}=(\overline{\boldsymbol{x}}_{1,\cdot},\overline{\boldsymbol{x}}_{2,\cdot},...,\overline{\boldsymbol{x}}_{H,\cdot})$ be the matrix consists of the sample means of slices. $\mathbf{Z}$ and $\mathbf{W}$ are defined similarly. Then

$$\widehat{\mathbf{\Lambda}}_p=\frac{1}{H-1}\sum_{h=1}^{H}(\overline{\boldsymbol{x}}_{h,\cdot}-\overline{\overline{\boldsymbol{x}}})(\overline{\boldsymbol{x}}_{h,\cdot}-\overline{\overline{\boldsymbol{x}}})^\tau = D+B,$$

where

$$D=\frac{1}{H}\mathbf{Z}\mathbf{Z}^\tau+\frac{1}{H}\mathbf{W}\mathbf{W}^\tau+\frac{1}{H(H-1)}\mathbf{X}\mathbf{X}^\tau-\frac{H}{H-1}\overline{\overline{\boldsymbol{x}}\overline{\boldsymbol{x}}}^\tau \tag{54}$$

$$B=\frac{1}{H}\mathbf{Z}\mathbf{W}^\tau+\frac{1}{H}\mathbf{W}\mathbf{Z}^\tau. \tag{55}$$

Let $\mathbf{\Sigma}=var(\boldsymbol{w})$ then $\mathbf{\Sigma}=\mathbf{I}-\boldsymbol{\beta}\boldsymbol{\beta}^\tau$ and $\mathbf{\Sigma}^{1/2}=\mathbf{\Sigma}$. One may let $\boldsymbol{w}=\mathbf{\Sigma}^{1/2}\boldsymbol{\epsilon}$ for some standard normal random variables $\boldsymbol{\epsilon}\sim N(0,\mathbf{I}_p)$ which is independent of $\boldsymbol{z}$ and y. One then has that $\sqrt{c}\overline{\boldsymbol{\epsilon}}_{h,\cdot}=\frac{1}{\sqrt{c}}\sum_{i=1}^{c}\boldsymbol{\epsilon}_{h,i}$ identically distributed as standard normal and $\overline{\boldsymbol{w}}_{h,\cdot}=\mathbf{\Sigma}^{1/2}\overline{\boldsymbol{\epsilon}}_{h,\cdot}$. i.e., $\mathbf{W}=\frac{1}{\sqrt{c}}\mathbf{\Sigma}^{1/2}\mathbf{E}$ where $\mathbf{E}=\sqrt{c}(\overline{\boldsymbol{\epsilon}}_{1,\cdot},\overline{\boldsymbol{\epsilon}}_{2,\cdot},...,\overline{\boldsymbol{\epsilon}}_{H,\cdot})$ is a $p\times H$ matrix with entries are i.i.d standard normal. Similarly, since $\boldsymbol{z}=\boldsymbol{\beta}z(y)$, one has $\mathbf{Z}=\boldsymbol{\beta}(\overline{z}_{1,\cdot},\overline{z}_{2,\cdot},...,\overline{z}_{H,\cdot})$. To ease notation, let $\boldsymbol{\theta}^\tau=(\overline{z}_{1,\cdot},\overline{z}_{2,\cdot},...,\overline{z}_{H,\cdot})$, then

$$\begin{aligned} D&=\frac{1}{H}\|\boldsymbol{\theta}\|^2\boldsymbol{\beta}\boldsymbol{\beta}^\tau+\frac{1}{n}\mathbf{\Sigma}^{1/2}\boldsymbol{E}\boldsymbol{E}^\tau\mathbf{\Sigma}^{1/2}-\frac{H}{H-1}\overline{\overline{\boldsymbol{x}}\overline{\boldsymbol{x}}}^\tau+\frac{1}{H(H-1)}\boldsymbol{X}\boldsymbol{X}^\tau \\ B&=\boldsymbol{\beta}\boldsymbol{u}^\tau+\boldsymbol{u}\boldsymbol{\beta}^\tau \text{ where } \quad \boldsymbol{u}=\frac{1}{H\sqrt{c}}\mathbf{\Sigma}^{1/2}\boldsymbol{E}\boldsymbol{\theta}. \end{aligned} \tag{56}$$

Let $0 < \alpha < arctan(\frac{1}{16})$ and

$$N_\alpha = \left\{ \boldsymbol{x} \in \mathbb{R}^p : \angle(\boldsymbol{x}, \boldsymbol{\beta}) \leq \alpha \text{ and } \|x\| = 1 \right\} \tag{57}$$

be the set of unit vectors making angle at most α where $\angle(\boldsymbol{x}, \boldsymbol{y})$ is the angle between the vectors $\boldsymbol{x}$ and $\boldsymbol{y}$.

To prove the second part of Theorem 3, we need the following lemmas.

LEMMA 6. *Let $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}_-$ be the principal eigenvector of $S_+ \triangleq D + B$ and $S_- \triangleq D - B$ respectively. There exists a positive constant $\omega(\alpha)$ such that for any $\widehat{\boldsymbol{\beta}} \in N_\alpha$, i.e., $\angle\left(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}\right) \leq \alpha$, one has*

$$\angle\left(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\beta}}_-\right) \geq \frac{1}{3}\omega(\alpha) \tag{58}$$

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

Assuming Lemma 6, note that S_+ and S_- have the same distribution (viewed as functions of random terms $\boldsymbol{E}$ and θ):

$$S_-(\boldsymbol{E}, \theta) = S_+(-\boldsymbol{E}, \theta).$$

Let $\mathcal{A}_\alpha$ denote the event $\left\{\angle\left(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}\right) \leq \alpha\right\} \cup \left\{\angle\left(\widehat{\boldsymbol{\beta}}_-, \boldsymbol{\beta}\right) \leq \alpha\right\}$, then

$$\begin{aligned}
\mathbb{E}[\angle\left(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}\right)] &\geq \mathbb{E}[\angle\left(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}\right), \mathcal{A}_\alpha^c] + \mathbb{E}[\angle\left(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}\right), \mathcal{A}_\alpha] \\
&\geq \mathbb{E}[\angle\left(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}\right), \mathcal{A}_\alpha^c] + \frac{1}{2}\mathbb{E}[\angle\left(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\beta}}_-\right), \mathcal{A}_\alpha] \\
&\geq \min\{\alpha, \frac{\omega(\alpha)}{6}\} > 0.
\end{aligned}$$

□

7.5. *Proof of Lemma 6.* We need the following lemmas.

LEMMA 7. *Recall that $\boldsymbol{u} = \frac{1}{H\sqrt{c}}\boldsymbol{\Sigma}^{1/2}\boldsymbol{E}\boldsymbol{\theta}$ defined as in (56), then there exist positive constants C_1 and C_2 such that*

$$0 < C_1 \leq \|\boldsymbol{u}\|_2 \leq C_2$$

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

LEMMA 8. *Assuming conditions in Theorem 3, let B and N_α be defined as in (56) and(57) respectively where $0 < \alpha < arctan(\frac{1}{16})$.*

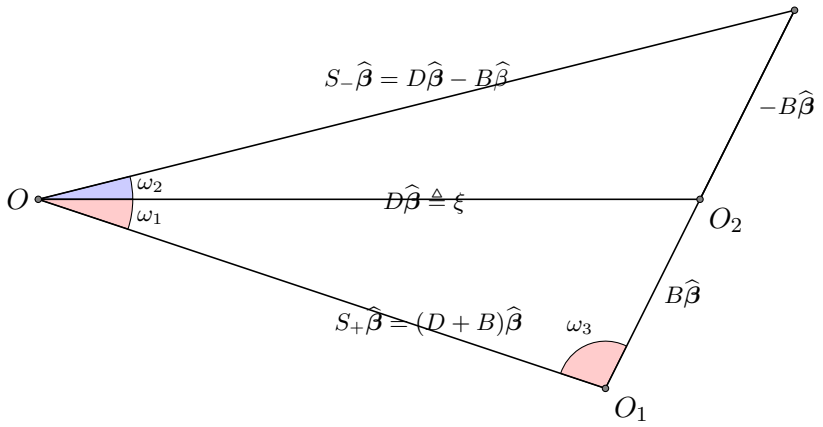

Fig 4: An illustrated graph

i) There exists positive constant C_1 such that for any $\boldsymbol{x} \in N_\alpha$, one has $\|B\boldsymbol{x}\| \geq C_1$ with probability converging to one as $n \to \infty$;

ii) For any $\boldsymbol{x} \in N_\alpha$, one has $\left|\cos \angle(\boldsymbol{x}, B\boldsymbol{x})\right| \leq 4\alpha$ with probability converging to one as $n \to \infty$.

The following lemma is borrowed from Johnstone and Lu [2004].

LEMMA 9. *Let $\boldsymbol{\xi}$ be a principal eigenvector of a non-zero symmetric matrix M. For any $\boldsymbol{\eta} \neq 0$,*

$$\angle(\boldsymbol{\eta}, M\boldsymbol{\eta}) \leq 3\angle(\boldsymbol{\eta}, \boldsymbol{\xi}).$$

The proof of Lemma 6 is made plausible by reference to the Figure 4.

Since

$$\sin\left(\angle\left(\widehat{\boldsymbol{\beta}}, S_-\widehat{\boldsymbol{\beta}}\right)\right) = \sin(\omega_1 + \omega_2) = \sin(\pi - \omega_1 - \omega_2) \tag{59}$$

$$\geq \min\left\{ \sin(\omega_1), \sin(\omega_3) \right\}, \tag{60}$$

one only needs to prove that there exists a positive small constant $\omega(\alpha)$ ($< \frac{\pi}{2}$) such that $\sin(\omega_1), \sin(\omega_2) \geq \sin(\omega(\alpha))$. In fact, if such $\omega(\alpha)$ exists, one may choose $\boldsymbol{M} = S_-, \xi = \widehat{\boldsymbol{\beta}}_-$ in Lemma 9 and get

$$\angle\left(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\beta}}_-\right) \geq \frac{1}{3}\angle\left(\widehat{\boldsymbol{\beta}}, S_-\widehat{\boldsymbol{\beta}}\right) \geq \frac{1}{3}\omega(\alpha).$$

For ω_3. From Lemma 8 ii), $\left|\cos\angle(\widehat{\boldsymbol{\beta}}, B\widehat{\boldsymbol{\beta}})\right| \leq 4\alpha$, we know that there exists positive constants $\delta(\alpha)(<\frac{\pi}{2})$ such that $\sin\omega_3 \geq sin\,(\delta(\alpha))$.
For ω_1. Applying the law of sines to the triangle $\triangle(O, O_1, O_2)$, one has

$$\frac{\sin\omega_1}{\|B\widehat{\boldsymbol{\beta}}\|} = \frac{\sin\omega_3}{\|D\widehat{\boldsymbol{\beta}}\|}\left(= \frac{\sin\angle\left(B\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\beta}}\right)}{\|D\widehat{\boldsymbol{\beta}}\|}\right). \tag{61}$$

Note that from Lemma 8 i), there exists a constant $C_1 > 0$ such that $\|B\widehat{\boldsymbol{\beta}}\| > C_1$ and

$$\|D\widehat{\beta}\| \leq \|D\| \leq \frac{1}{H}\|\theta\|^2 + \|\frac{1}{n}EE^\tau\| + \|\frac{H}{H-1}\overline{\overline{XX}}^\tau\| + O(\frac{1}{H}),$$

is bounded by an absolute constant C given $\lim_{n\to\infty}\frac{p}{n} = \rho \neq 0$ and sliced stable condition(or Lemma 2). Then (61) implies

$$\sin\omega_1 = \frac{\|B\widehat{\boldsymbol{\beta}}\|\sin\angle(B\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\beta}})}{\|D\widehat{\boldsymbol{\beta}}\|} \geq \frac{C_1\sin\delta(\alpha)}{C} \geq \sin\omega' > 0$$

where ω' ($< \frac{\pi}{2}$)is an small angle such that the last inequality holds. In particular, we have $\omega_1 \geq \omega'$. Hence

$$\angle(\widehat{\boldsymbol{\beta}}, S_-\widehat{\boldsymbol{\beta}}) \geq \omega' \wedge \delta(\alpha) \triangleq \omega(\alpha)$$

□

7.6. *Proof of Lemma 7 and 8*

7.6.1. *Proof of Lemma 7* *Proof* : In fact, let $\boldsymbol{T}$ be an orthogonal matrix such that $\boldsymbol{T}\boldsymbol{\beta} = (1, 0\cdots, 0)^\tau$ and $\boldsymbol{M} = \boldsymbol{T}\boldsymbol{\beta}\boldsymbol{\beta}^\tau\boldsymbol{T}^\tau$, then

$$\begin{aligned} cH^2\boldsymbol{u}^\tau\boldsymbol{u} &= \boldsymbol{\theta}^\tau\boldsymbol{E}^\tau\boldsymbol{\Sigma}E\boldsymbol{\theta} \\ &= \boldsymbol{\theta}^\tau\boldsymbol{E}^\tau\boldsymbol{E}\boldsymbol{\theta} - \boldsymbol{\theta}^\tau\boldsymbol{E}^\tau\boldsymbol{\beta}\boldsymbol{\beta}^\tau\boldsymbol{E}\boldsymbol{\theta} \\ &= \boldsymbol{\theta}^\tau\boldsymbol{E}^\tau\boldsymbol{T}^\tau\boldsymbol{T}\boldsymbol{E}\boldsymbol{\theta} - \boldsymbol{\theta}^\tau\boldsymbol{E}^\tau\boldsymbol{T}^\tau(\boldsymbol{T}\boldsymbol{\beta})\boldsymbol{\beta}^\tau\boldsymbol{T}^\tau\boldsymbol{T}\boldsymbol{E}\boldsymbol{\theta} \\ &\overset{d.}{=} \boldsymbol{\theta}^\tau\boldsymbol{E}^\tau\boldsymbol{E}\boldsymbol{\theta} - \boldsymbol{\theta}^\tau\boldsymbol{E}^\tau\boldsymbol{M}\boldsymbol{E}\boldsymbol{\theta} \\ &\overset{d.}{=} \frac{p-1}{p}\boldsymbol{\theta}^\tau\boldsymbol{E}^\tau\boldsymbol{E}\boldsymbol{\theta}, \end{aligned}$$

where $\overset{d.}{=}$ means equal in distribution. Note that $\boldsymbol{E}^\tau\boldsymbol{E}$ is full rank $H \times H$ matrix, combining with Lemma 17 , one knows that

$$C_1(1 - \sqrt{\frac{H}{p}})^2 \leq \lambda_{min}(\frac{1}{p}\boldsymbol{E}^\tau\boldsymbol{E}) \leq \lambda_{max}(\frac{1}{p}\boldsymbol{E}^\tau\boldsymbol{E}) \leq C_2(1 + \sqrt{\frac{H}{p}})^2$$

for some positive constants C_1 and C_2 with probability at least $1-2\exp(-p/8)$. Note that $\lim \frac{p}{n} = \rho > 0$ as $n \to \infty$ and $n = Hc$, we know there exists positive constants C_1 and C_2 such that

$$C_1 \frac{1}{H}\|\boldsymbol{\theta}\|^2 \leq \|\boldsymbol{u}\|^2 \leq C_2 \frac{1}{H}\|\boldsymbol{\theta}\|^2$$

with probability at least $1 - 2\exp(-p/8)$.

On the other hand, the sliced stable condition (or Lemma 2) implies that $\lim \frac{1}{H}\|\boldsymbol{\theta}\|^2$ exists ($\neq 0$), so $\|\boldsymbol{u}\|^2$ is bounded away from 0 and ∞ with probability 1 as $n \to \infty$. □

7.6.2. *Proof of Lemma 8* *Proof*: For i), $\forall \boldsymbol{x} \in N_\alpha$, let

$$\boldsymbol{x} = \cos(\delta)\boldsymbol{\beta} + \sin(\delta)\boldsymbol{\eta} \text{ where } \boldsymbol{\eta} \perp \boldsymbol{\beta}, \|\boldsymbol{\eta}\| = 1, \delta \leq \alpha.$$

Since $B\boldsymbol{x} = \cos(\delta)\boldsymbol{u} + (\boldsymbol{u}^\tau\boldsymbol{\eta})\sin(\delta)\boldsymbol{\beta}$, we have:

$$\begin{aligned}\|B\boldsymbol{x}\| &\geq \cos(\delta)\|\boldsymbol{u}\| - \sin(\delta)\|\boldsymbol{u}\| \geq \frac{1}{2}\cos(\delta)\|\boldsymbol{u}\| \\ &\geq \frac{1}{2}\cos(\alpha)\|\boldsymbol{u}\| > \frac{C_1}{4} > 0\end{aligned}$$

for some positive constant C_1.

For ii), since

$$\boldsymbol{x}^\tau B\boldsymbol{x} = 2(\boldsymbol{u}^\tau\boldsymbol{\eta})\cos(\delta)\sin(\delta)$$

we have that uniformly over N_δ,

$$|\boldsymbol{x}^\tau B\boldsymbol{x}| \leq |\boldsymbol{u}^\tau\boldsymbol{\eta}|\sin(2\delta)$$

which in turn implies:

$$\left|\cos(\angle(B\boldsymbol{x}, \boldsymbol{x}))\right| = \frac{|\boldsymbol{x}^\tau B\boldsymbol{x}|}{\|\boldsymbol{x}\|\|B\boldsymbol{x}\|} \leq \frac{\sin(2\delta)|\boldsymbol{u}^\tau\boldsymbol{\eta}|}{\frac{1}{2}\cos(\delta)\|\boldsymbol{u}\|} \leq 4\delta \leq 4\alpha.$$

□

8. Appendix B: Proofs of Theorems 4 to 6. In this section, to ease the notation and avoid the tedious arguments, we introduce the following condition.

- Condition **C** : $var(\boldsymbol{m}(k, y)) = O(1)$; $c\ var(\boldsymbol{m}(k, y)) \to \infty$ when $n \to \infty$; $H = o\left(\log(n) \wedge \log(p)\right)$. Here $\boldsymbol{m}(k, y)$ denotes the k-th coordinate of the central curve $\boldsymbol{m}(y)$.

It can be easily verified that, in the scenarios of our interest, this condition are automatically satisfied.

8.1. *Outline of Proof of Theorem 4.* In the decomposition (15), let $\boldsymbol{\epsilon} = \boldsymbol{v} + \boldsymbol{w}$ and $\overline{\boldsymbol{\epsilon}}_{h,\cdot} = \frac{1}{c}\sum_i \boldsymbol{\epsilon}_{h,i}$, $\overline{\overline{\boldsymbol{\epsilon}}}(k) = \frac{1}{n}\sum_{i=1}^n \boldsymbol{\epsilon}_i(k)$. From Lemma 20 and the assumption **T2)**, we know that $\boldsymbol{\epsilon}(k)$ are sub-Gaussian. Let us denote $\sigma^2 = sup_k \mathbb{E}[\boldsymbol{\epsilon}(k)^2]$, which is bounded by a positive constant.

LEMMA 10. *Assuming the Condition* **C**, *we have the large deviation properties of* $var_{H,c}(\boldsymbol{x}(k))$. *Recall that* $\mathcal{T}$ *consists of coordinate* k *such that* $var(\mathbb{E}[\boldsymbol{x}(k)|y]) \neq 0$.

i) if $k \notin \mathcal{T}$, *then for any* $b = O(1)$, *such that* $\sqrt{n}(cb/4 - \sigma^2) \to \infty$, *one has*

$$\mathbb{P}(var_{H,c}(\boldsymbol{x}(k)) > b) \leq C_1 \exp\left(-C_2 \frac{cb}{H} + C_3 \log(H)\right) \tag{62}$$

for some positive constants C_1, C_2 *and* C_3.

ii) if $k \in \mathcal{T}$, *then one has*

$$|var_{H,c}(\boldsymbol{x}(k)) - var(\mathbb{E}[\boldsymbol{x}|y])| \geq \frac{1}{2} var(\boldsymbol{m}(k,y))$$

with probability at least

$$1 - C_1 \exp\left(-C_2 \frac{c\ var(\boldsymbol{m}(k,y))}{H} + C_3 \log(H)\right)$$

for some positive constants C_1, C_2 *and* C_3.

Proof of Theorem 4. Recall that

$$\begin{aligned}
\mathcal{T} &= \left\{ k \;\middle|\; \mathbb{E}[\boldsymbol{x}(k)|y] \text{ is not constant.} \right\} \\
\mathcal{I}_p(t) &= \left\{ k \;\middle|\; var_{H,c}(\boldsymbol{x}(k)) > t \right\} \\
\mathcal{E}_p(t) &= \left\{ k \;\middle|\; var_{H,c}(\boldsymbol{x}(k)) \leq t \right\}.
\end{aligned}$$

and $|\mathcal{T}| \leq Cs$ for some positive constant C. According to Lemma 10 and the Bonferroni's inequality, one has

$$\begin{aligned}
\mathbb{P}\left(\mathcal{T}^c \subset \mathcal{E}_p(t)\right) &\geq 1 - \sum_{k \in \mathcal{T}^c} \mathbb{P}\left(\ var_{H,c}(\boldsymbol{x}(k)) > t\right) \\
&\geq 1 - C_1 \exp\left(-C_2 \frac{ct}{H} + C_3 \log(H) + \log(p-s)\right).
\end{aligned} \tag{63}$$

and
(64)

$$\begin{aligned}\mathbb{P}(\mathcal{T} \subset I_p(t)) &\geq \mathbb{P}\left(\bigcap_{k\in\mathcal{T}}\left\{\ var_{H,c}(\boldsymbol{x}(k)) \geq \frac{1}{2}var(\boldsymbol{m}(k,y))\ \right\}\right)\\ &\geq 1-\sum_{k\in\mathcal{T}}\mathbb{P}\left(var_{H,c}\left(\boldsymbol{x}(k)\right) < \frac{1}{2}var\left(\boldsymbol{m}(k,y)\right)\right)\\ &\geq 1-C_1\exp\left(-C_2\frac{c\ var(\boldsymbol{m}(k,y))}{H}+C_3\log(H)+\log\left(Cs\right)\right).\end{aligned}$$

Since $var(\mathbb{E}[\boldsymbol{x}(k)|y]) > \frac{C}{s^\omega}$, one may choose $t = \frac{a}{s^\omega}$ for sufficiently small positive constant a such that for any k, $t < \frac{1}{2}var(\mathbb{E}[\boldsymbol{x}(k)|y]$, then one has (13) and (14) hold.

□

8.2. *Proof of Lemma 10*

8.2.1. *Proof of Lemma 10 i)* To avoid heavier notation, in this subsection, we pretend there are only $c-1$ samples in the H-th slice. Reader can easily adapt it into a rigorous way.

If $k \notin \mathcal{T}$ which means $\boldsymbol{m}(k,y) = 0$ (or equivalently $var(\boldsymbol{m}(k,y)) = 0$), we have

$$var_{H,c}(\boldsymbol{x}(k)) - var(\boldsymbol{m}(k,y)) = \frac{1}{H-1}\sum_h \bar{\boldsymbol{\epsilon}}_{h,\cdot}(k)^2 - \frac{H}{H-1}\bar{\bar{\boldsymbol{\epsilon}}}(k)^2.$$

Note that

$$\bar{\boldsymbol{\epsilon}}_{h,\cdot}(k) = \frac{c-1}{c}\frac{1}{c-1}\sum_{i=1}^{c-1}\boldsymbol{\epsilon}_{h,i}(k) + \frac{1}{c}\boldsymbol{\epsilon}_{h,c}(k),$$

one has:

$$\mathbb{P}(var_{H,c}(\boldsymbol{x}(k)) > b) \qquad (\text{ Let } b' = \tfrac{H-1}{H} b \text{ })$$

(65)

$$\begin{aligned}
&\leq \mathbb{P}(\frac{1}{H}\sum_{h} \bar{\boldsymbol{\epsilon}}_{h,\cdot}(k)^2 > b') \\
&\leq \mathbb{P}\left(\frac{2}{H}\sum_{h}\left(\frac{1}{c-1}\sum_{i=1}^{c-1}\boldsymbol{\epsilon}_{h,i}(k)\right)^2 + \frac{2}{Hc^2}\sum_{h}\boldsymbol{\epsilon}_{h,c}(k)^2 > b'\right) \\
&\leq \mathbb{P}\left(\frac{1}{H}\sum_{h}\left(\frac{1}{c-1}\sum_{i=1}^{c-1}\boldsymbol{\epsilon}_{h,i}(k)\right)^2 > b'/4\right) + \mathbb{P}\left(\frac{1}{Hc^2}\sum_{h}\boldsymbol{\epsilon}_{h,c}(k)^2 > b'/4\right) \\
&\leq \mathbb{P}\left(\frac{1}{H}\sum_{h}\left(\frac{1}{c-1}\sum_{i=1}^{c-1}\boldsymbol{\epsilon}_{h,i}(k)\right)^2 > b'/4\right) + \mathbb{P}\left(\frac{1}{cn}\sum_{i=1}^{n}\boldsymbol{\epsilon}_{i}(k)^2 > b'/4\right) \\
&\triangleq I + II
\end{aligned}$$

For I, from Lemma 12, we know that $\boldsymbol{\epsilon}_{h,i}(k)$ can be treadted as $c-1$ i.i.d. samples from $\boldsymbol{\epsilon}(k)|_{y\in S_h}$. Lemma 20 (iii) implies that

$$\mathbb{P}\left(\boldsymbol{\epsilon}(k)|_{y\in S_h} > t\right) \leq CH\exp\left(-\frac{t^2}{K^2}\right)$$

for some positive constant C. Since $\mathbb{E}[\boldsymbol{x}(k)|y] = 0$, one has $\mathbb{E}[\boldsymbol{x}(k)|y \in S_h] = 0$. According to Lemma 20 (iv),

$$\mathbb{P}\left(|\frac{1}{c-1}\sum_{i=1}^{c-1}\boldsymbol{\epsilon}_{h,i}(k)| > \sqrt{b'}/2\right) \leq 2C\exp\left(\frac{-b'(c-1)}{8CHK^2 + 4\sqrt{b'}K}\right).$$

Now we have:

$$\begin{aligned}
I \leq &\sum_{h=1}^{H}\mathbb{P}\left(|\frac{1}{c-1}\sum_{i=1}^{c-1}\boldsymbol{\epsilon}_{h,i}(k)| > \sqrt{b'}/2\right) \\
\leq &2CH\exp\left(\frac{-b'(c-1)}{8CHK^2 + 4\sqrt{b'}K}\right) \\
\leq &C_1\exp\left(-C_2\frac{cb}{H} + C_3\log(H)\right)
\end{aligned}$$

for some positive constants C_1, C_2 and C_3.

For II, note that $\boldsymbol{\epsilon}_i(k)$ are i.i.d. samples from a sub-Gaussian distribution $\boldsymbol{\epsilon}(k)$ with mean 0 and upper-exponentially bounded by $2K$. Lemma 22 gives us

$$II \leq \mathbb{P}\left(\left|\frac{1}{n}\sum_i \boldsymbol{\epsilon}_i(k)^2 - \mathbb{E}[\boldsymbol{\epsilon}(k)^2]\right| \geq cb'/4 - \mathbb{E}[\boldsymbol{\epsilon}(k)^2]\right) \tag{66}$$

$$\leq \mathbb{P}\left(\left|\frac{1}{n}\sum_i \boldsymbol{\epsilon}_i(k)^2 - \mathbb{E}[\boldsymbol{\epsilon}(k)^2]\right| \geq cb'/4 - 4K^2\right) \tag{67}$$

$$\leq C_1 \exp\left(-C_2 \frac{\sqrt{n}(cb'/4 - 4K^2)}{K^2}\right) \tag{68}$$

$$\leq C_1 \exp\left(-C_2 \frac{cb}{H} + C_3 \log(H)\right). \tag{69}$$

for some positive constants C_1, C_2 and C_3. We remind that in (67), we have used that $\mathbb{E}[\boldsymbol{\epsilon}(k)^2] \leq 4K^2$.

To summarize, if $b = O(1)$ and $\sqrt{n}(cb/4 - \mathbb{E}[\boldsymbol{\epsilon}(k)^2]) \to \infty$, we have

$$\mathbb{P}(var_{H,c}(\boldsymbol{x}(k)) > b) \leq C_1 \exp\left(-C_2 \frac{cb}{H} + C_3 \log(H)\right)$$

for some positive constants C_1, C_2 and C_3.

8.2.2. *Proof of Lemma 10 ii)* If $k \in \mathcal{T}$ which means $\boldsymbol{m}(k, y) \neq 0$ (or equivalently $var(\boldsymbol{m}(k, y)) \neq 0$), one has

$$var_{H,c}(\boldsymbol{x}(k)) - var(\boldsymbol{m}(k, y)) = G_1 + A_5 \tag{70}$$

where

$$G_1 = \frac{1}{H}\sum_h \overline{\boldsymbol{x}}_{h,\cdot}(k)^2 - var(\boldsymbol{m}(k, y)) \tag{71}$$

$$A_5 = \frac{1}{H(H-1)}\sum_h \overline{\boldsymbol{x}}_{h,\cdot}(k)^2 - \frac{H}{H-1}\overline{\overline{\boldsymbol{x}}}(k)^2. \tag{72}$$

Note that

$$\begin{aligned} |G_1| =& \Big|\frac{1}{H}\sum_h \overline{\boldsymbol{m}}_{h,\cdot}(k)^2 + \frac{2}{H}\sum_h \overline{\boldsymbol{m}}_{h,\cdot}(k)\overline{\boldsymbol{\epsilon}}_{h,\cdot}(k) + \frac{1}{H}\sum_h \overline{\boldsymbol{\epsilon}}_{h,\cdot}(k)^2 \\ & - var(\boldsymbol{m}(k, y))\Big| \\ \leq & A_1 + A_2 + A_3 + A_4, \end{aligned}$$

where

$$
\begin{aligned}
A_1 &= \Big|\frac{1}{H}\sum_h \mu_h(k)^2 - var(\boldsymbol{m}(k,y))\Big|,\\
A_2 &= \frac{1}{H}\sum_h \Big|\overline{\boldsymbol{m}}_{h,\cdot}(k)^2 - \mu_h(k)^2\Big|,\\
A_3 &= \frac{1}{H}\sum_h \bar{\boldsymbol{\epsilon}}_{h,\cdot}(k)^2,\\
A_4 &= (\frac{1}{H}\sum_h \overline{\boldsymbol{m}}_{h,\cdot}(k)^2)^{1/2}(\frac{1}{H}\sum_h \bar{\boldsymbol{\epsilon}}_{h,\cdot}(k)^2)^{1/2},
\end{aligned}
\tag{73}
$$

The deviation properties of A_i's are summarized in the following lemma.

LEMMA 11. *Let A_i's be defined as in equation (73) and (72). Assuming Condition* **C**, *then each of the following events*

i) $\Theta_1 = \left\{ A_1 \le \frac{1}{8} var(\boldsymbol{m}(k,y)) \right\}$,

ii) $\Theta_2 = \left\{ A_2 \le \frac{1}{4} var(\boldsymbol{m}(k,y)) \right\}$,

iii) $\Theta_3 = \left\{ A_3 \le \frac{1}{32} var(\boldsymbol{m}(k,y)) \right\}$,

iv) $\Theta_4 = \left\{ A_4 \le \frac{1}{16} var(\boldsymbol{m}(k,y)) \right\}$,

v) $\Theta_5 = \left\{ A_5 \le \frac{1}{64} var(\boldsymbol{m}(k,y)) \right\}$,

occurs with probability at least

$$
1 - C_1 \exp\left(-C_2 \frac{c\ var(\boldsymbol{m}(k,y))}{H} + C_3 \log(H)\right) \tag{74}
$$

for some positive constant C_1, C_2 and C_3.

Bonferroni's inequality implies that

$$
\mathbb{P}\left(\bigcap \Theta_i\right) \ge 1 - C_1 \exp\left(-C_2 \frac{c\ var(\boldsymbol{m}(k,y))}{H} + C_3 \log(H)\right). \tag{75}
$$

for some positive constant C_1, C_2 and C_3. The actual values of C_1,C_2 and C_3 might be different between (74) and (75).

Assuming Lemma 11, we know

$$
|var_{H,c}(\boldsymbol{x}(k)) - var(\boldsymbol{m}(k))| \ge \frac{1}{2} var(\boldsymbol{m}(k,y))
$$

with probability at least

$$1 - C_1 \exp\left(-C_2 \frac{c\ var(\boldsymbol{m}(k,y))}{H} + C_3 \log(H)\right).$$

. □

8.2.3. *Proof of Lemma 11.*

8.2.3.1. Proof of **i)** : Let $\epsilon = \frac{1}{Hn_0+1}$ and event $E(\epsilon)$ be defined as in Lemma 14 in Section 9.

When $E(\epsilon)^c$ happens, for sufficiently large H, one has

$$\left|\frac{1}{H}\sum_h \mu_h(k)^2 - var(\boldsymbol{m}(k,y))\right| \leq \frac{1}{8} var(\boldsymbol{m}(k,y))$$

by choosing $\boldsymbol{\beta} = e_k$ in (33). According to Lemma 14,

$$\mathbb{P}(E(\epsilon)^c) \geq 1 - CH^2\sqrt{Hc+1}\exp\left(-(Hc+1)\frac{\epsilon^2}{32}\right). \tag{76}$$

Note that $var(\boldsymbol{m}(k,y)) = O(1)$, $H = o(\log(n))$ and $\epsilon = \frac{1}{Hn_0+1}$. Then

(77)
$$\exp\left(-C_2 \frac{c\ var(\boldsymbol{m}(k,y))}{H} + C_3 \log(H)\right) \succ \exp\left(-\frac{Hc+1}{32(Hn_0+1)^2} + \log(H^2\sqrt{Hc+1})\right).$$

for some positive constant C_1, C_2 and C_3.

8.2.3.2. Proof of **ii)** : Choose $\boldsymbol{\beta} = e_k$ in (38), then one has

$$A_2 \leq I + II + III + IV$$

where

$$I = \frac{1}{H}\sum_{h=1}^{H-1}\left|\overline{\boldsymbol{m}}_{h,1:(c-1)}(k)^2 - \mu_h(k)^2\right| + \frac{1}{H}\left|\overline{\boldsymbol{m}}_{H,\cdot}(k)^2 - \mu_H(k)^2\right| \tag{78}$$

$$II = \frac{2}{Hc}\sum_{h=1}^{H}\mu_h(k)^2 \tag{79}$$

$$III = \frac{2(c-1)}{c}\sqrt{\frac{1}{H}\sum_{h=1}^{H-1}\overline{\boldsymbol{m}}_{h,1:(c-1)}(k)^2 + \frac{1}{H}\overline{\boldsymbol{m}}_{H,\cdot}(k)^2}\sqrt{IV} \tag{80}$$

$$IV = \frac{1}{Hc^2}\sum_{h=1}^{H}\boldsymbol{m}_{h,c}(k)^2 \tag{81}$$

One only needs to prove that each term, with high probability, is smaller than a multiple of $(var(m(y)))$ as n $\to \infty$.

For I, Let $\eta(k)$ be the maximal value in $\left|\overline{\boldsymbol{m}}_{h,1:(c-1)}(k)-\mu_h(k)\right|, h=1,...,H-1$ and $\left|\overline{\boldsymbol{m}}_{H,\cdot}(k)-\mu_H(k)\right|$. Since $\boldsymbol{m}_{h,i}(k), i=1,..,c-1$ can be treated as i.i.d. random samples of $\boldsymbol{m}(k)|_{y\in S_h}$ for $h=1,...,H-1$ and $\boldsymbol{m}_{H,i}(k), i=1,..,c$ can be treated as i.i.d. random samples of $\boldsymbol{m}(k)|_{y\in S_H}$.

Let

$$E_1(a)=\left\{\ \eta(k)>a\sqrt{var(\boldsymbol{m}(k,y))}\ \right\}. \tag{82}$$

According to Lemma 20 (iv) and Bonferroni's inequality, one has

$$\mathbb{P}\left(E_1(a)\right)\leq 2H\exp\left(\frac{-(c-1)a^2var(\boldsymbol{m}(k,y))}{2CHK^2+2a\sqrt{var(m(y))}K}\right) \tag{83}$$

$$\leq C_1\exp\left(-C_2\frac{a^2c\ var(\boldsymbol{m}(k,y))}{H}+\log(H)\right) \tag{84}$$

for some positive constants C_1 and C_2.

On the event $E(\epsilon)^c\cap E_1(a)^c$, combining with (36), we have

$$\begin{aligned} I\leq&\frac{1}{H}\sum_h\eta(k)(\eta(k)+2|\mu_h(k)|)\leq\eta(k)^2+\frac{2\eta(k)}{H}\sum_h|\mu_h(k)|\\ \leq&\left(a^2+2a\left(1+\frac{C'}{H^\kappa}\right)^{1/2}\right)var(\boldsymbol{m}(k,y))\\ \leq&\frac{1}{32}var(\boldsymbol{m}(k,y)), \end{aligned}$$

when choosing a sufficiently small a.

Remark 6. From above, on the event $E(\epsilon)^c\cap E_1(a)^c$, one has

$$\frac{1}{H}\sum_{h=1}^{H-1}\overline{\boldsymbol{m}}_{h,1:(c-1)}(k)^2+\frac{1}{H}\overline{\boldsymbol{m}}_{H,\cdot}(k)^2\leq\frac{33}{32}var(\boldsymbol{m}(k,y)). \tag{85}$$

For II, apply a similar argument near (66), we know that

$$\begin{aligned} \mathbb{P}\left(II>\frac{1}{1024}var(\boldsymbol{m}(k,y))\right)\leq&\mathbb{P}(\frac{1}{nc}\sum_i m_i^2>var(\boldsymbol{m}(k,y)))\\ \leq&C_1\exp\left(-C_2\sqrt{n}(c\ var(\boldsymbol{m}(k,y))-K^2)\right)\\ \leq&C_1\exp\left(-C_2\frac{c\ var(\boldsymbol{m}(k,y))}{H}+C_3\log(H)\right) \end{aligned}$$

for some positive constant C_1, C_2 and C_3

For III, Let

$$E_2 \triangleq \left\{ \ II > \frac{1}{1024} var(\boldsymbol{m}(k,y)) \ \right\}.$$

When the event $E(\epsilon)^c \cap E_1(a)^c \cap E_2^c$ occurs, according to equation (85),

$$III \leq \frac{2(c-1)}{c}\sqrt{\frac{33}{32}}\frac{1}{32} var(\boldsymbol{m}(k,y)) < \frac{1}{8} var(\boldsymbol{m}(k,y)).$$

For VI, When the event $E(\epsilon)^c \cap E_1\,(a)^c \cap E_2^c$ occurs, from (33), we know

$$VI = \frac{2}{Hc}\sum_h \mu_h(k)^2 \leq \frac{9}{4c} var(\boldsymbol{m}(k,y)) < \frac{1}{16} var(\boldsymbol{m}(k,y))$$

when c is sufficiently large.

To summarize, we know that there exist positive constant C_1, C_2, C_3 and C_4 such that

$$A_2 \leq I + II + III + VI \leq \frac{1}{4} var(\boldsymbol{m}(k,y))$$

holds on the event $E(\epsilon)^c \cap E_1(a)^c \cap E_2^c$ which is with probability at least

$$1 - C_1 \exp\left(-C_2 \frac{c\ var(\boldsymbol{m}(k,y))}{H} + C_3 \log(H)\right)$$

for some positive constants C_1,C_2 and C_3.

8.2.3.3. Proof of **iii**) : Similar to the proof of Lemma 10 (i) one has

$$\mathbb{P}(A_3 > b) \leq C_1 H \exp\left(\frac{-(c-1)b}{8C_2 H K_1^2 + 4\sqrt{b}K_2}\right) + C_3 \exp(-bc/(16K_1^2)).$$

for some positive constants C_1, C_2 and C_3. In particular , if we take $b = \frac{1}{16} var(\boldsymbol{m}(k,y))$, we know that

$$A_3 \leq \frac{1}{16} var(\boldsymbol{m}(k,y))$$

with probability at least

$$1 - C_1 \exp\left(-C_2 \frac{c\ var(\boldsymbol{m}(k,y))}{H} + C_3 \log(H)\right)$$

for some positive constant C_1, C_2 and C_3 .

REMARK 7. In the proof, one actually sees that for any positive constant a,

$$A_3 \leq a\ var(\boldsymbol{m}(k,y))$$

with probability at least

$$1 - C_1 \exp\left(-C_2 \frac{c\ var(\boldsymbol{m}(k,y))}{H} + C_3 \log(H)\right)$$

for some positive constant C_1, C_2 and C_3 (which depend on a). We will use it in the proof of **iv**).

8.2.3.4. Proof of **iv**) : Let

$$B_1 \triangleq \frac{1}{H}\sum_h \overline{\boldsymbol{m}}_{h,\cdot}(k)^2 \quad B_2 \triangleq A_3 = \frac{1}{H}\sum_h \bar{\boldsymbol{\epsilon}}_{h,\cdot}(k)^2$$

Consequently,

$$\begin{aligned} &\mathbb{P}\left(B_1^{1/2}B_2^{1/2} > \frac{1}{32} var(\boldsymbol{m}(k,y))\right) \\ &\leq \mathbb{P}\left(|B_1| > \frac{11}{8} var(\boldsymbol{m}(k,y))\right) + \mathbb{P}\left(B_2 > \frac{var(\boldsymbol{m}(k,y))}{32 * 44}\right) \end{aligned} \tag{86}$$

Note that

$$|B_1 - var(\boldsymbol{m}(k,y))| \leq A_2 + A_1$$

According to (i), (ii), and Remark 7, the right hand side of (86) is bounded by

$$C_1 \exp\left(-C_2 \frac{c\ var(\boldsymbol{m}(k,y)}{H} + C_3 \log(H)\right)$$

for some positive constants C_1, C_2 and C_3.

8.2.3.5. Proof of **v**) : From **i**) to **vi**), we know that

$$\frac{1}{H}\sum_h \overline{\boldsymbol{x}}_{h,\cdot}(k)^2 \leq \frac{3}{2} var(\boldsymbol{m}(k,y))$$

with probability at least

$$1 - C_1 \exp\left(-C_2 \frac{c\ var(m(y))}{H} + C_3 \log(H)\right)$$

for some positive constants $C_1, ..., C_3$. Since $x_1, ..., x_n$ are i.i.d. samples of a sub-Gaussian random variable with mean 0 and upper-exponentially bounded by K, we know that

$$\mathbb{P}(|\overline{\overline{\boldsymbol{x}}}(x)| > \sqrt{var(\boldsymbol{m}(k,y))/128}) \leq C_4 \exp\left(-C_5 n \ var(\boldsymbol{m}(k,y))\right) \tag{87}$$

$$\leq 1 - C_1 \exp\left(-C_2 \frac{c \ var(\boldsymbol{m}(k,y))}{H} + C_3 \log(H)\right) \tag{88}$$

for some positive constants $C_1, ..., C_5$

In particular, we have

$$\begin{aligned}
&\mathbb{P}(A_5 > \frac{1}{64} var(\boldsymbol{m}(k,y))) \\
&\quad \leq \mathbb{P}(|\frac{1}{H}\sum_h \overline{x}_{h\cdot}^2| > \frac{(H-1)var(\boldsymbol{m}(k,y))}{128}) + \mathbb{P}(|\overline{\overline{\boldsymbol{x}}}(x)| > \sqrt{\frac{var(\boldsymbol{m}(k,y))}{128}}) \\
&\quad \leq C_1 \exp\left(-C_2 \frac{c \ var(\boldsymbol{m}(k,y))}{H} + C_3 \log(H)\right)
\end{aligned}$$

□

8.3. *Proof of Theorem 5* By choosing H, c and $t = \frac{a}{s^\omega}$ properly, from Theorem 4, one has

$$P(\widehat{\mathcal{T}} = \mathcal{T}) \geq 1 - C_1 \exp\left(-C_2 \frac{n}{H^2 s^\omega} + C_3 \log(H) + \log(p-s)\right)$$

for some positive constants $C_i, i = 1, 2, 3$. When $\widehat{\mathcal{T}} = \mathcal{T}$, we have $|\widehat{\mathcal{T}}| = O(s)$. For the n samples $(Y_i, X_i^{\widehat{\mathcal{T}}})$, apply Theorem 1, one has

$$\|e(\widehat{\boldsymbol{\Lambda}}_p^{\mathcal{T},\mathcal{T}}) - \boldsymbol{\Lambda}_p\|_2 \leq \|\widehat{\boldsymbol{\Lambda}}_p^{\mathcal{T},\mathcal{T}} - \boldsymbol{\Lambda}_p^{\mathcal{T},\mathcal{T}}\|_2 \leq O_P(\frac{dH^2}{\sqrt{n}} + \frac{1}{H^{1\wedge\kappa}} + \frac{H^2 s}{n} + \sqrt{\frac{H^2 s}{n}}).$$

In particular, with probability converging to one, we have

$$\|e(\widehat{\boldsymbol{\Lambda}}_p^{\mathcal{T},\mathcal{T}}) - \boldsymbol{\Lambda}_p\|_2 \to 0 \text{ as } n \to \infty.$$

8.4. *Proof of Theorem 6.* The proof is similar to the proof of Theorem 2, except we use the Theorem 1 in Bickel and Levina [2008].

9. Appendix C

9.1. *Assisting Lemmas*

DEFINITION 2. *A set of random variables $x_1, ..., x_n$ can be treated as i.i.d random samples from a random variable x, if for any n variates symmetric function $f(w_1, ..., w_n)$, $f(x_1, ..., x_n)$ is identically distributed as $f(z_1, ..., z_n)$ where $z_1, ..., z_n$ are i.i.d random samples from x.*

LEMMA 12. *Let (x_i, y_i) be n i.i.d random samples from a joint distribution (x, y). Sort these samples according to the order statistics of y_i's and denote the sorted samples by $(x_{(1)}, y_{(1)}), (x_{(2)}, y_{(2)}), ..., (x_{(n)}, y_{(n)})$. Then for any a, b $(1 \leq a \leq b + 1 \leq n)$, $x_{(a+1)}, ..., x_{(b)}$ can be treated as $b - a$ i.i.d samples from $x\Big|\left(y \in [y_{(a)}, y_{(b+1)}]\right)$.*

PROOF. In fact, one only needs to prove that $y_{(a+1)}, \cdots, y_{(b)}$ can be treated as $b - a$ i.i.d. samples of $y\Big|(y \in [y_{(a)}, y_{(b+1)}])$. The latter only needs to be proved for uniform distribution which can be verified directly.

COROLLARY 1. *In the slicing inverse regression contexts, recall that S_h denotes the h-th interval $(y_{h-1,c}, y_{h,c}]$ for $2 \leq h \leq H-1$ and $S_1 = (-\infty, y_{1,c}]$, $S_H = (y_{H-1,c}, \infty)$. One has that $x_{h,i}, i = 1, \cdots, c-1$ can be treated as $c-1$ random samples of $x\Big|(y \in S_h)$ for $h = 1, ..., H-1$ and $x_{H,1}, ..., x_{H,c}$ can be treated as c random samples of $x\Big|(y \in S_H)$.*

LEMMA 13. *Suppose that (x, y) are defined over σ-finite space $\mathcal{X} \times \mathcal{Y}$ and g is a non-negative function such that $\mathbb{E}[g(x)]$ exists. For any fixed positive constants $C_1 < 1 < C_2$, there exists a constant C which only depends on C_1, C_2 such that for any partition $\mathbb{R} = \bigcup_{h=1}^{H} S_h$ where S_h are intervals satisfying*

$$\frac{C_1}{H} \leq \mathbb{P}(y \in S_h) \leq \frac{C_2}{H}, \forall h, \tag{89}$$

one has

$$\sup_h \mathbb{E}(g(x)\Big|y \in S'_h) \leq CH\mathbb{E}[g(x)].$$

PROOF. According to Fubini's Theorem, for any h,

$$\begin{aligned}\mathbb{E}[g(x)] =& \sum_k \mathbb{P}(y \in S_k) \int_{\mathcal{X}} g(x)p(x|y \in S_k)dx \\ \geq& \mathbb{P}(y \in S_h) \int_{\mathcal{X}} g(x)p(x|y \in S_h)dx.\end{aligned}$$

Due to the condition (89), there exists a positive constant C such that

$$\int_{\mathcal{X}} g(x)p(x|y \in S_h)dx \leq CH\mathbb{E}[g(x)].$$

□

COROLLARY 2. *Let $\boldsymbol{x}$ be a multivariate random variable with covariance matrix $\boldsymbol{\Sigma}$. For any partition satisfying (89), there exists a constant C such that*

$$var(\boldsymbol{\beta}^\tau \boldsymbol{x}|_{y\in S'_h}) \leq CHvar(\boldsymbol{\beta}^\tau \boldsymbol{x}), \text{for any unit vector } \boldsymbol{\beta},$$

and

$$\lambda_{max}\left(var\left(\boldsymbol{x}\middle|y \in S'_h\right)\right) \leq CH\lambda_{max}\left(var\left(\boldsymbol{x}\right)\right).$$

COROLLARY 3. *Let x be a sub-Gaussian random variable which is upper-exponentially bounded by K. Then for any partition satisfying (89), there exists a constant C such that*

$$\mathbb{E}[\exp\left(\frac{x^2}{K^2}\right)\Big|y \in S'_h] \leq CH\mathbb{E}[\exp\left(\frac{x^2}{K^2}\right)].$$

Recall the definition of the random intervals $S_h, h = 1, 2, \cdots, H$ and random variable $\delta_h = \delta_h(\omega) = \int_{y\in S_h(\omega)} f(y)dy$.

LEMMA 14. *Define the event $E(\epsilon) = \left\{ \omega \;\middle|\; |\delta_h - \frac{1}{H}| > \epsilon, \forall h \right\}$. There exists a positive constant C such that, for any $\epsilon > \frac{4}{Hc-1}$ one has*

$$P(E(\epsilon)) \leq CH^2\sqrt{Hc+1}\exp(-(Hc+1)\frac{\epsilon^2}{32}) \tag{90}$$

for sufficient large H and c.

PROOF. The proof is deferred to the end of this paper.

9.2. *Some Results from Random Matrices Theory.* We collect some direct corollaries of the non-asymptotic random matrices theory (e.g., Rudelson and Vershynin [2013]).

LEMMA 15. *Let $\boldsymbol{M}$ be any $p \times n$ matrix $(n > p)$ whose columns $\boldsymbol{M}_i$ are independent sub-Gaussian random vectors in $\mathbb{R}^p$ with second moment $\mathbf{I}_p$ and $\lambda^+_{sing,min}(\boldsymbol{M})$, $\lambda_{sing,max}(\boldsymbol{M})$ be the minimal non-zero and maximal singular value of $\boldsymbol{M}$. Then for every t, with probability at least $1 - 2\exp(-C't^2)$ one has:*

$$\sqrt{n} - C\sqrt{p} - t \leq \lambda^+_{sing,min}(\boldsymbol{M}) \leq \lambda_{sing,max}(\boldsymbol{M}) \leq \sqrt{n} + C\sqrt{p} + t.$$

LEMMA 16. *Let $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n$ be n i.i.d. samples from a p-dimensional sub-Gaussian random variable with covariance matrix $\boldsymbol{\Sigma}$ and $\rho = \frac{p}{n}$. If there exists positive constants C_1 and C_2 such that*

$$C_1 \leq \lambda_{\min}(\boldsymbol{\Sigma}_{\boldsymbol{x}}) \leq \lambda_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{x}}) \leq C_2.$$

Let $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{x}} = \frac{1}{n}\sum_i \boldsymbol{x}_i \boldsymbol{x}_i^\tau$. Then

$$\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{x}} - \boldsymbol{\Sigma}_{\boldsymbol{x}}\|_2 \to 0 \text{ if } \rho = 0 \text{ when } n \to \infty.$$

It is also easy to see that, given the boundedness condition on $\boldsymbol{\Sigma}_X$, $\|\widehat{\boldsymbol{\Sigma}}_X^{-1} - \boldsymbol{\Sigma}_X^{-1}\|_2 \to 0$ if $\rho = \frac{p}{n} \to 0$ when $n \to \infty$.

PROOF. Let $\boldsymbol{x}_i = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{1/2}\boldsymbol{m}_i$ where $\boldsymbol{m}_i$ is sub-Gaussian random variable with covariance matrix $\mathbf{I}_p$ and $\boldsymbol{M} = (\boldsymbol{m}_1, ..., \boldsymbol{m}_n)$. From Lemma 15, one has

$$\|\frac{1}{n}\boldsymbol{M}\boldsymbol{M}^\tau - \mathbf{I}_p\|_2 \to 0$$

and

$$\|\widehat{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma}^{-1}\|_2 = \|\boldsymbol{\Sigma}^{-1/2}\|_2 \|\frac{1}{n}\boldsymbol{M}\boldsymbol{M}^\tau - \mathbf{I}_p\|_2 \|\boldsymbol{\Sigma}^{-1/2}\|_2 \to 0,$$

with probability converges to 1 as $n \to \infty$ because

$$\lambda_{max}\left(\frac{1}{n}\boldsymbol{M}\boldsymbol{M}^\tau\right) \leq \left(1 + \frac{(C+1)\sqrt{p}}{\sqrt{n}}\right)^2$$

and

$$\lambda_{min}\left(\frac{1}{n}\boldsymbol{M}\boldsymbol{M}^\tau\right) \geq \left(1 - \frac{(C+1)\sqrt{p}}{\sqrt{n}}\right)^2$$

with probability at least $1 - 2\exp(-C'p)$. □

The following lemma is well known in the non-asymptotic random matrix theory (Vershynin [2010] Proposition 5.34) which is slightly different from the Lemma 15.

LEMMA 17. *Let $\boldsymbol{E}_{p\times H}$ 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 :*

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

, and

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

COROLLARY 4. *One has*

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

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

PROOF. Choosing $t = \sqrt{p}/2$, according to Lemma 17, one has:

$$\mathbb{P}\left(\frac{\lambda_{max}(E_H)}{\sqrt{p}+\sqrt{H}} \leq \frac{3}{2}\right) \geq \mathbb{P}\left(\frac{\lambda_{max}(E_H)}{\sqrt{p}+\sqrt{H}} \leq 1+\frac{\sqrt{p}}{2\sqrt{p}+2\sqrt{H}}\right)$$

and

$$\mathbb{P}\left(\frac{\lambda_{min}^{+}(E_H)}{\sqrt{p}-\sqrt{H}} \geq \frac{1}{2}\right) \geq \mathbb{P}\left(\frac{\lambda_{max}(E_H)}{\sqrt{p}-\sqrt{H}} \geq 1-\frac{\sqrt{p}}{2\sqrt{p}-2\sqrt{H}}\right)$$

with probability at least $1-2\exp(-p/8)$. i.e., With probability converging to one, one has

$$\frac{1}{2}(\sqrt{p}-\sqrt{H}) \leq \lambda_{min}^{-}(E_{p\times H}) \leq \lambda_{max}(E_{p\times H}) \leq \frac{3}{2}(\sqrt{p}+\sqrt{H}).$$

□

9.3. *Basic Properties of sub-Gaussian random variables.* We rephrased several equivalent definitions of the sub-Gaussian distribution here (See e.g., Vershynin [2010]):

DEFINITION 3. *Let* x *be a random variable. Then the following properties are equivalent with parameters* K_i*'s differing from each other by at most an absolute constant factor,*

1. *Tails:* $\mathbb{P}(|x| > t) \leq \exp(1-t^2/K_1^2)$ *for all* $t \geq 0$.
2. *Moments:* $(\mathbb{E}[|x|^p])^{1/p} \leq K_2\sqrt{p}$ *for all* $p \geq 1$.
3. *Super-exponential moment:* $\mathbb{E}\exp(x^2/K_3^2) \leq e$.

Moreover, if $\mathbb{E}[x] = 0$, *then the properties* $1-3$ *are also equivalent to the following one:*

4. *Moment generating function:* $\mathbb{E}[\exp(tx)] \leq \exp(t^2K_4^2)$.

DEFINITION 4. *For a sub-Gaussian random variable* x *with the constants* $K_i, i = 1,2,3,4$ *given in Definition 3, we will call a constant* K *an upper-exponential bound of* x *or* x *is upper-exponentially bounded by* K *if* $K > \max_i\{K_1, K_2, K_3, K_4\}$.

We summarize some properties regarding the sub-Gaussian distributions into the following lemmas.

LEMMA 18. *Let $\delta_1, ..., \delta_n$ be n (not necessarily independent or with mean zero) sub-Gaussian random variables upper-exponentially bounded by K.*

- **i)** *$\frac{1}{n}\sum_{i=1}^n \delta_i$ is sub-Gaussian and upper-exponentially bounded by K.*
- **ii)** *$\delta_1 - \mathbb{E}[\delta_1]$ is sub-Gaussian upper-exponentially bounded by $2K$.*
- **iii)** *If they are independent and with mean zero , then $\frac{1}{\sqrt{n}}\sum_i \delta_i$ is sub-Gaussian and upper-exponentially bounded by K.*
- **iv)** *If they are i.i.d., then one has the concentration inequality:*

$$\mathbb{P}\left(\left|\frac{\sum_{i=1}^n \delta_i}{n} - \mathbb{E}[x]\right| > t\right) \leq 2\exp\left(\frac{-nt^2}{2K^2e + 2tK}\right).$$

PROOF. **i)** follows from the linear property of expectation and the the convexity of exponential function. i.e.,

$$\mathbb{E}[\exp(\frac{1}{nK^2}\sum_i \delta_i^2)] \leq \mathbb{E}[\frac{1}{n}\sum_h \exp(\frac{\delta_i^2}{K^2})] \leq \max_i \mathbb{E}[\exp(\frac{\delta_i^2}{K^2})] \leq e.$$

ii) From Definition 3, we know that $|\mathbb{E}[\delta_i]| \leq K$ which gives us the desired upper-exponential bound of $\delta_i - \mathbb{E}[\delta_i]$.
iii) is trivial as $\delta_1, \cdots, \delta_c$ are independent and with mean zero.
iv) Since δ_1 is sub-Gaussian upper-exponentially bounded by K, we have:

$$\begin{aligned}
\mathbb{E}[|\delta_1|^p] &= \int_0^\infty pt^{p-1}\mathbb{P}(|\delta_1| > t)dt \leq \int_0^\infty pt^{p-1}\exp\left(1 - \frac{t^2}{K^2}\right)dt \\
&= \frac{ep}{2}\Gamma(\frac{p}{2})K^p && \text{for any } p \geq 1 \\
&\leq p!K^{p-2}\frac{(K^2e)}{2} && \text{for any } p \geq 2
\end{aligned}$$

Recall the well known Bernstein inequality.

LEMMA 19. **(Bernstein Inequality).** *If there exists positive constants V and b such that for any integers $m \geq 2$,*

$$E[|\delta_1|^m] \leq m!b^{m-2}V/2$$

then:

$$\mathbb{P}\left(\left|\frac{\sum_{i=1}^n \delta_i}{n} - \mathbb{E}[x]\right| > t\right) \leq 2\exp\left(\frac{-nt^2}{2V + 2tb}\right). \tag{91}$$

By chooing $V = K^2 e$ and $b = K$, we get the desired concentration inequality. □

LEMMA 20. *Suppose that (x, y) are defined over σ-finite space $\mathcal{X} \times \mathcal{Y}$ and x is sub-Gaussian with mean 0 and upper exponentially bounded by K, let $m(y) = \mathbb{E}[x|y]$, $\epsilon(y) = x - m(y)$, then we have*

i) *$m(y)$ and $\epsilon(y)$ are sub-Gaussian and upper-exponentially bounded by K and $2K$ respectively.*

ii) *Let $\mathcal{Z}$ consists of points y such that $x|_y$ is not sub-Gaussian, i.e.,*

$$\mathcal{Z} \triangleq \left\{ y \mid \exists t \in (0, t_0] \text{ such that } \int_{\mathcal{X}} \exp(tx^2) p(x|y) p(y) dx = \infty \right\}$$

, then $\mathbb{P}(y \in \mathcal{Z}) = 0$.

iii) *For any fixed positive constants $C_1 < 1 < C_2$ and any partition $\mathbb{R} = \bigcup_{h=1}^{H} S_h$ where S_h are intervals satisfying*

$$\frac{C_1}{H} \leq \mathbb{P}(y \in S_h) \leq \frac{C_2}{H}, \forall h$$

there exists a constant C such that

$$\sup_h \mathbb{P}(x|_{y \in S_h} > t) \leq CH \exp\left(1 - \frac{t^2}{K^2}\right).$$

As a direct corollary, we know that there exists a positive constant C such that

$$\mathbb{E}\left[\exp\left(\frac{(x|_{y \in S_h})^2}{2K^2}\right)\right] \leq CH,$$

and

$$\mathbb{E}\left[\left|(x|_{y \in S_h})\right|^m\right] \leq CHmK^m \Gamma(\frac{m}{2})/2.$$

vi) *Suppose that $x|_{y \in S_h}$ is defined as in* **iii**). *Let x_i, i=1,...,c be c samples from $x|_{y \in S_h}$, $\bar{x}_h = \frac{1}{c} \sum_i x_i$ and $\mu_h = \mathbb{E}[x|_{y \in S_h}]$, one has*

$$\mathbb{P}[|\bar{x}_h - \mu_h| > t] \leq 2 \exp\left(\frac{-ct^2}{2CHK^2 + 2tK}\right).$$

PROOF. **i)** By Jensen's inequality, we have

$$\mathbb{E}[\exp(t\mathbb{E}[x|y])] \leq \mathbb{E}[\mathbb{E}[\exp(tx)|y]] = \mathbb{E}[\exp(tx)] \leq \exp(t^2 K_1^2).$$

i.e., $m(y)$ is sub-Gaussian and upper-exponentially bounded by K_1. Since x , $m(y)$ is sub-Gaussian and upper-exponentially bounded by K_1, we know that $\epsilon = x - m(y)$ is sub-Gaussian and upper-exponentially bounded by $2K_1$.

ii) Let $p(x, y)$ be the joint density function of (x, y) and $p(x)$, $p(y)$ be the marginal distribution of x, y. Since x is sub-Gaussian, we know there exists $t_0 > 0$ such that

$$\int_{\mathcal{X}} \exp(tx^2) \int_{\mathcal{Y}} p(x|y)p(y)dydx \leq e \text{ for } 0 \leq t \leq t_0.$$

By Fubini Theorem, we know

$$\int_{\mathcal{Y}} p(y) \int_{\mathcal{X}} \exp(tx^2)p(x|y)dxdy \leq e \text{ for } 0 \leq t \leq t_0. \tag{92}$$

Recall that we have $\mathcal{Z} \triangleq \{y|\exists t \in (0, t_0] \text{ such that } \int_{\mathcal{X}} \exp(tx^2)p(x|y)p(y)dx = \infty\}$, from equation (92), we know $\mathbb{P}(y \in \mathcal{Z})$=0. In particular, we know that for any $y \notin \mathcal{Z}$, $x|_y$ is sub-Gaussian. [1]

iii) From Lemma 13, we know that there exists a positive constant C such that

$$\int_{\mathcal{X}} \exp(tx^2)p(x|y \in S_h)dx \leq CHe.$$

For simplicity if notation, we will denote $x|_{y\in S_h}$ by z through out this lemma. So for $0 \leq t \leq t_0 = \frac{1}{K}$, one has

$$\mathbb{P}(z > a) \leq \frac{\mathbb{E}[\exp((tz)^2)]}{\exp(t^2a^2)} \leq CHe \exp(-t^2a^2).$$

From the above tail bounds, one has that for any integer $m > 0$

$$\begin{aligned}\mathbb{E}[|z|^m] = \int_0^\infty \mathbb{P}(|z| > t)(m)t^{m-1}dt \leq CHm \int_0^\infty exp(-\frac{t^2}{K^2})t^{m-1}dt \\ \leq CHmK^m\Gamma(m/2)/2.\end{aligned}$$

We then have

$$\begin{aligned}\mathbb{E}\left[\exp\left(tz^2\right)\right] &\leq \sum_{m=0}^{\infty} \frac{\mathbb{E}\left[t^m z^{2m}\right]}{m!} \leq \sum_{m=0}^{\infty} \frac{\mathbb{E}\left[t^m z^{2m}\right]}{m!} \\ &\leq \sum_{m=0}^{\infty} \frac{t^m CHmK^{2m}\Gamma(m)}{m!} = CH \sum_{m=0}^{\infty} t^m K^{2m}\end{aligned}$$

[1] However, the norm (e.g., sub-exponential norm)of $x|_y$ might be varying along with y and , as a function of y , it might be not bounded.

From which we know if $0 \leq t < \frac{1}{2}K^{-2}$, the R.H.S is bounded by CH for a positive constant C.

vi) From the previous proof, we know that for any integer $m \geq 2$

$$\mathbb{E}[|z|^m] \leq CHm!K^m = m!K^{m-2}(2CHK^2)/2.$$

By the Bernstein inequality (91), we have:

$$\mathbb{P}\left(\left|\frac{\sum_{i=1}^{c} z_i}{c} - \mathbb{E}[z]\right| > t\right) \leq 2\exp\left(\frac{-ct^2}{2CHK^2 + 2tK}\right).$$

We remind that we use C denote an absolute constant which the actual value may vary from case to case.

□

LEMMA 21. *Let $z_i, i = 1, \cdots, n$ be i.i.d. samples of a sub-Gaussian distribution exponentially upper bounded by K, then there exist positive constants C_1, C_2 such that, if $\sqrt{n}\epsilon \to \infty$, one has*

$$\mathbb{P}(|\frac{1}{n}\sum_i (z_i - \bar{z})^2 - var(z)| > \epsilon) \leq C_1 \exp(-C_2 \frac{\epsilon\sqrt{n}}{K^2}),$$

where $\bar{z} = \frac{1}{n}\sum_i z_i$.

PROOF. Recall the following Hanson-Wright inequality in Rudelson and Vershynin [2013]

LEMMA 22. *Let $\boldsymbol{v} = (\boldsymbol{x}(1), \cdots, \boldsymbol{x}(n))$ be a sub-Gaussian random vector with independent components $\boldsymbol{x}(k)$ such that $\mathbb{E}[\boldsymbol{x}(k)] = 0$ and $\|\boldsymbol{x}(k)\|_{\psi_2} \leq K$. Let $\boldsymbol{A}$ be an $n \times n$ matrix. Then there exists a positive constant C such that for any $t > 0$,*

$$\mathbb{P}\{|\boldsymbol{x}^\tau \boldsymbol{A}\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}^\tau A\boldsymbol{x}]| > t\} \leq 2\exp\left(-C \ \min\left(\frac{t^2}{K^4\|\boldsymbol{A}\|_{HS}^2}, \frac{t}{K^2\|\boldsymbol{A}\|_{HS}}\right)\right).$$

Here the ψ_2 norm of a random variable z is defined as $\|z\|_{\psi_2} \triangleq \sup_p p^{-1/2}(\mathbb{E}|z|^p)^{1/p}$ and the HS norm of a matrix $\boldsymbol{A}$ is defined as $\|\boldsymbol{A}\|_{HS} = (\sum_{i,j}|a_{i,j}|^2)^{1/2}$.

Since

$$\begin{aligned}
&\mathbb{P}\left(|\frac{1}{n}\sum_i (z_i-\bar{z})^2 - var(z)| > 2\epsilon\right)\\
&= \mathbb{P}\left(|\frac{1}{n}\sum_i (z_i-\mathbb{E}[z])^2 - (\mathbb{E}[z]-\bar{z})^2 - \mathbb{E}[(z-E[z])^2]| > 2\epsilon\right)\\
&\leq \mathbb{P}\left(|\frac{1}{n}\sum_i (z_i-\mathbb{E}[z])^2 - \mathbb{E}[(z-\mathbb{E}[z])^2]| > \epsilon\right) + \mathbb{P}\left((\mathbb{E}[z]-\bar{z})^2 > \epsilon\right),
\end{aligned}$$

and $z_i - \mathbb{E}[z]$ are sub-Gaussian with mean 0, from Lemma 22 by choosing $\boldsymbol{A} = \frac{1}{n}\mathbf{I}_p$ and $\boldsymbol{z}^\tau = (z_1 - \mathbb{E}[z], .., z_p - \mathbb{E}[z])$, we have

$$\mathbb{P}\left(|\frac{1}{n}\sum_i (z_i-\mathbb{E}[z])^2 - \mathbb{E}[(z-\mathbb{E}[z])^2]| > \epsilon\right) \leq 2\exp\left(-C\frac{\sqrt{n}\epsilon}{K^2}\right), \tag{93}$$

since $\sqrt{n}\epsilon \to \infty$.

The following follows from the usual deviation argument:

$$\mathbb{P}\left((\mathbb{E}[z]-\bar{z})^2 > \epsilon\right) \leq C_1 \exp\left(-C_2 n\epsilon\right).$$

Combining with the estimate (93), we know there exists positive constants C_1 and C_2 such that

$$\mathbb{P}\left(|\frac{1}{n}\sum_i (z_i-\bar{z})^2 - var(z)| > \epsilon\right) \leq C_1 \exp\left(-C_2\frac{\sqrt{n}\epsilon}{K^2}\right),$$

for sufficiently large n since $\sqrt{n}\epsilon \to \infty$. □

9.4. *Proof of Lemma 14.* We only need to prove this lemma for n i.i.d. sample y_i's from a uniform distribution over $[0,1]$. We slightly change the notation of order statistics $y_{(i)}$ to $y_{(i,n)}$ so that we can keep track of the sample size. Since y is uniform distribution on $[0,1]$, it is well known that $y_{(i,n)} \sim Beta(i, n-i+1)$ with expectation $\frac{i}{n+1}$ and mode $\frac{i-1}{n-1}$. Lemma 14 is a direct corollary of the following lemma.

LEMMA 23. *Suppose there are $n = Hc$ i.i.d. samples from uniform distribution over $[0,1]$, when H,c are sufficiently large, we have the following large deviation inequalities of $y_{(kc,Hc)}$, $k = 1, \cdots, (H-1)$.*

i) There exists a positive constant C, such that for any $\epsilon > \frac{1}{Hc-1}$, one has

$$\mathbb{P}\left(y_{(kc,Hc)} > \frac{k}{H} + \epsilon\right) \leq CH\sqrt{Hc+1}\exp\left(-(Hc+1)\frac{\epsilon^2}{2}\right);$$

ii) There exists a positive constant C, such that for any $\epsilon > \frac{2}{Hc-1}$, one has

$$\mathbb{P}\left(y_{(kc,Hc)} < \frac{k}{H} - \epsilon\right) \leq CH\sqrt{Hc+1}\exp\left(-(Hc+1)\frac{\epsilon^2}{8}\right);$$

iii) Let $\delta(k,H,c) = |y_{((k-1)c,Hc)} - y_{(kc,Hc)}|$, for $2 \leq k \leq H-1$, $\delta(1,H,c) = |y_{(c,Hc)}|$ and $\delta(H,H,c) = |1 - y_{((H-1)c,Hc)}|$. There exists a positive constant C, such that for any $\epsilon > \frac{4}{Hc-1}$, one has for any $1 \leq k \leq H$:

$$\mathbb{P}\left(|\delta(k,H,c) - \frac{1}{H}| > \epsilon\right) \leq CH\sqrt{Hc+1}\exp\left(-(Hc+1)\frac{\epsilon^2}{32}\right).$$

We will prove lemma 23 later. Assuming it, we have

$$\begin{aligned}\mathbb{P}\left(E(\epsilon)\right) &\leq \sum_{k=1}^{H}\mathbb{P}\Big(|\delta(k,H,c) - \frac{1}{H}| > \epsilon\Big) \\ &\leq CH^2\sqrt{Hc+1}\exp\left(-(Hc+1)\frac{\epsilon^2}{32}\right).\end{aligned}$$

□

9.4.1. *Proof of Lemma 23. The first part:* For any $1 \leq k \leq H-1$, we note that

$$\begin{aligned}&\mathbb{P}\left(y_{(kc,Hc)} > \frac{k}{H} + \epsilon\right) \leq \mathbb{P}\left(y_{(kc,Hc)} > \frac{kc}{Hc+1} + \epsilon\right) \\ &= \frac{1}{B(kc, Hc-kc+1)}\int_{x > \frac{kc}{Hc+1}+\epsilon}^{1} x^{kc-1}(1-x)^{Hc-kc}dx.\end{aligned}$$

When $\epsilon > \frac{1}{Hc-1}$, we know the mode $x_M = \frac{kc-1}{Hc-1} < x_D \triangleq \frac{kc}{Hc+1} + \epsilon$, so we have

$$\begin{aligned}\mathbb{P}\left(y_{(kc,Hc)} > \frac{k}{H} + \epsilon\right) &\leq \frac{(x_D)^{kc-1}(1-x_D)^{Hc-kc+1}}{B(kc, Hc-kc+1)} \\ &\leq H\frac{(x_D)^{kc}(1-x_D)^{Hc-kc+1}}{B(kc, Hc-kc+1)}.\end{aligned} \tag{94}$$

The last inequality due to $Hx_D \geq 1$. If $\epsilon + \frac{k}{H} \geq 1$, then $\mathbb{P}(y_{(kc,Hc)} > \frac{k}{H} + \epsilon) = 0$ and Lemma 23 holds automatically.

We may assume that $\epsilon + \frac{k}{H} < 1$ below. Let us denote the right hand side of the inequality (94) by A, then

$$\begin{aligned}\log(A) =& \log(H) + kc\log(E+\epsilon) + (Hc - kc + 1)\log(1 - E - \epsilon) \\ &+ \log(Hc+1)! - \log(kc)! - \log(Hc - kc + 1)! \\ &- \log(Hc+1) + \log(kc) + \log(Hc - kc + 1),\end{aligned}$$

where $E = \frac{kc}{Hc+1}$. Recall the following form of Stirling formula:

$$\log(n!) = n\log(n) - n + \frac{1}{2}\log(2\pi n) + O(\frac{1}{n}).$$

So, when m is sufficiently large, we have:

$$\begin{aligned}\log(A) =& \log(H) + (Hc+1)(E\log(1 + \frac{\epsilon}{E}) + (1-E)\log(1 - \frac{\epsilon}{1-E})) \\ &- \frac{1}{2}(\log(Hc+1) - \log(kc) - \log(Hc - kc + 1)) \\ &- \frac{1}{2}\log(2\pi) + O(\frac{1}{kc}) + O(\frac{1}{Hc - kc + 1}) \\ \leq& \log(H) - \frac{(Hc+1)\epsilon^2}{2(1-E)} - \frac{1}{2}\log(2\pi) \\ &- \frac{1}{2}(\log(Hc+1) - \log(kc) - \log(Hc - kc + 1)) + O(\frac{1}{c}) \\ \leq& \log(H) - \frac{(Hc+1)\epsilon^2}{2(1-E)} - \frac{1}{2}(\log(Hc+1) - \log(kc) - \log(Hc - kc + 1)),\end{aligned} \tag{95}$$

where we use the fact $\frac{kc}{Hc+1} \leq E + \epsilon < 1$ and the following elementary lemma which can be proved by Taylor expansion:

LEMMA 24. *Suppose x,y are positive number such that x+y=1, then for any $0 < \epsilon < y$, we have:*

$$x\log(1 + \frac{\epsilon}{x}) + y\log(1 - \frac{\epsilon}{y}) \leq -\frac{\epsilon^2}{2y}.$$

Now we know that there exists a positive constant C such that for any

$1 \le k \le H-1$ and for any $\epsilon > \frac{1}{Hc-1}$, the following holds:

$$\begin{aligned}
&\mathbb{P}\Big(y_{(kc,Hc)} > \frac{k}{H} + \epsilon\Big) \\
&\qquad \le CH\sqrt{\frac{(kc)(Hc-kc+1)}{Hc+1}} \exp\left(-(Hc+1)\frac{\epsilon^2}{2(1-E)}\right) \\
&\qquad \le CH\sqrt{Hc+1} \exp\left(-(Hc+1)\frac{\epsilon^2}{2(1-E)}\right) \\
&\qquad \le CH\sqrt{Hc+1} \exp\left(-(Hc+1)\frac{\epsilon^2}{2}\right).
\end{aligned}$$

The last inequality follows from $\frac{\epsilon^2}{1-E} \ge \epsilon^2$ since $\frac{1}{H+1} \le E \le \frac{H-1}{H}$.

The second part: The proof of the second part is similar. For completeness, we sketch some calculations below. For any $1 \le k \le H-1$, when $\epsilon > \frac{2}{Hc-1}$, we have

$$\begin{aligned}
&\mathbb{P}\left(y_{(kc,Hc)} < \frac{k}{H} - \epsilon\right) \le \mathbb{P}\left(y_{(kc,Hc)} < \frac{kc}{Hc+1} - \epsilon/2\right) \\
&= \frac{1}{B(kc, Hc-kc+1)} \int_{x < \frac{kc}{Hc+1} - \epsilon} x^{kc-1}(1-x)^{Hc-kc} dx.
\end{aligned}$$

Since $\epsilon > \frac{1}{Hc-1}$, we know the mode $x_M = \frac{kc-1}{Hc-1} > x_{D'} \triangleq \frac{kc}{Hc+1} - \epsilon/2$, so we have

$$\begin{aligned}
\mathbb{P}\left(y_{(kc,Hc)} \le \frac{k}{H} - \epsilon\right) \le& \frac{(x_{D'})^{kc}(1-x_{D'})^{Hc-kc}}{B(kc, Hc-kc+1)} \\
\le& H\frac{(x_{D'})^{kc}(1-x_{D'})^{Hc-kc+1}}{B(kc, Hc-kc+1)}.
\end{aligned}$$

The last inequality due to $H(1-x_{D'}) \ge 1$.

The rest is similar to the first part. We have that for any $1 \le k \le H-1$ and for any $\epsilon > \frac{2}{Hc-1}$,

$$\mathbb{P}\left(y_{(kc,Hc)} < \frac{k}{H+1} - \epsilon\right) \le CH\sqrt{Hc+1} \exp\left(-(Hc+1)\frac{\epsilon^2}{8}\right). \tag{96}$$

The third part: The third part is a direct corollary of the first two

parts. Note that for any $2 \le k \le H-2$, for any $\epsilon > \frac{4}{Hc-1}$

$$\begin{aligned}
&\mathbb{P}\Big(|\delta(k,H,c) - \frac{1}{H}| > \epsilon\Big) \\
&\qquad = \mathbb{P}\left(\Big|y_{(k+1)c,Hc} - \frac{k+1}{H} - (y_{kc,Hc} - \frac{k}{H})\Big| > \epsilon\right) \\
&\qquad \le \mathbb{P}\left(|y_{(k+1)c,Hc} - \frac{k+1}{H}| > \frac{\epsilon}{2}\right) + \mathbb{P}\left(|y_{kc,Hc} - \frac{k}{H}| > \frac{\epsilon}{2}\right) \\
&\qquad \le CH\sqrt{Hc+1}\exp\left(-(Hc+1)\frac{\epsilon^2}{32}\right).
\end{aligned}$$

When k=1, we have

$$\begin{aligned}
\mathbb{P}\Big(|\delta(1,H,c) - \frac{1}{H}| > \epsilon\Big) &= \mathbb{P}\left(|y_{c,Hc} - \frac{1}{H}| > \epsilon\right) \\
&\le \mathbb{P}\left(y_{c,Hc} - \frac{1}{H} > \epsilon\right) + \mathbb{P}\left(y_{c,Hc} - \frac{1}{H} < -\epsilon\right) \\
&\le CH\sqrt{Hc+1}\exp\left(-(Hc+1)\frac{\epsilon^2}{8}\right) \\
&\le CH\sqrt{Hc+1}\exp\left(-(Hc+1)\frac{\epsilon^2}{32}\right).
\end{aligned}$$

When k=H, we have

$$\begin{aligned}
\mathbb{P}\Big(|\delta(H-1,H,c) - \frac{1}{H}| > \epsilon\Big) &= \mathbb{P}\left(|y_{(H-1)c,Hc} - \frac{H-1}{H}| > \epsilon\right) \\
&\le \mathbb{P}\left(y_{(k+1)c,Hc} - \frac{H-1}{H} > \epsilon\right) + \mathbb{P}\left(y_{(H-1)c,Hc} - \frac{H-1}{H} < -\epsilon\right) \\
&\le CH\sqrt{Hc+1}\exp\left(-(Hc+1)\frac{\epsilon^2}{8}\right) \\
&\le CH\sqrt{Hc+1}\exp\left(-(Hc+1)\frac{\epsilon^2}{32}\right).
\end{aligned}$$

□

References

R. Adamczak, O. Guédon, A. Litvak, A. Pajor, and N. Tomczak-Jaegermann. Smallest singular value of random matrices with independent columns. *Comptes Rendus Mathematique*, 346(15):853–856, 2008.

P. J. Bickel and E. Levina. Covariance regularization by thresholding. *The Annals of Statistics*, pages 2577–2604, 2008.

T. T. Cai, C. Zhang, and H. Zhou. Optimal rates of convergence for covariance matrix estimation. *The Annals of Statistics*, 38(4):2118–2144, 2010.

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.

R. D. Cook. Graphics for regressions with a binary response. *Journal of the American Statistical Association*, 91(435):983–992, 1996.

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.

Hengjian Cui, Runze Li, and Wei Zhong. Model-free feature screening for ultrahigh dimensional discriminant analysis. *Journal of the American Statistical Association*, (just-accepted):00–00, 2014.

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

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

I. M. Johnstone and A. Y. Lu. Sparse principal components analysis. 2004.

I. M. Johnstone and A. Y. Lu. On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486), 2009.

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

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

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

X. Luo, L. A. Stefanski, and D. D. Boos. Tuning variable selection procedures by adding noise. *Technometrics*, 48(2):165–175, 2006.

M. Rudelson and R. Vershynin. Hanson-wright inequality and sub-gaussian concentration. 2013.

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

R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Y. Wu, D. D. Boos, and L. A. Stefanski. Controlling variable selection by the addition of pseudovariables. *Journal of the American Statistical Association*, 102(477), 2007.

Z. Yu, L. Zhu, H. Peng, and L. Zhu. Dimension reduction and predictor selection in semiparametric models. *Biometrika*, page ast005, 2013.

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

L. Zhu and K. Fang. Asymptotics for kernel estimate of sliced inverse regression. *The Annals of Statistics*, 24(3):1053–1068, 1996.

L. Zhu, L. Li, R. Li, and L. Zhu. Model-free feature screening for ultrahigh-dimensional data. *Journal of the American Statistical Association*, 106(496), 2011a.

L. P. Zhu, L. Li, R. Li, and L. X. Zhu. Model-free feature screening for ultrahigh-dimensional data. *Journal of the American Statistical Association*, 106(496), 2011b.

L. X. Zhu and K. W. Ng. Asymptotics of sliced inverse regression. *Statistica Sinica*, 5(2): 727–736, 1995.

L. X. Zhu, B. Miao, and H. 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.

Qian Lin
Center of Mathematical Sciences
and Applications
Harvard University
One Oxford Street
Cambridge, MA 02138
USA
E-mail: qianlin@cmsa.fas.harvard.edu

Zhigen Zhao
Department of Statistics
Temple University
346 Speakman Hall
1810 N. 13th Street
Philadelphia, PENNSYLVANIA, 19122
USA
E-mail: zhaozhg@temple.edu

Jun S. Liu
Department of Statistics
Harvard University
1 Oxford Street
Cambridge, MA 02138
USA
E-mail: jliu@stat.harvard.edu

This figure "GraphOfRho.png" is available in "png" format from:

http://arxiv.org/ps/1507.03895v1