The Annals of Statistics
2018, Vol. 46, No. 2, 580–610
https://doi.org/10.1214/17-AOS1561

ON CONSISTENCY AND SPARSITY FOR SLICED INVERSE REGRESSION IN HIGH DIMENSIONS

BY QIAN LIN[*,†,1] ZHIGEN ZHAO[‡,2] AND JUN S. LIU[*,†,3]

Tsinghua University [*], *Harvard University*[†] *and 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 [*J. Amer. Statist. Assoc.* **86** (1991) 316–342]. 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 eigenspace of $\operatorname{var}(\mathbb{E}[\boldsymbol{x}|y])$, the covariance matrix of the conditional expectation. The desired dimension reduction space is then obtained by multiplying the inverse of the covariance matrix on the eigenspace. 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 an 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 (1992), Zhu and Ng (1995) and Zhu and Fang (1996). Later, Zhu, Miao and Peng (2006) have proved the consistency when $p = o(\sqrt{n})$, a condition also appearing in two recent work Zhong et al. (2012) and Jiang and Liu (2014). When $p > n$, a common strategy pursued by recent researchers is to make

Received July 2015; revised January 2017.
[1]Supported in part by the Center of Mathematical Sciences and Applications at Harvard University.
[2]Supported in part by NSF Grant DMS-12-08735 and NSF Grant IIS-16-33283.
[3]Supported in part by NSF Grants DMS-11-20368 and DMS-16-13035.
MSC2010 subject classifications. Primary 62J02; secondary 62H25.
Key words and phrases. Dimension reduction, random matrix theory, sliced inverse regression.

sparsity assumptions that only a few predictors play a role in explaining and predicting y and apply various regularization methods. For instance, Li (2007), Li and Nachtsheim (2006) 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, Forzani and Rothman (2012). In this paper, we prove that, under certain technical assumptions, the SIR estimator is consistent if and only if $\rho = \lim \frac{p}{n} = 0$. This behavior 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. In this paper we provide a new framework to study the phase transition behavior of SIR. The technical tools developed here can be extended to study the phase transition behavior of other SDR estimators. The phase transition phenomenon provides theoretical justifications for imposing certain structural assumptions such as sparsity in high-dimensional settings.

The second part of this paper aims at extending the original SIR to the scenario with ultra-high dimension [i.e., $p = o(\exp(n^{\xi}))$]. Based on equation (3) in Section 2, the central space can be estimated in two steps: (i) obtain $\widehat{\boldsymbol{V}}_H$, the SIR estimate of $\operatorname{var}(\mathbb{E}[\boldsymbol{x}|y])$ as the top d eigenvectors of $\widehat{\boldsymbol{\Lambda}}_H$; (ii) estimate the precision matrix of $\boldsymbol{x}$ as $\widehat{\boldsymbol{\Sigma}}^{-1}$, and estimate the central subspace as $\widehat{\boldsymbol{\Sigma}}^{-1}\operatorname{col}(\widehat{\boldsymbol{V}}_H)$, where $\operatorname{col}(\widehat{\boldsymbol{V}}_H)$ represents the subspace formed by the column vectors of $\widehat{\boldsymbol{V}}_H$. The phase transition phenomenon indicates that $\operatorname{col}(\widehat{\boldsymbol{V}}_H)$ is not a consistent estimate of $\operatorname{col}(\operatorname{var}(\mathbb{E}[\boldsymbol{x}|y]))$ when $\lim \frac{p}{n} \neq 0$. Thus, we select variables according to the diagonal elements of $\widehat{\Lambda}_H$ and then estimate $\operatorname{col}(\operatorname{var}(\mathbb{E}[\boldsymbol{x}|y]))$ by applying SIR to these selected variables. We name this procedure as **D**iagonal **T**hresholding **SIR** (DT-SIR), and have shown that DT-SIR is consistent in estimating the central space under certain regularity conditions. Extensive simulation studies have demonstrated that DT-SIR performs better than its competitors and is computationally efficient.

The rest of the paper is organized as follows. In Section 2, we briefly describe the SIR procedure and introduce the notation. In Section 3, we 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 A, B and the Supplementary Material [Lin, Zhao and Liu (2018)].

2. Preliminaries and notation.

2.1. *Sliced inverse regression.* Consider the multiple index model

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

where $\boldsymbol{x} \in \mathbb{R}^p$, ϵ is the noise and f is an unknown link function. Without loss of generality, we assume that $\mathbb{E}[\boldsymbol{x}] = 0 \in \mathbb{R}^p$. Although the $p \times d$ matrix $\boldsymbol{V} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d)$ is not identifiable, the space spanned by the $\boldsymbol{\beta}$'s, which is called the column space of $\boldsymbol{V}$ and denoted as $\operatorname{col}(\boldsymbol{V})$, might be. Li (1991) proposed the *Sliced Inverse Regression* (SIR) procedure to estimate the central space $\operatorname{col}(\boldsymbol{V})$ without knowing $f(\cdot)$. SIR can be summarized as follows: given n i.i.d. samples $(y_i, \boldsymbol{x}_i)$, $i = 1, \ldots, n$, first divide them into H equal-sized slices according to the order statistics $y_{(i)}$.[4] Re-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 a sample in the hth slice, that is,

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

Here, $\boldsymbol{x}_{(k)}$ is the concomitant of $y_{(k)}$. Let $\overline{\boldsymbol{x}}_{h,\cdot}$ be the sample mean of the hth slice, and $\overline{\overline{\boldsymbol{x}}}$ be the overall mean of all the data. Then $\boldsymbol{\Lambda} \triangleq \operatorname{var}(\mathbb{E}[\boldsymbol{x}|y])$ can be estimated by

$$\widehat{\boldsymbol{\Lambda}}_H = \frac{1}{H} \sum_{h=1}^{H} \bar{\boldsymbol{x}}_{h,\cdot} \bar{\boldsymbol{x}}_{h,\cdot}^\tau. \tag{2}$$

Based on the observation that

$$\operatorname{col}(\boldsymbol{\Lambda}) = \boldsymbol{\Sigma}_x \operatorname{col}(\boldsymbol{V}), \tag{3}$$

the central space $\operatorname{col}(\boldsymbol{V})$ is estimated as $\widehat{\boldsymbol{\Sigma}}_x^{-1} \operatorname{col}(\widehat{\boldsymbol{V}}_H)$ where $\widehat{\boldsymbol{V}}_H$ is the matrix formed by the top d eigenvectors of $\widehat{\boldsymbol{\Lambda}}_H$. Throughout the paper we assume that d is fixed and the dth largest eigenvalue λ_d of $\boldsymbol{\Lambda}$ is bounded away from 0 when $n, p \to \infty$.

For SIR to be consistent in estimating the central space, Li (1991) imposed the following two conditions:

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

[4] To ease notation and arguments, we assume that $n = cH$ and $H = o(\log(n) \wedge \log(p))$ throughout the paper.

2.2. *Further notation.* Let S_h be the hth 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, 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. For $\mathcal{I} \subset \{1, \ldots, n\}$, $\mathcal{J} \subset \{1, \ldots, p\}$ and a $n \times p$ matrix $\boldsymbol{A}$, $\boldsymbol{A}^{\mathcal{I},\mathcal{J}}$ denotes the $|\mathcal{I}| \times |\mathcal{J}|$ submatrix formed by restricting the rows of $\boldsymbol{A}$ to $\mathcal{I}$ and columns to $\mathcal{J}$. In particular, $\boldsymbol{A}^{-,\mathcal{J}}$ denotes the submatrix formed by restricting the columns to $\mathcal{J}$. For any matrix $\boldsymbol{B} = \boldsymbol{A}^{\mathcal{I},\mathcal{J}} \in \mathbb{R}^{|\mathcal{I}| \times |\mathcal{J}|}$, let $e(\boldsymbol{B})$ be the embedded matrix into $\mathbb{R}^{p \times p}$ by putting 0 on entries outside $\mathcal{I} \times \mathcal{J}$. Similar notation are used for vectors. For two positive numbers a and b, let $a \vee b \equiv \max\{a, b\}$ and $a \wedge b \equiv \min\{a, b\}$. Let $\tau(x, t) = x \times 1(|x| > t)$ be the hard thresholding function. Throughout the paper, C, C_1 and C_2 denote generic absolute constants, though the actual value may vary from case to case. For a vector $\boldsymbol{x}$, the kth entry is denoted 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 $\{a_n\}$, $\{b_n\}$, $a_n \ll b_n$ stands for $a_n = O(b_n^{\epsilon})$ for some positive $\epsilon < 1$ and $a_n \succ b_n$ stands for $\lim \frac{b_n}{a_n} = 0$.

3. Consistency of SIR. To study the consistency of SIR, we impose the following boundedness condition **(A3)** on the predictors' covariance matrix in addition to the tail condition (sub-Gaussian) on their joint distribution. We also need a condition **(A4)** for the central curve.

- **(A3)** *Boundedness condition*: $\boldsymbol{x}$ is sub-Gaussian, and there exist positive constants C_1, C_2 such that
$$C_1 \le \lambda_{\min}(\boldsymbol{\Sigma}_{\boldsymbol{x}}) \le \lambda_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{x}}) \le C_2,$$
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.
- **(A4)** The central curve $\boldsymbol{m}(y) \triangleq \mathbb{E}[\boldsymbol{x}|y]$ has finite fourth moment and is ϑ-sliced stable (defined below) with respect to y and $\boldsymbol{m}(y)$.

DEFINITION 1. For any two positive constants $\boldsymbol{\gamma}_1 < 1 < \boldsymbol{\gamma}_2$, let $\mathcal{A}_H(\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2)$ be the collection of all the partitions $-\infty = a_0 < a_1 < \cdots < a_{H-1} < a_H = \infty$ of $\mathbb{R}$ satisfying that
$$\frac{\boldsymbol{\gamma}_1}{H} \le P(a_i \le y < a_{i+1}) \le \frac{\boldsymbol{\gamma}_2}{H}.$$
The central curve $\boldsymbol{m}(y) = \mathbb{E}[\boldsymbol{x}|y]$ is called ϑ-sliced stable with respect to y for some $\vartheta > 0$ if there exist positive constants $\boldsymbol{\gamma}_i, i = 1, 2, 3$ such that for any $\boldsymbol{\beta}$ in $\mathbb{R}^p$ and any partition in $\mathcal{A}_H(\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2)$,

$$\frac{1}{H}\left|\sum_{h=0}^{H-1} \operatorname{var}\left(\boldsymbol{\beta}^{\tau}\boldsymbol{m}(y) \mid a_h \le y \le a_{h+1}\right)\right| \le \frac{\boldsymbol{\gamma}_3}{H^{\vartheta}} \operatorname{var}\left(\boldsymbol{\beta}^{\tau}\boldsymbol{m}(y)\right) \tag{4}$$

for sufficiently large H. The central curve is sliced stable if it is ϑ-sliced stable for some positive constant ϑ.

REMARK 1. Note that we only need (4) to hold for all unit vectors $\mathbb{R}^p$ by rescaling. In particular, we have the following two useful properties of the slice-stability:

(i) By choosing $\boldsymbol{\beta}^\tau = (0, \ldots, 0, 1, 0, \ldots, 0)$ with 1 at the kth position, we have

$$\left| \sum_{h=0}^{H} \operatorname{var}\big(\boldsymbol{m}(y,k) \mid a_h \le y \le a_{h+1}\big) \right| \le \boldsymbol{\gamma}_3 H^{1-\vartheta} \operatorname{var}\big(\boldsymbol{m}(y,k)\big),$$

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

(ii) Since equation (4) holds for all unit vector $\boldsymbol{\beta}$, we have

$$\left\| \sum_{h=0}^{H} \operatorname{var}\big(\boldsymbol{m}(y) \mid a_h \le y \le a_{h+1}\big) \right\|_2 \le \boldsymbol{\gamma}_3 H^{1-\vartheta} \big\| \operatorname{var}\big(\boldsymbol{m}(y)\big) \big\|_2.$$

The sliced stable condition is satisfied by a large family of distributions. Here are some examples:

(i) If y is Gaussian, then y $(= \mathbb{E}[y|y])$ is sliced stable with respect to y. In fact, let $Y \sim N(0,1)$, then $\mathbb{E}[Y^4] < \infty$ and $y^4 \mathbb{P}(Y \ge y) \to 0$ as $y \to \infty$. We now prove that Y is $\frac{1}{2}$-sliced stable with respect to Y. Let us fix two positive constants $\gamma_1 < 1 < \gamma_2$. We want to prove that for any partition $-\infty = a_0 < \cdots < a_H = \infty$ satisfying $\frac{\gamma_1}{H} \le \mathbb{P}(a_i < Y < a_{i+1}) \le \frac{\gamma_2}{H}$, we have

$$\frac{1}{H} \sum_{h=0}^{H-1} \operatorname{var}(Y | a_h \le Y \le a_{h+1}) \le \frac{1}{H^{1/2}}.$$

To avoid tedious notation, we only prove it for the partition $\{a_j\}$ where a_j is the j/Hth quantile of Y, that is, $\mathbb{P}(Y \le a_j) = j/H$. It is easy to verify that

$$\operatorname{var}(Y | a_j \le Y \le a_{j+1}) \le \begin{cases} (a_{j+1} - a_j)^2 & \text{if } 1 \le j \le H-2, \\ o(1)\sqrt{H} & \text{if } j = 0, H-1. \end{cases}$$

Thus,

$$\frac{1}{H} \sum_{h=2}^{H-2} \operatorname{var}(Y | a_h \le Y \le a_{h+1}) \le \frac{1}{H}(a_1 - a_{H-1})^2 \le \frac{4}{H} \max\{a_1^2, a_{H-1}^2\}.$$

Since $a_1^4 P(Y < a_1) \to 0$, we know that $a_1^2 = o(\sqrt{H})$. Same argument shows that $a_{H-1}^2 = o(\sqrt{H})$. To summarize, we have

$$\frac{1}{H} \sum_{h=0}^{H-1} \operatorname{var}(Y | a_h \le Y \le a_{h+1}) \le \frac{1}{H^{1/2}}. \tag{5}$$

As a direct corollary, if $m(\cdot)$ is a function with bounded first derivative, then $m(Y)$ is $\frac{1}{2}$-sliced stable with respect to Y.

(ii) Let $y = \beta^\tau \boldsymbol{x} + \epsilon$, where $\boldsymbol{x}$ follows a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$ and ϵ is a normal error. Simple calculation shows that

$$\mathbb{E}[\boldsymbol{x}|y] = \frac{y\boldsymbol{\Sigma}\boldsymbol{\beta}}{\operatorname{var}(y)}. \tag{6}$$

Thus, $\mathbb{E}[\boldsymbol{x}|y]$ is a vector governed by the Gaussian random variable y and is sliced stable with respect to y according to example (i).

(iii) If y is bounded, then for any function $m(\cdot)$ with continuous first derivative, $m(y)$ is sliced stable with respect to y because $\operatorname{var}(m(y)|y \in [a,b]) \le C|a-b|^2$.

(iv) If $m(y)$ is sliced stable with respect to y, then for any monotone transform $y = g(z)$, $m(g(z))$ is sliced stable with respect to z because $\operatorname{var}(m(g(z))|z \in [a,b]) = \operatorname{var}(m(y)|y \in [g(a), g(b)])$ and $\mathbb{P}(g(a) \le Y \le g(b)) = \mathbb{P}(a \le z \le b)$. Especially, assume that $Y = f(\beta^\tau x + \epsilon)$ where $f(\cdot)$ is a monotone function, $\boldsymbol{x}$ is multivariate Gaussian and ϵ is a normal error. Then $m(y) = \mathbb{E}[\boldsymbol{x}|y]$ is sliced stable. In fact, let $z = f^{-1}(Y)$, then $n(z) = \mathbb{E}[\boldsymbol{x}|z]$ is sliced stable according to (ii). Thus $m(y) = n(f^{-1}(y))$ is sliced stable.

REMARK 2. Suppose $\mathbb{E}[\boldsymbol{m}(y)] = 0$ and 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 similarly to $\boldsymbol{x}_{h,i}$ and $\overline{\boldsymbol{x}}_{h,\cdot}$, respectively. On one hand, we have the classic consistent estimator $\frac{1}{n}\sum_i \boldsymbol{m}_i \boldsymbol{m}_i^\tau$ of $\operatorname{var}(\boldsymbol{m}(y))$. On the other hand, a necessary condition that the slice-based estimate $\frac{1}{H}\sum_h \overline{\boldsymbol{m}}_{h,\cdot}\overline{\boldsymbol{m}}_{h,\cdot}^\tau$ of $\operatorname{var}(\boldsymbol{m}(y))$ is consistent is the average loss of variance in each slice decreases to zero as H increases, that is,

$$\frac{1}{H}\sum_h \bar{\boldsymbol{m}}_{h,\cdot}\bar{\boldsymbol{m}}_{h,\cdot}^\tau - \frac{1}{n}\sum_i \boldsymbol{m}_i\boldsymbol{m}_i^\tau = \frac{1}{H}\sum_h \frac{1}{c}\sum_i (\bar{\boldsymbol{m}}_{h,\cdot} - \bar{\boldsymbol{m}}_{h,i})^2 \to 0. \tag{7}$$

The slice-stability ensure the left-hand side converges to zero at a power rate of H. It would be easily seen that if $\boldsymbol{m}$ is smooth and y is compactly supported then (7) holds automatically. For general curve $\boldsymbol{m}$ and random variable y, the slice-stability is a condition on smoothness of the central curve $\boldsymbol{m}$ and the tail distribution of $\boldsymbol{m}(y)$. This is not surprising because the smoonthness and tail conditions are commonly assumed for the consistency of SIR estimate.

The most widely used smoothness and tail condition is the following one proposed by Hsing and Carroll (1992) [later used in Zhu, Miao and Peng (2006), Zhu and Ng (1995)]. For $B > 0$ and $n \ge 1$, let $\Pi_n(B)$ be the collection of all the n-point partitions $-B \le y_{(1)} \le \cdots \le y_{(n)} \le B$ of $[-B, B]$. First, they assumed that the central curve $\boldsymbol{m}(y)$ satisfies the following smoothness condition:

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

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

$$\begin{aligned} &\tilde{m}^4(y)P\big(|Y| > y\big) \to 0 \qquad \text{as } y \to \infty, \\ &\big\|\boldsymbol{m}(y) - \boldsymbol{m}(y')\big\|_2 \le \big|\tilde{m}(y) - \tilde{m}(y')\big| \qquad \text{for } y, y' \in (-\infty, -B_0) \cup (B_0, \infty). \end{aligned} \tag{8}$$

By changing the tail condition (8) to a slightly stronger condition $\mathbb{E}[\tilde{m}(y)^4] < \infty$, Neykov, Lin and Liu (2015) proved that the modified condition implies the slice-stability. Now, we are ready to state our main results.

THEOREM 1. *Under conditions* (**A1**), (**A2**), (**A3**) *and* (**A4**), *we have*

$$\|\widehat{\boldsymbol{\Lambda}}_H - \boldsymbol{\Lambda}\|_2 = O_P\bigg(\frac{1}{H^{\vartheta}} + \frac{H^2 p}{n} + \sqrt{\frac{H^2 p}{n}}\bigg). \tag{9}$$

As a direct consequence of Theorem 1, we observe that if $\rho = \lim_{n\to\infty} \frac{p}{n} = 0$, we may choose $H = \log(n/p)$ such that the right-hand side of equation (9) converges to 0. Thus, Theorem 1 implies that $\widehat{\boldsymbol{\Lambda}}_H$ is a consistent estimate of $\boldsymbol{\Lambda}$ if $\rho = 0$.

REMARK 3 (More on convergence rate). Note that the convergence rate in (9) depends on the choice of H. This may seem not very desirable at the first glance. However, what we are really interested is the convergence rate of $\operatorname{col}(\widehat{\boldsymbol{V}}_H)$ which actually does not depend on H. In fact,

$$\widehat{\boldsymbol{\Lambda}}_H - \boldsymbol{\Lambda} = (\widehat{\boldsymbol{\Lambda}}_H - P_{\operatorname{col}(\boldsymbol{\Lambda})}\widehat{\boldsymbol{\Lambda}}_H P_{\operatorname{col}(\boldsymbol{\Lambda})}) + (P_{\operatorname{col}(\boldsymbol{\Lambda})}\widehat{\boldsymbol{\Lambda}}_H P_{\operatorname{col}(\boldsymbol{\Lambda})} - \boldsymbol{\Lambda}). \tag{10}$$

From the proof of Theorem 1, we can easily check that the first term is of convergence rate $\frac{pH^2}{n} + \sqrt{\frac{pH^2}{n}}$ and the second term is of rate $\frac{1}{H^{\vartheta}}$. Since $P_{\operatorname{col}(\boldsymbol{\Lambda})}\widehat{\boldsymbol{\Lambda}}_H P_{\operatorname{col}(\boldsymbol{\Lambda})}$ and $\boldsymbol{\Lambda}$ share the same column space, if we are only interested in estimating $P_{\operatorname{col}(\boldsymbol{\Lambda})}$, then the convergence rate of the second term does not matter provided that H is a large enough integer, which may depend on ϑ and $\boldsymbol{\gamma}_3$ but does not depend on n and p. For such an H, if $\mathcal{A}_H(\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2)$ is nonempty, Theorem 1 and (10) hold for both categorical and continuous response variable Y.

EXAMPLE 1. We consider a toy example to show that the convergence rate of $\widehat{\boldsymbol{\Lambda}}_H$ is different $\operatorname{col}(\widehat{\boldsymbol{\Lambda}}_H)$. Consider the following noiseless toy model:

$$y = 1 * \boldsymbol{x}(1) + 0 * \boldsymbol{x}(2) + 0 * \epsilon, \tag{11}$$

where $\boldsymbol{x}(1), \boldsymbol{x}(2)$ and $\epsilon \sim N(0, 1)$. It is easy to see that $\boldsymbol{\Lambda} = \operatorname{var}(\mathbb{E}[\boldsymbol{x}|y]) = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right)$, and its SIR estimate $\widehat{\boldsymbol{\Lambda}}_H(i, j) = \frac{1}{H}\sum_h \boldsymbol{x}_{h,\cdot}(i)\boldsymbol{x}_{h,\cdot}(j)$, where $\boldsymbol{x}_{h,\cdot}(1) = y_{h,\cdot}$ and $\boldsymbol{x}_{h,\cdot}(2) \sim N(0, \frac{1}{c})$. To ensure $\widehat{\boldsymbol{\Lambda}}_H$ is a consistent estimate of $\boldsymbol{\Lambda}$,

$$\big|\widehat{\boldsymbol{\Lambda}}_H(1,1) - 1\big| \to 0 \quad \text{and} \quad \big|\widehat{\boldsymbol{\Lambda}}_H(i,j)\big| \to 0 \qquad \forall (i,j) \neq (1,1). \tag{12}$$

In order for $\widehat{\mathbf{\Lambda}}_H(1,1)$ to be a consistent estimate of 1, we may need $H \to \infty$. However, if we are only interested in the first eigenvector (the basis of the central subspace in this toy model) of $\mathbf{\Lambda}$, we only need that $\widehat{\mathbf{\Lambda}}_H(1,1) - 1$ is sufficiently small. In summary, to get a consistent estimate of the central subspace using SIR, H must be large enough, but finite.

THEOREM 2. *Under conditions* (**A1**), (**A2**), (**A3**), (**A4**) *and assuming that* $\rho = \lim \frac{p}{n} = 0$, *we have*

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

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

The proofs of Theorems 1 and 2 are in Appendix B. We define the distance $\mathcal{D}(\boldsymbol{V}_1, \boldsymbol{V}_2)$ of two d-dimensional subspaces $\boldsymbol{V}_1$ and $\boldsymbol{V}_2$ as the operator norm (or Frobenius norm) of the difference between $P_{\boldsymbol{V}_1}$ and $P_{\boldsymbol{V}_2}$. Simple linear algebra shows that if the $\widetilde{\boldsymbol{\beta}}_i$'s satisfy $\mathbf{\Sigma}_{\boldsymbol{x}}\widetilde{\boldsymbol{\beta}}_i = \lambda_i \mathbf{\Lambda}\widetilde{\boldsymbol{\beta}}_i$, then

$$\operatorname{col}(\boldsymbol{V}) = \operatorname{span}\{\widetilde{\boldsymbol{\beta}}_1, \ldots, \widetilde{\boldsymbol{\beta}}_d\}.$$

Let $\widehat{\boldsymbol{V}}$ be the matrix formed by the top d generalized eigenvectors of $(\widehat{\mathbf{\Sigma}}_{\boldsymbol{x}}^{-1}, \widehat{\mathbf{\Lambda}}_H)$. Recall that the dth eigenvalue of $\mathbf{\Lambda}$ is assumed to be bounded away from 0. Therefore, Theorem 2 implies that $\mathcal{D}(P_{\widehat{\boldsymbol{V}}}, P_{\boldsymbol{V}}) \to 0$ when $\rho = 0$.

We have already shown that 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. *Under conditions* (**A1**), (**A2**), (**A4**) *and assuming that* $\boldsymbol{x} \sim N(0, \mathbf{I}_p)$ *for the single index model*

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

we have:

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

(ii) *Let* $\widehat{\boldsymbol{\beta}}$ *be the principal eigenvector of the SIR estimator* $\widehat{\mathbf{\Lambda}}_H$. *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 converging to one.

The proof is left in Appendix C in the Supplementary Material [Lin, Zhao and Liu (2018)]. We illustrate this result via a numerical study of the linear model

$$y = \boldsymbol{x}^\tau \boldsymbol{\beta} + \epsilon \qquad \text{where } \boldsymbol{\beta}^\tau = (1, 0, \ldots, 0), \boldsymbol{x} \sim N(0, \mathbf{I}_p), \epsilon \sim N(0, 1). \tag{13}$$

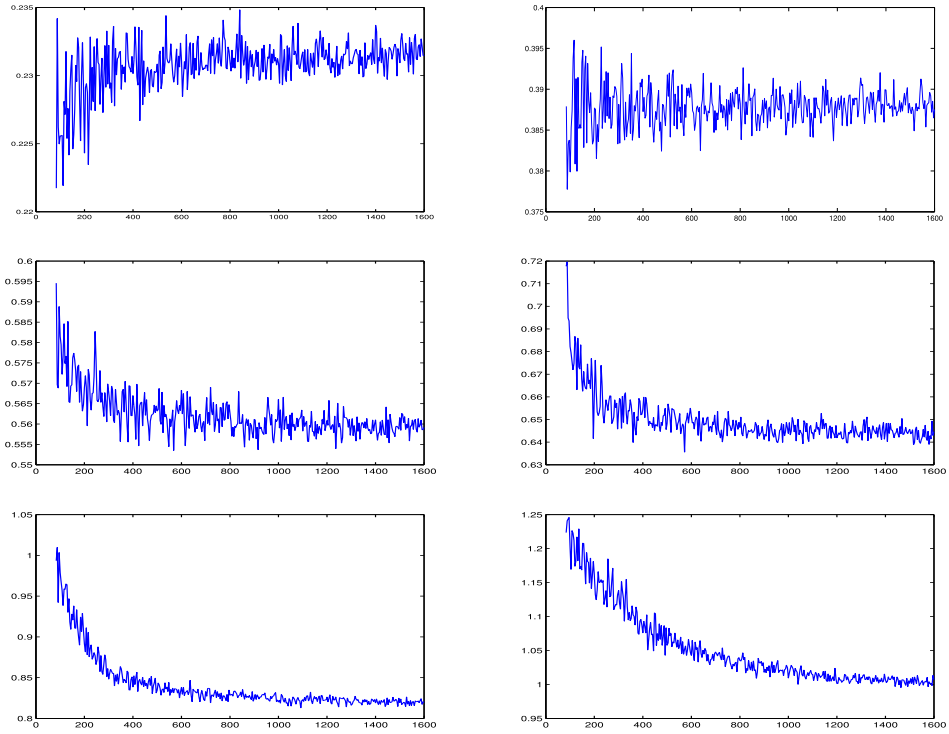

FIG. 1. *Numerical approximations of* $\mathbb{E}\angle(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta})$ *for model* (13) *as a function of dimension* p *for* $\rho = 0.1, 0.3, 0.7, 1, 2$ *and* 4, *respectively* (*upper left*, *upper right*, *middle left*, *middle right*, *lower left*, *lower right*), *where* $\widehat{\boldsymbol{\beta}}$ *is estimated by SIR*.

Figure 1 shows how $\mathbb{E}\angle(\boldsymbol{\beta}, \widehat{\boldsymbol{\beta}})$ is related to the dimension p for fixed ratio $\rho = \frac{p}{n}$ (taking values in $\{0.1, 0.3, 0.7, 1, 2, 4\}$), where $\widehat{\boldsymbol{\beta}}$ is calculated using SIR with the slice number $H = 10$. For each p, $\mathbb{E}\angle(\boldsymbol{\beta}, \widehat{\boldsymbol{\beta}})$ is calculated based on 100 iterations. It is seen that this expected angle converges to a positive number when the ratio ρ is nonzero. 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 ρ approaches zero, and increases when ρ increases.

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 when $p > n$.

4. SIR in ultra-high dimension. As we have shown in Section 3, the SIR estimator is not consistent when $\rho = \lim \frac{p}{n} \neq 0$. Hence, when $p \gg n$, some structural assumptions are necessary for getting a consistent estimate of the central space. In

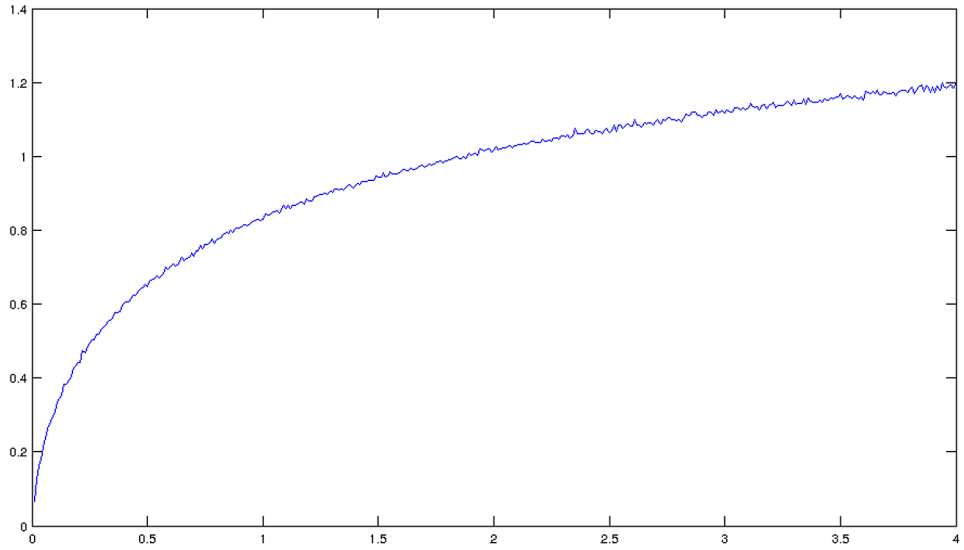

FIG. 2. *The relationship of* $\mathbb{E}\angle(\boldsymbol{\beta}, \widehat{\boldsymbol{\beta}})$ *and the ratio* p/n *where* $\widehat{\boldsymbol{\beta}}$ *is estimated by SIR.*

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 future work. For $\boldsymbol{\beta}_i$'s, we impose the following prevalent sparsity condition:

- **(A5)** $s = |\mathcal{S}| \ll p$ where $\mathcal{S} = \{i \mid \boldsymbol{\beta}_j(i) \neq 0 \text{ for some } j, 1 \leq j \leq d\}$ 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, Zhang and Zhou (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 notation and arguments, we choose a slightly stronger condition:

- **(A6)** $\boldsymbol{\Sigma}_{\boldsymbol{x}} \in \mathcal{U}(\epsilon_0, \alpha, C)$ and $\max_{1 \leq i \leq p} r_i$ is bounded where r_i is the number of nonzero elements in the ith row of $\boldsymbol{\Sigma}_{\boldsymbol{x}}$.

Let $\mathcal{T} = \{k \mid \operatorname{var}(\mathbb{E}[\boldsymbol{x}(k)|y]) \neq 0\}$. If $k \in \mathcal{T}$, there exists $\boldsymbol{\eta} \in \operatorname{col}(\boldsymbol{\Lambda})$ such that $\boldsymbol{\eta}(k) \neq 0$. Since we have (3),

$$\boldsymbol{\Sigma}_{\boldsymbol{x}} \operatorname{col}(\boldsymbol{V}) = \operatorname{col}(\boldsymbol{\Lambda}),$$

there exists a $\boldsymbol{\beta} \in \operatorname{col}(\boldsymbol{V})$ such that $\boldsymbol{\eta} = \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{\beta}$. Thus if $k \in \mathcal{T}$, then $k \in \operatorname{supp}(\boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{\beta})$ for some $\boldsymbol{\beta} \in \operatorname{col}(\boldsymbol{V})$. In particular, with the above sparsity assumptions **(A5)** and

(A6), we have $|\mathcal{T}| \le s \max_{1\le i\le p} r_i = O(s)$.[5] Note that our goal here is to recover the column space $\mathrm{col}(V)$ rather than $\mathcal{S}$. The key for recovering $\mathrm{col}(V)$ is to consistently recover the set $\mathcal{T}$.

At the population level, $\mathrm{var}(\mathbb{E}(\boldsymbol{x}(k))|y)$ can separate $\mathcal{T}$ from $\mathcal{T}^c$. When there are only finite samples, we use

$$\mathrm{var}_H\big(\boldsymbol{x}(k)\big) = \frac{1}{H}\sum_{h=1}^{H} \bar{\boldsymbol{x}}_{h,\cdot}(k)^2 \tag{14}$$

as an estimate of $\mathrm{var}(\mathbb{E}(\boldsymbol{x}(k))|y)$. These are the diagonal elements of the matrix $\widehat{\boldsymbol{\Lambda}}_H$. Note that these quantities depend on the sliced sample means, which are neither independent nor identically distributed. Thus, the usual concentration inequalities for χ^2 are no longer applicable. We have thus developed the concentration inequalities accordingly which is one of the main technical contributions of this paper, and can be further generalized.

REMARK 4. The link function $f(\cdot)$ is not used explicitly in the definition of $\mathrm{var}_H(\boldsymbol{x}(k))$. This nonparametric characteristic of the method is of particular interest to us and will be further investigated in future researches. Screening statistics inspired by the sliced inverse regression idea have been proposed in various formats, such as those in Jiang and Liu (2014), Zhu et al. (2011) and Cui, Li and Zhong (2015).

With the quantities $\mathrm{var}_H(\mathbb{E}[\boldsymbol{x}(k)|y])$'s, we define the inclusion set $\mathcal{I}_p(t)$ and the exclusion set $\mathcal{E}_p(t)$ below, which depend on a thresholding value t:

$$\mathcal{I}_p(t) = \big\{k \mid \mathrm{var}_H\big(\boldsymbol{x}(k)\big) > t\big\}$$

and

$$\mathcal{E}_p(t) = \big\{k \mid \mathrm{var}_H\big(\boldsymbol{x}(k)\big) \le t\big\}.$$

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 such that p/n is sufficiently small, the SIR estimator $\widehat{\boldsymbol{\Lambda}}^{\widehat{\mathcal{T}},\widehat{\mathcal{T}}}$ is a consistent estimate of $\boldsymbol{\Lambda}^{\mathcal{T},\mathcal{T}}$. Let $\widehat{\boldsymbol{V}}^{\widehat{\mathcal{T}}}$ be the matrix formed by the top d eigenvectors of $\widehat{\boldsymbol{\Lambda}}^{\widehat{\mathcal{T}},\widehat{\mathcal{T}}}$. We then use $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}\mathrm{col}(e(\widehat{\boldsymbol{V}}^{\widehat{\mathcal{T}}}))$ to estimate the central space $\mathrm{col}(\boldsymbol{V})$, where $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}$ is a consistent estimate of $\boldsymbol{\Sigma}_{\boldsymbol{x}}$. Estimating the covariance matrix and precision matrix in a high-dimensional setting is a challenging problem by itself and is not a main focus of this paper. We just employ the methods of Bickel and Levina (2008). In summary, we propose the following **D**iagonal **T**hresholding screening **SIR** (DT-SIR) algorithm (see Algorithm 1):

[5]We could introduce $\xi = \max_{1\le i\le p} r_i$, then $|\mathcal{T}| \le s\xi$. The arguments below still work, except we might need $s\xi = o(p)$.

Algorithm 1 DT-SIR

1. Calculate $\mathrm{var}_H(\boldsymbol{x}(k))$ according (14) for $k = 1, 2, \ldots, p$;
2. Let $\widehat{\mathcal{T}} = \{k \mid \mathrm{var}_H(\boldsymbol{x}(k)) > t\}$ for an appropriate t;
3. Let $\widehat{\boldsymbol{\Lambda}}_H^{\widehat{\mathcal{T}},\widehat{\mathcal{T}}}$ be the SIR estimator of the conditional covariance matrix for the data $(y, \boldsymbol{x}^{-,\widehat{\mathcal{T}}})$ according to equation (2);
4. Let $\widehat{\boldsymbol{V}}^{\widehat{\mathcal{T}}}$ be the matrix formed by the top d eigenvectors of $\widehat{\boldsymbol{\Lambda}}^{\widehat{\mathcal{T}},\widehat{\mathcal{T}}}$;
5. $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}\,\mathrm{col}(e(\widehat{\boldsymbol{V}}^{\widehat{\mathcal{T}}}))$ is the estimate of $\mathrm{col}(\boldsymbol{V})$

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

- (**S1**) $\exists C > 0$ and $\omega > 0$ such that $\mathrm{var}(\mathbb{E}[\boldsymbol{x}(k)|y]) > Cs^{-\omega}$ when $\mathbb{E}[\boldsymbol{x}(k)|y]$ is not a constant.

THEOREM 4. *Under conditions* **(A1)**–**(A6)** *and* **(S1)**, *and let* $t = as^{-\omega}$ *for some constant* $a > 0$ *such that* $t < \frac{1}{2}\mathrm{var}(m(y,k)), \forall k \in \mathcal{T}$, *we have*:

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

$$1 - C_1 \exp\left(-C_2 \frac{n}{H^2 s^\omega} + C_3 \log(H) + \log(p-s)\right); \tag{15}$$

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

$$1 - C_4 \exp\left(-C_5 \frac{n}{H^2 s^\omega} + C_6 \log(H) + \log(s)\right), \tag{16}$$

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

This theorem has a simple implication. If $\frac{n}{s^\omega} \succ \log(p) + \log(s)$, we may choose $H = \log(\frac{n}{s^\omega \log(p)})$, so that

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

Thus, we know $\mathcal{T} = \mathcal{I}_p$ with probability converging to one. Next, we have results for the consistency of DT-SIR.

THEOREM 5. *Under the same assumptions and choosing the same* t *as Theorem* 4, *if* $\frac{n}{s^\omega} \succ \log(p) + \log(s)$, *we have*

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

with probability converging to one, where $\widehat{\mathcal{T}} = \mathcal{I}(t)$ *and* $H = \log(\frac{n}{s^\omega \log(p)})$.

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, we have*

$$\|\widehat{\mathbf{\Sigma}}_{\boldsymbol{x}}^{-1} e(\widehat{\mathbf{\Lambda}}_H^{\widehat{\mathcal{T}},\widehat{\mathcal{T}}}) - \mathbf{\Sigma}_{\boldsymbol{x}}^{-1}\mathbf{\Lambda}_p\|_2 \to 0 \qquad \text{as } n \to \infty$$

with probability converging to one.

The proofs of Theorems 4 to 6 are left in Appendix D in the Supplementary Material [Lin, Zhao and Liu (2018)].

5. Simulation studies. 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 $\epsilon_i \overset{\text{i.i.d.}}{\sim} N(0, 1)$;
- **Setting II.** $y_i = \sum_{j=1}^{7} x_{ij} * \exp(x_{i8} + x_{i9}) + 0.5 * \epsilon_i$ where $\epsilon_i \overset{\text{i.i.d.}}{\sim} N(0, 1)$;
- **Setting III.** $y_i = \sum_{j=1}^{10} x_{ij} * \exp(\sum_{i=11}^{20} x_{ij}) + \epsilon_i$ where $\epsilon_i \overset{\text{i.i.d.}}{\sim} N(0, 1)$.

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 $\epsilon_i \overset{\text{i.i.d.}}{\sim} N(0, 1)$ and $\mathbf{\Sigma} = (\sigma_{ij})$ is tri-diagonal 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 $\epsilon_i \overset{\text{i.i.d.}}{\sim} N(0, 1)$, 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 tri-diagonal 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 VII.** Assume the same setting as in Setting V except that $\mathbf{\Sigma} = (\sigma_{ij})$ is given as $\sigma_{ij} = \rho^{|i-j|}$.

DT-SIR first screens all the predictors according to the statistic $\text{var}_H(\boldsymbol{x}(k))$, which requires a tuning parameter t. We chose t by using an auxiliary variable method based on an idea first proposed by Luo, Stefanski and Boos (2006) and extended by Wu, Boos and Stefanski (2007) and Zhu et al. (2011). 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 simulation studies. It is known that $\boldsymbol{y}$ and $\boldsymbol{z}$ are independent. The threshold t can be chosen as

$$\widehat{t} = \max_{1 \le k \le p'} \{\text{var}_H(\boldsymbol{z}(k))\}.$$

In DT-SIR, when $n > 1000$, H is chosen as 20; when $n \le 1000$, H is chosen as 10 in the screening step and 20 in the SIR step.

We also consider the following alternative methods in the screening step: Sure Independent Ranking and Screening (SIRS) in Zhu et al. (2011), SIR for variable selection via Inverse modeling (SIRI) in Jiang and Liu (2014), and trace pursuit

in Yu, Dong and Zhu (2016). As a comparison, we also considered two screening methods that are not based on the sliced regression: Distance correlation in Székely, Rizzo and Bakirov (2007) and SURE independence in Fan and Lv (2008). For SIRS, the threshold is chosen according to the auxiliary statistic (2.9) of Zhu et al. (2011). For SIRI, the predictors are chosen according to 10-fold cross validation. The threshold values $\bar{c}^{\mathrm{SIR}}$ and $\underline{c}^{\mathrm{SIR}}$ are chosen as the 10th and 5th quantile of a weighted χ^2 distribution given in Theorem 3.1 of Yu, Dong and Zhu (2016). In both SURE and DC screening, the top $\lfloor \gamma n \rfloor$ predictors where $\gamma = 0.01$ are kept for subsequent analyses.

After the screening step, similar to DT-SIR, we then applied the SIR algorithm (steps 3–5 of DT-SIR) to estimate $\mathrm{col}(\boldsymbol{V})$. These alternative methods are denoted as SIRS-SIR, SIRI-SIR, SURE-SIR, DC-SIR and TP-SIR, respectively, in the following discussions. Another method that we compared with is the sparse SIR, abbreviated as SpSIR, proposed in Li (2007). After obtaining an estimator $\mathrm{col}(\widehat{\boldsymbol{V}})$, we calculated $\mathcal{D}(P_{\mathrm{col}(\hat{V})}, P_{\mathrm{col}(V)})$ as a measure of the estimation error. We replicated this step 100 times, and calculated the average distance for the estimation result from each method and reported these numbers in Tables 1–3. For each setting, the average distance of the optimal method is highlighted using bold fonts. We further ran a two-sample T-test to test if the actual estimation error of each method is significantly different from that of the best method for that example at 1% level of significance.

Under all settings, the average distance obtained by DT-SIR is much smaller than that obtained by SpSIR and SURE-SIR. The p-values for comparing DT-SIR and SpSIR/SURE-SIR are all significant at the 0.01 level. 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 orthogonal to the true space spanned by β.

Under settings II–IV, DT-SIR performs either the best or not significantly worse than the best method. For all other cases, DT-SIR performs the best except for a few cases: Setting I when $n = 500$, $p = 1000$, setting V when $n = 500$, $p = 6000$, setting VI when $n = 500$, $p = 6000$ and setting VII when $n = 1000$, $p = 1000$.

When $p = 6000, n = 500$, both DT-SIR and SIRI-SIR are the winners. Under Setting III, DT-SIR performs better than SIRI-SIR; under settings V and VI, SIRI-SIR performs better than DT-SIR; under other settings, these two methods are comparable.

To graphically show the performance of various methods, we considered setting IV with $d = 1$. Consider two cases when $(n, p) = (2000, 1000)$ and $(n, p) = (500, 100)$. We calculated the estimated directions $\widehat{\boldsymbol{\beta}}$ using various methods and computed the angle between $\langle \widehat{\boldsymbol{\beta}} \rangle$ and $\langle \boldsymbol{\beta} \rangle$. We replicated this step 100 times to calculate the average angles for each method. The results are displayed in Figure 3, which shows clearly that DT-SIR performed better than its competitors.

Additionally, DT-SIR is computationally efficient. To show this, we reported the computing time for one replication under Setting II for various pairs of (n, p) in

TABLE 1

The average distance of the space estimated by each of the 7 *methods tested to the true space* $\mathrm{col}(\boldsymbol{V})$ *under various settings with* $p = 1000$. *The boldfaced number in each row represents the best result for that simulation scenario, and the "*" in cells represents that the p-value of the two-sample T-test comparing the estimation error of the corresponding method with that of the best method is less than* 0.01

	n	DT-SIR	SIRI-SIR	SIRS-SIR	SpSIR	SURE-SIR	DC-SIR	TP-SIR
I	500	0.655(*)	0.751(*)	**0.492**	2(*)	1.39(*)	0.731(*)	1.18(*)
	1000	**0.3**	0.431(*)	0.309	2(*)	1.29(*)	0.632(*)	0.94(*)
	2000	**0.221**	0.341(*)	0.226	1.58(*)	1.04(*)	0.655(*)	0.784(*)
	3000	0.167	0.245(*)	**0.149**	1.48(*)	0.816(*)	0.641(*)	0.713(*)
II	500	0.383	0.396	**0.371**	2(*)	1.64(*)	1.08(*)	0.389
	1000	0.235	**0.227**	0.256	2(*)	1.36(*)	0.266(*)	0.318(*)
	2000	0.161	**0.157**	0.189(*)	1.25(*)	1.25(*)	0.387(*)	0.264(*)
	3000	0.134	**0.129**	0.153(*)	0.975(*)	1.12(*)	0.404(*)	0.23(*)
III	500	1.15	1.48(*)	1.38(*)	2(*)	1.97(*)	1.85(*)	**1.13**
	1000	**0.426**	0.974(*)	0.596(*)	2(*)	1.94(*)	1.57(*)	0.429
	2000	**0.263**	0.403(*)	0.29(*)	1.33(*)	1.89(*)	0.996(*)	0.338(*)
	3000	**0.214**	0.297	0.238(*)	1.06(*)	1.82(*)	0.475(*)	0.299(*)
IV	500	0.263	**0.257**	0.333	1.41(*)	0.335(*)	0.334(*)	0.332(*)
	1000	**0.219**	0.447(*)	0.25	1.41(*)	0.436(*)	0.459(*)	0.469(*)
	2000	**0.161**	0.4(*)	0.196(*)	0.42(*)	0.442(*)	0.469(*)	0.452(*)
	3000	**0.134**	0.377(*)	0.177(*)	0.297(*)	0.43(*)	0.458(*)	0.438(*)
V	500	0.546	**0.529**	0.562	2(*)	1.62(*)	1.24(*)	1.09(*)
	1000	0.401	0.463(*)	0.514(*)	2(*)	1.15(*)	**0.367**	0.615(*)
	2000	**0.288**	0.418(*)	0.341(*)	1.51(*)	0.926(*)	0.569(*)	0.54(*)
	3000	**0.249**	0.399(*)	0.284(*)	1.24(*)	0.691(*)	0.597(*)	0.511(*)
VI	500	0.568	**0.535**	0.566	2(*)	1.64(*)	1.24(*)	1.08(*)
	1000	0.427	0.524(*)	0.548(*)	2(*)	1.22(*)	**0.39**	0.641(*)
	2000	**0.311**	0.469(*)	0.351(*)	1.51(*)	0.927(*)	0.598(*)	0.583(*)
	3000	**0.265**	0.456(*)	0.307(*)	1.25(*)	0.807(*)	0.622(*)	0.56(*)
VII	500	0.556	**0.534**	0.585(*)	2(*)	1.66(*)	1.26(*)	1.11(*)
	1000	0.436(*)	0.528(*)	0.545(*)	2(*)	1.22(*)	**0.39**	0.643(*)
	2000	**0.303**	0.465(*)	0.358(*)	1.51(*)	0.747(*)	0.589(*)	0.579(*)
	3000	**0.258**	0.468(*)	0.319(*)	1.25(*)	0.698(*)	0.63(*)	0.558(*)

Table 4. All computations were done on a computer with Intel Xeon(R) E5-1620 CPU@3.70G Hz and 16 GB memory. It is clearly seen that DT-SIR performed as fast as SURE-SIR, and both were much faster than other competitors. Consider the case when $p = 3000, n = 2000$. The computation time of DT-SIR is only 30 seconds; while that for DC-SIR is 21 minutes and 38 seconds, and the that for TP-SIR is 6 minutes and 17 seconds.

TABLE 2
The average distance of the space estimated by each of the 7 methods we tested to the true space $\mathrm{col}(\boldsymbol{V})$ *under various settings with* $n = 2000$

	p	DT-SIR	SIRI-SIR	SIRS-SIR	SpSIR	SURE-SIR	DC-SIR	TP-SIR
I	500	0.213	0.312(*)	**0.206**	1.44(*)	0.903(*)	0.629(*)	0.772(*)
	1000	**0.221**	0.341(*)	0.226	1.58(*)	1.04(*)	0.655(*)	0.784(*)
	2000	0.241	0.29	**0.214**	2(*)	1.07(*)	0.677(*)	0.793(*)
	3000	0.23	0.278	**0.218**	2(*)	1.17(*)	0.683(*)	0.797(*)
II	500	0.163	**0.16**	0.19(*)	0.83(*)	1.22(*)	0.369(*)	0.26(*)
	1000	0.161	**0.157**	0.189(*)	1.25(*)	1.25(*)	0.387(*)	0.264(*)
	2000	0.172	**0.159**	0.196(*)	2(*)	1.23(*)	0.404(*)	0.259(*)
	3000	0.164	**0.158**	0.199(*)	2(*)	1.3(*)	0.414(*)	0.261(*)
III	500	**0.272**	0.353	0.29(*)	0.916(*)	1.84(*)	0.846(*)	0.341(*)
	1000	**0.263**	0.403(*)	0.29(*)	1.33(*)	1.89(*)	0.996(*)	0.338(*)
	2000	**0.262**	0.368	0.285(*)	2(*)	1.92(*)	0.98(*)	0.339(*)
	3000	**0.269**	0.344	0.291(*)	2(*)	1.93(*)	1.09(*)	0.339(*)
IV	500	**0.145**	0.409(*)	0.182(*)	0.248(*)	0.406(*)	0.433(*)	0.438(*)
	1000	**0.161**	0.4(*)	0.196(*)	0.42(*)	0.442(*)	0.469(*)	0.452(*)
	2000	**0.16**	0.395(*)	0.198(*)	1.41(*)	0.472(*)	0.506(*)	0.447(*)
	3000	**0.15**	0.395(*)	0.216(*)	1.41(*)	0.49(*)	0.527(*)	0.447(*)
V	500	**0.272**	0.434(*)	0.353(*)	1.09(*)	0.876(*)	0.547(*)	0.539(*)
	1000	**0.288**	0.418(*)	0.341(*)	1.51(*)	0.926(*)	0.569(*)	0.54(*)
	2000	**0.289**	0.418(*)	0.351(*)	2(*)	0.868(*)	0.596(*)	0.537(*)
	3000	**0.3**	0.417(*)	0.372(*)	2(*)	0.968(*)	0.605(*)	0.544(*)
VI	500	**0.307**	0.479(*)	0.368(*)	1.1(*)	0.858(*)	0.566(*)	0.583(*)
	1000	**0.311**	0.469(*)	0.351(*)	1.51(*)	0.927(*)	0.598(*)	0.583(*)
	2000	**0.309**	0.461(*)	0.399(*)	2(*)	1.08(*)	0.617(*)	0.585(*)
	3000	**0.31**	0.46(*)	0.408(*)	2(*)	1(*)	0.638(*)	0.587(*)
VII	500	**0.299**	0.482(*)	0.343(*)	1.09(*)	0.818(*)	0.564(*)	0.583(*)
	1000	**0.303**	0.465(*)	0.358(*)	1.51(*)	0.747(*)	0.589(*)	0.579(*)
	2000	**0.309**	0.455(*)	0.383(*)	2(*)	0.966(*)	0.622(*)	0.578(*)
	3000	**0.308**	0.46(*)	0.357(*)	2(*)	0.858(*)	0.626(*)	0.58(*)

6. Conclusion. When the dimension p diverges to infinity, classical statistical procedures often fail unless additional structures such as sparsity conditions are imposed. Understanding boundary conditions of a statistical procedure provides us theoretical justification and practical guidance for our modeling efforts. In this paper we provide a new framework to show that $\rho = \lim \frac{p}{n}$ is the phase transition parameter for the SIR procedure. Under certain conditions, it is shown that the SIR estimator is consistent if and only if $\rho = 0$. When $\rho > 0$, where the original SIR fails to be consistent, we propose a two-stage method, DT-SIR for variable screening and selection in ultra-high dimension situations and show that

TABLE 3
The average distance of the space estimated by each of the 7 methods tested to the true space $\mathrm{col}(\boldsymbol{V})$ *under various settings with* $n = 500$ *and* $p = 6000$

	DT-SIR	SIRI-SIR	SIRS-SIR	SpSIR	SURE-SIR	DC-SIR	TP-SIR
I	0.694	0.631	**0.606**	2(*)	1.43(*)	0.97(*)	1.19(*)
II	0.446	0.462	0.414	2(*)	1.74(*)	1.08(*)	**0.4**
III	**1.35**	1.56(*)	1.56(*)	2(*)	1.99(*)	1.88(*)	1.37
IV	0.163	**0.122**	0.245(*)	1.41(*)	0.27(*)	0.305(*)	0.195(*)
V	0.481(*)	**0.431**	0.486(*)	2(*)	1.62(*)	1.1(*)	0.995(*)
VI	0.463(*)	**0.423**	0.494(*)	2(*)	1.62(*)	1.11(*)	0.999(*)
VII	0.44	**0.412**	0.477(*)	2(*)	1.61(*)	1.1(*)	1.03(*)

the method is 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.

It is natural to ask if similar phase transition phenomena occur for other SDR algorithms. For simplicity, let us assume that $\boldsymbol{x} \sim N(0, \mathbf{I})$. If we decompose $\boldsymbol{x} = P_{\mathcal{S}}\boldsymbol{x} + P_{\mathcal{S}^\perp}\boldsymbol{x}$, then $y \perp\!\!\!\perp P_{\mathcal{S}^\perp}\boldsymbol{x}$. The SIR procedure accumulates the signal along direction $P_{\mathcal{S}}\boldsymbol{x}$ and averages out the noise along direction $P_{\mathcal{S}^\perp}\boldsymbol{x}$. It is clear that if $\lim \frac{p}{n} \neq 0$, the averaging-out idea fails. Thus, we cannot expect that SIR can produce a consistent estimate of $\mathcal{S}$. This intuitive argument could apply to those SDR algorithms that inherit the sliced modeling characteristics. However, such a development relies on the higher moment and is technical challenging.

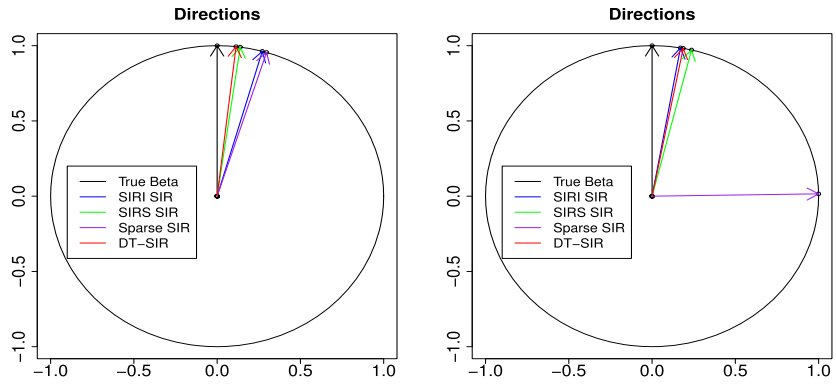

FIG. 3. *Simulated value of* $E\angle(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta})$ *for the various methods. Left panel*: $(n, p) = (2000, 1000)$. *Right panel*: $(n, p) = (500, 1000)$.

TABLE 4
Comparison of computing time under setting II

	DT-SIR	SIRI-SIR	SIRS-SIR	SpSIR	SURE-SIR	DC-SIR	TP-SIR
n				$p = 1000$			
500	$1''$	$1'12''$	$7''$	$11''$	$1''$	$24''$	$29''$
1000	$2''$	$2'2''$	$20''$	$11''$	$1''$	$1'52''$	$1'2''$
2000	$3''$	$3'27''$	$1'14''$	$13''$	$2''$	$7'38''$	$2'18''$
3000	$4''$	$4'59''$	$2'45''$	$15''$	$3''$	$6'51''$	$3'7''$
p				$n = 2000$			
500	$1''$	$2'48''$	$35''$	$2''$	$1''$	$3'46''$	$1'7''$
1000	$3''$	$3'27''$	$1'14''$	$13''$	$2''$	$7'38''$	$2'18''$
2000	$12''$	$4'55''$	$2'35''$	$1'39''$	$12''$	$14'24''$	$3'22''$
3000	$30''$	$6'0''$	$4'10''$	$5'19''$	$30''$	$21'38''$	$6'17''$

APPENDICES

The following two sections provide details about our theoretical derivations. But some more tedious intermediate steps (organized as Lemmas 6–21) are in the Supplementary Material, which is available on-line [Lin, Zhao and Liu (2018)].

APPENDIX A: THE KEY LEMMA

The following lemma plays an important role in developing the high-dimensional theory for sliced inverse regression. Here, keep in mind that H and v (if they are not constants) grow at very slow rate compared with c and n [e.g., polynomial of $\log(n)$]. Let $\boldsymbol{m}(y) = \mathbb{E}[\boldsymbol{x}|y]$, and $\boldsymbol{x} = \boldsymbol{m}(y) + \epsilon$. Notation $\boldsymbol{m}_{h,j}$, $\overline{\boldsymbol{m}}_{h,\cdot}$, $\overline{\overline{\boldsymbol{m}}}$ and $\boldsymbol{\epsilon}_{h,j}$, $\overline{\boldsymbol{\epsilon}}_{h,\cdot}$, $\overline{\overline{\boldsymbol{\epsilon}}}$ are similarly defined as $\boldsymbol{x}_{h,j}$, $\overline{\boldsymbol{x}}_{h,\cdot}$ and $\overline{\overline{\boldsymbol{x}}}$ that were introduced before.

LEMMA 1. *Assume the conditions* **(A1)**, **(A2)**, **(A3)** *and* **(A4)** *hold. Let* $\boldsymbol{x} \in \mathbb{R}^p$ *be a sub-Gaussian random variable which is upper exponentially bounded by* K (*see Definition* 4). *For any unit vector* $\boldsymbol{\beta} \in \mathbb{R}^p$, *let* $\boldsymbol{x}(\boldsymbol{\beta}) = \langle \boldsymbol{x}, \boldsymbol{\beta} \rangle$ *and* $\boldsymbol{m}(\boldsymbol{\beta}) = \langle \boldsymbol{m}, \boldsymbol{\beta} \rangle = \mathbb{E}[\boldsymbol{x}(\boldsymbol{\beta}) \mid y]$, *we have the following*:

(i) *If* $\operatorname{var}(\boldsymbol{m}(\boldsymbol{\beta})) = 0$, *there exists positive constants* C_1, C_2 *and* C_3 *such that for any* $b = O(1)$ *and sufficiently large* H, *we have*

$$\mathbb{P}\big(\operatorname{var}_H(\boldsymbol{x}(\boldsymbol{\beta})) > b\big) \le C_1 \exp\bigg(-C_2 \frac{nb}{H^2} + C_3 \log(H)\bigg).$$

(ii) *If* $\operatorname{var}(\boldsymbol{m}(\boldsymbol{\beta})) \neq 0$, *there exists positive constants* C_1, C_2 *and* C_3 *such that, for any* $v > 1$, *we have*

$$\big|\operatorname{var}_H(\boldsymbol{x}(\boldsymbol{\beta})) - \operatorname{var}(\boldsymbol{m}(\boldsymbol{\beta}))\big| \ge \frac{1}{2v} \operatorname{var}(\boldsymbol{m}(\boldsymbol{\beta}))$$

with probability at most

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

Here, we choose H *such that* $H^{\vartheta} > C_4 v$ *for some sufficiently large constant* C_4.

A.1. Proof of Lemma 1(i). If $\boldsymbol{m}(\boldsymbol{\beta}) = 0$ [or equivalently $\operatorname{var}(\boldsymbol{m}(\boldsymbol{\beta})) = 0$], since

$$\begin{aligned}\bar{\boldsymbol{\epsilon}}_{h,\cdot}(\boldsymbol{\beta})^2 &= \left(\frac{c-1}{c}\frac{1}{c-1}\sum_{i=1}^{c-1}\boldsymbol{\epsilon}_{h,i}(\boldsymbol{\beta}) + \frac{1}{c}\boldsymbol{\epsilon}_{h,c}(\boldsymbol{\beta})\right)^2 \\ &\le 2\left(\frac{1}{c-1}\sum_{i=1}^{c-1}\boldsymbol{\epsilon}_{h,i}(\boldsymbol{\beta})\right)^2 + 2\left(\frac{1}{c}\boldsymbol{\epsilon}_{h,c}(\boldsymbol{\beta})\right)^2\end{aligned}$$

for $h = 1, \ldots, H-1$ and $\bar{\boldsymbol{\epsilon}}_{H,\cdot}(\boldsymbol{\beta}) = \frac{1}{c}\sum_{i=1}^{c}\boldsymbol{\epsilon}_{H,i}(\boldsymbol{\beta})$, we have

$$\begin{aligned}&\operatorname{var}_H(\boldsymbol{x}(\boldsymbol{\beta})) - \operatorname{var}(\boldsymbol{m}(\boldsymbol{\beta})) \\ &= \frac{1}{H}\sum_{h}^{H-1}\bar{\boldsymbol{\epsilon}}_{h,\cdot}(\boldsymbol{\beta})^2 + \frac{1}{H}\bar{\boldsymbol{\epsilon}}_{H,\cdot}(\boldsymbol{\beta})^2 \\ &\le \frac{2}{H}\left(\sum_{h}^{H-1}\left(\frac{1}{c-1}\sum_{i=1}^{c-1}\boldsymbol{\epsilon}_{h,i}(\boldsymbol{\beta})\right)^2 + \bar{\boldsymbol{\epsilon}}_{H,\cdot}(\boldsymbol{\beta})^2\right) + \frac{2}{Hc^2}\sum_{h}^{H-1}\boldsymbol{\epsilon}_{h,c}(\boldsymbol{\beta})^2 \\ &\triangleq 2I + 2II.\end{aligned}$$

Thus

$$\mathbb{P}\left(\operatorname{var}_H(\boldsymbol{x}(\boldsymbol{\beta})) > b\right) \le \mathbb{P}(I > b/4) + \mathbb{P}(II > b/4). \tag{17}$$

Lemma 17(iii) in the Supplementary Material [Lin, Zhao and Liu (2018)] implies that

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

for some positive constant C. Since $\mathbb{E}[\boldsymbol{x}(\boldsymbol{\beta})|y] = 0$, we have $\mathbb{E}[\boldsymbol{x}(\boldsymbol{\beta})|y \in S_h] = 0$. From Lemma 9, we know that for $1 \le h \le H-1$, $\boldsymbol{\epsilon}_{h,i}(\boldsymbol{\beta})$ can be treated as $c-1$ i.i.d. samples from $\boldsymbol{\epsilon}(\boldsymbol{\beta})|_{y\in S_h}$. According to Lemma 17(iv),

$$\mathbb{P}\left(\left|\frac{1}{c-1}\sum_{i=1}^{c-1}\boldsymbol{\epsilon}_{h,i}(\boldsymbol{\beta})\right| > \sqrt{b}/2\right) \le C_1\exp\left(\frac{-b(c-1)}{8C_2HK^2 + 4\sqrt{b}K}\right).$$

Similarly, we have

$$\mathbb{P}\left(\left|\frac{1}{c}\sum_{i=1}^{c}\boldsymbol{\epsilon}_{H,i}(\boldsymbol{\beta})\right| > \sqrt{b}/2\right) \le C_1\exp\left(\frac{-bc}{8C_2HK^2 + 4\sqrt{b}K}\right).$$

Thus, if $b = O(1)$ and H is sufficiently large, we have

$$\begin{aligned}\mathbb{P}\Big(I > \frac{b}{4}\Big) &\le C_1\Big((H-1)\exp\Big(\frac{-b(c-1)}{8C_2HK^2 + 4\sqrt{b}K}\Big) \\ &\quad + \exp\Big(\frac{-bc}{8C_3HK^2 + 4\sqrt{b}K}\Big)\Big) \\ &\le C_1\exp\Big(-C_2\frac{cb}{H} + C_3\log(H)\Big)\end{aligned}$$

for some positive constants C_1, C_2 and C_3.

Since $\boldsymbol{\epsilon}_i(\boldsymbol{\beta})$ are i.i.d. samples from a sub-Gaussian distribution $\boldsymbol{\epsilon}(\boldsymbol{\beta})$ with mean 0 and upper-exponentially bounded by $2K$. Lemma 19 implies that if $b = O(1)$ and H is sufficiently large, we have

$$\begin{aligned}\mathbb{P}(II > b/4) &\le \mathbb{P}\Big(\frac{1}{n}\sum_i \boldsymbol{\epsilon}_i(\boldsymbol{\beta})^2 > bc/4\Big) \\ &\le \mathbb{P}\Big(\frac{1}{n}\sum_i \boldsymbol{\epsilon}_i(\boldsymbol{\beta})^2 - \mathbb{E}[\boldsymbol{\epsilon}(\boldsymbol{\beta})^2] > bc/4 - \mathbb{E}[\boldsymbol{\epsilon}(\boldsymbol{\beta})^2]\Big) \\ &\le \mathbb{P}\Big(\Big|\frac{1}{n}\sum_i \boldsymbol{\epsilon}_i(\boldsymbol{\beta})^2 - \mathbb{E}[\boldsymbol{\epsilon}(\boldsymbol{\beta})^2]\Big| \ge cb/4 - 4K^2\Big) \\ &\le C_1\exp\Big(-C_2\frac{\sqrt{n}(cb/4 - 4K^2)}{K^2}\Big) \\ &\le C_1\exp\Big(-C_2\frac{cb}{H} + C_3\log(H)\Big)\end{aligned}$$

for some positive constants C_1, C_2 and C_3 if H is sufficiently large. We used in above the fact that $\mathbb{E}[\boldsymbol{\epsilon}(\boldsymbol{\beta})^2] \le 4K^2$.

To summarize, if $b = O(1)$ and H is sufficiently large, we have

$$\mathbb{P}(\mathrm{var}_H(\boldsymbol{x}(\boldsymbol{\beta})) > b) \le C_1\exp\Big(-C_2\frac{cb}{H} + C_3\log(H)\Big)$$

for some positive absolute constants C_1, C_2 and C_3.

A.2. Proof of Lemma 1(ii). Let $\mu_h = \mathbb{E}[m(y \mid y \in S_h)]$. Since $\boldsymbol{x}$ is sub-Gaussian and $\boldsymbol{\beta}$ is unit vector, we know that $\mathrm{var}(\boldsymbol{m}(\boldsymbol{\beta})) = O(1)$. If $\boldsymbol{m}(\boldsymbol{\beta}) \ne 0$ [or equivalently $\mathrm{var}(\boldsymbol{m}(\boldsymbol{\beta})) \ne 0$], we have

$$\begin{aligned}\big|\mathrm{var}_H(\boldsymbol{x}(\boldsymbol{\beta})) - \mathrm{var}(\boldsymbol{m}(\boldsymbol{\beta}))\big| &= \Big|\frac{1}{H}\sum_h \overline{\boldsymbol{x}}_{h,\cdot}(\boldsymbol{\beta})^2 - \mathrm{var}(\boldsymbol{m}(\boldsymbol{\beta}))\Big| \\ &= \Big|\frac{1}{H}\sum_h \overline{\boldsymbol{m}}_{h,\cdot}(\boldsymbol{\beta})^2 + \frac{2}{H}\sum_h \overline{\boldsymbol{m}}_{h,\cdot}(\boldsymbol{\beta})\overline{\boldsymbol{\epsilon}}_{h,\cdot}(\boldsymbol{\beta})\end{aligned}$$

$$+\frac{1}{H}\sum_h \overline{\boldsymbol{\epsilon}}_{h,\cdot}(\boldsymbol{\beta})^2 - \operatorname{var}(\boldsymbol{m}(\boldsymbol{\beta}))\Big|$$
$$\le A_1 + A_2 + A_3 + A_4,$$

where

$$\begin{aligned}
A_1 &= \Big|\frac{1}{H}\sum_h \mu_h(\boldsymbol{\beta})^2 - \operatorname{var}(\boldsymbol{m}(\boldsymbol{\beta}))\Big|,\\
A_2 &= \frac{1}{H}\sum_h |\overline{\boldsymbol{m}}_{h,\cdot}(\boldsymbol{\beta})^2 - \mu_h(\boldsymbol{\beta})^2|,\\
A_3 &= \frac{1}{H}\sum_h \bar{\boldsymbol{\epsilon}}_{h,\cdot}(\boldsymbol{\beta})^2,\\
A_4 &= \Big(\frac{1}{H}\sum_h \overline{\boldsymbol{m}}_{h,\cdot}(\boldsymbol{\beta})^2\Big)^{1/2}\Big(\frac{1}{H}\sum_h \overline{\boldsymbol{\epsilon}}_{h,\cdot}(\boldsymbol{\beta})^2\Big)^{1/2}.
\end{aligned} \tag{18}$$

Lemma 1(ii) is a direct corollary of the following properties of A_i's.

LEMMA 2. *Let the A_i's be defined as in equation (18). There exist positive constants C_1, C_2 and C_3, such that for any $\nu > 1$ and H satisfying $H^{\vartheta} = N_1\nu$ for sufficiently large N_1, we have that each of the following events*:

(i) $\Theta_1 = \{A_1 \le \frac{1}{4\nu}\operatorname{var}(\boldsymbol{m}(\boldsymbol{\beta}))\}$,
(ii) $\Theta_2 = \{A_2 \le \frac{1}{8\nu}\operatorname{var}(\boldsymbol{m}(\boldsymbol{\beta}))\}$,
(iii) $\Theta_3 = \{A_3 \le \frac{1}{16\nu}\operatorname{var}(\boldsymbol{m}(\boldsymbol{\beta}))\}$,
(iv) $\Theta_4 = \{A_4 \le \frac{1}{16\nu}\operatorname{var}(\boldsymbol{m}(\boldsymbol{\beta}))\}$,

occurs with probability at least

$$1 - C_1\exp\Big(-C_2\frac{c\operatorname{var}(\boldsymbol{m}(\boldsymbol{\beta}))}{H\nu^2} + C_3\log(H)\Big). \tag{19}$$

A.2.1. *Proof of Lemma 2.*

A.2.1.1. Proof of (i). Recall definitions of the random intervals S_h, $h = 1, 2, \ldots, H$ and random variable $\delta_h = \delta_h(\omega) = \int_{y\in S_h(\omega)} f(y)\,dy$. We have

$$\begin{aligned}
&\Big|\frac{1}{H}\sum_h \big(\mu_h(\boldsymbol{\beta})\big)^2 - \operatorname{var}(\boldsymbol{m}(\boldsymbol{\beta}))\Big|\\
&\quad\le \Big|\operatorname{var}(\boldsymbol{m}(\boldsymbol{\beta})) - \sum_h \delta_h\big(\mu_h(\boldsymbol{\beta})\big)^2\Big| + \Big|\frac{1}{H}\sum_h \big(\mu_h(\boldsymbol{\beta})\big)^2 - \sum_h \delta_h\big(\mu_h(\boldsymbol{\beta})\big)^2\Big|\\
&\quad\triangleq B_1 + B_2.
\end{aligned}$$

Let $\epsilon = \frac{1}{Hn_0+1}$ where $n_0 = N_2\nu$ for some sufficiently large constant N_2 and let event $E(\epsilon)$ be defined as in Lemma 11 in Section E, that is, $E(\epsilon) = \{\omega \mid |\delta_h - \frac{1}{H}| > \epsilon, \forall h\}$. For any $\omega \in E(\epsilon)^c$, we have

$$B_1 = \sum_h \delta_h(\omega)\operatorname{var}\big(\boldsymbol{m}(\boldsymbol{\beta})|y \in S_h(\omega)\big)$$

$$\leq \left(\frac{1}{H} + \epsilon\right)\sum_h \operatorname{var}\big(\boldsymbol{m}(\boldsymbol{\beta})|y \in S_h(\omega)\big) \tag{20}$$

$$\leq (1 + H\epsilon)\frac{\boldsymbol{\gamma}_3}{H^{\vartheta}}\operatorname{var}\big(\boldsymbol{m}(\boldsymbol{\beta})\big) \tag{21}$$

$$\leq \frac{2\boldsymbol{\gamma}_3}{N_1\nu}\operatorname{var}\big(\boldsymbol{m}(\boldsymbol{\beta})\big), \tag{22}$$

where inequality (20) follows from the fact that $\delta_h(\omega) \leq \frac{1}{H} + \epsilon$, inequality (21) follows from the sliced stable condition (4) and inequality (22) follows from the requirement that $H^{\vartheta} > N_1\nu$, and the fact

$$B_2 \leq \epsilon\sum_h \big(\boldsymbol{\beta}^{\tau}\mu_h\big)^2 = \sum_h \frac{\epsilon}{\delta_h}\delta_h\big(\boldsymbol{\beta}^{\tau}\mu_h\big)^2$$

$$\leq \frac{H\epsilon}{1 - H\epsilon}\sum_h \delta_h\big(\boldsymbol{\beta}^{\tau}\mu_h\big)^2 \tag{23}$$

$$\leq \frac{2}{N_2\nu}\sum_h \delta_h\big(\boldsymbol{\beta}^{\tau}\mu_h\big)^2,$$

where inequality (23) follows from the fact $\delta_h \geq \frac{1}{H} - \epsilon$.

From (22), we observe that

$$\sum_h \delta_h\big(\mu_h(\boldsymbol{\beta})\big)^2 \leq \left(1 + \frac{2\boldsymbol{\gamma}_3}{N_1\nu}\right)\operatorname{var}\big(\boldsymbol{m}(\boldsymbol{\beta})\big). \tag{24}$$

Combining with (23), we then have

$$B_2 \leq \frac{2}{N_2\nu}\left(1 + \frac{2\boldsymbol{\gamma}_3}{N_1\nu}\right)\operatorname{var}\big(\boldsymbol{m}(\boldsymbol{\beta})\big).$$

So when $E(\epsilon)^c$ occurs, we have

$$B_1 + B_2 \leq \left(\frac{2\boldsymbol{\gamma}_3}{N_1\nu} + \frac{2}{N_2\nu}\left(1 + \frac{2\boldsymbol{\gamma}_3}{N_1\nu}\right)\right)\operatorname{var}\big(\boldsymbol{m}(\boldsymbol{\beta})\big).$$

Note that N_1 and N_2 can be chosen sufficiently large so that

$$B_1 + B_2 \leq \frac{4\boldsymbol{\gamma}_3}{N_1\nu}\operatorname{var}\big(\boldsymbol{m}(\boldsymbol{\beta})\big) \leq \frac{1}{4\nu}\operatorname{var}\big(\boldsymbol{m}(\boldsymbol{\beta})\big). \tag{25}$$

Consequently, conditioning on $E(\epsilon)^c$ where $\epsilon = \frac{1}{HN_2\nu+1}$, if we choose $H^\vartheta > N_1\nu$, then

$$\left|\frac{1}{H}\sum_h \big(\mu_h(\boldsymbol{\beta})\big)^2 - \operatorname{var}\big(\boldsymbol{m}(\boldsymbol{\beta})\big)\right| \le \frac{1}{4\nu}\operatorname{var}\big(\boldsymbol{m}(\boldsymbol{\beta})\big). \tag{26}$$

Since $\operatorname{var}(\boldsymbol{m}(\boldsymbol{\beta})) = O(1)$, $H^\vartheta > N_1\nu$ and $\epsilon = \frac{1}{HN_2\nu+1}$, the desired probability bound follows from Lemma 11, that is,

$$\begin{aligned}\mathbb{P}\big(E(\epsilon)\big) &\le C_1 \exp\bigg(-\frac{Hc+1}{32(Hn_0+1)^2} + \log\big(H^2\sqrt{Hc+1}\big)\bigg)\\ &\le C_1 \exp\bigg(-C_2\frac{c\operatorname{var}(\boldsymbol{m}(\boldsymbol{\beta}))}{H\nu^2} + C_3\log(H)\bigg),\end{aligned}$$

for some positive constants C_1, C_2 and C_3.

REMARK 5. From (26), conditioning on $E(\epsilon)^c$, we obtain the following two inequalities:

$$\frac{1}{H}\sum_h \big(\mu_h(\boldsymbol{\beta})\big)^2 \le \bigg(1 + \frac{4\boldsymbol{\gamma}_3}{H^\vartheta}\bigg)\operatorname{var}\big(\boldsymbol{m}(\boldsymbol{\beta})\big) \tag{27}$$

and

$$\frac{1}{H}\sum_h \big|\mu_h(\boldsymbol{\beta})\big| \le \bigg(\bigg(1 + \frac{4\boldsymbol{\gamma}_3}{H^\vartheta}\bigg)\operatorname{var}\big(\boldsymbol{m}(\boldsymbol{\beta})\big)\bigg)^{1/2}. \tag{28}$$

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

A.2.1.2. Proof of (ii). Denote $\frac{1}{c-1}\sum_{i=1}^{c-1}\boldsymbol{m}_{h,i}(\boldsymbol{\beta})$ by $\overline{\boldsymbol{m}}'_h(\boldsymbol{\beta})$ and $\overline{\boldsymbol{m}}_{H,\cdot}(\boldsymbol{\beta})$ by $\overline{\boldsymbol{m}}'_H(\boldsymbol{\beta})$, we have

$$\begin{aligned}A_2 &\le \frac{1}{H}\sum_{h=1}^H \big|\overline{\boldsymbol{m}}'_h(\boldsymbol{\beta})^2 - \mu_h(\boldsymbol{\beta})^2\big| + \frac{1}{Hc^2}\sum_{h=1}^H \boldsymbol{m}_{h,c}(\boldsymbol{\beta})^2\\ &\quad + \frac{2(c-1)}{c}\bigg(\frac{1}{H}\sum_{h=1}^H \overline{\boldsymbol{m}}'_h(\boldsymbol{\beta})^2\bigg)^{1/2}\bigg(\frac{1}{Hc^2}\sum_{h=1}^H \boldsymbol{m}_{h,c}(\boldsymbol{\beta})^2\bigg)^{1/2}\\ &\quad + \frac{2}{Hc}\sum_{h=1}^H \mu_h(\boldsymbol{\beta})^2\\ &\triangleq I + II + III + IV.\end{aligned}$$

Before we start proving this part, we need to introduce two events and bound their probabilities. First, let

$$E_1(N_3,\nu) = \bigg\{\eta(\boldsymbol{\beta}) > \frac{1}{N_3\nu}\sqrt{\operatorname{var}\big(\boldsymbol{m}(\boldsymbol{\beta})\big)}\bigg\}, \tag{29}$$

where $\eta(\boldsymbol{\beta}) = \max_{1\le h\le H}\{|\overline{\boldsymbol{m}}'_h(\boldsymbol{\beta}) - \mu_h(\boldsymbol{\beta})|\}$. According to Lemma 17(i), (iv) and Bonferroni's inequality, we have

$$\mathbb{P}\big(E_1(N_3,\nu)\big) \le 2H\exp\bigg(\frac{1}{(N_3\nu)^2}\frac{-(c-1)\operatorname{var}(\boldsymbol{m}(\boldsymbol{\beta}))}{2CHK^2 + \frac{2}{N_3\nu}\sqrt{\operatorname{var}(m(\boldsymbol{\beta}))}K}\bigg) \tag{30}$$

$$\le C_1\exp\bigg(-C_2\frac{c\operatorname{var}(\boldsymbol{m}(\boldsymbol{\beta}))}{H\nu^2} + C_3\log(H)\bigg) \tag{31}$$

for some positive constants C_1, C_2 and C_3. Second, let

$$E_2(N_4,\nu) \triangleq \bigg\{II > \frac{1}{N_4\nu}\operatorname{var}(\boldsymbol{m}(\boldsymbol{\beta}))\bigg\},$$

then

$$\begin{aligned}\mathbb{P}\big(E(N_4,\nu)\big) &\le \mathbb{P}\bigg(\frac{1}{nc}\sum_i m_i^2 > \frac{\operatorname{var}(\boldsymbol{m}(\boldsymbol{\beta}))}{N_4\nu}\bigg)\\ &\le C_1\exp\bigg(-C_2\sqrt{n}\bigg(c\frac{\operatorname{var}(\boldsymbol{m}(\boldsymbol{\beta}))}{\nu}\bigg) - K^2\bigg)\\ &\le C_1\exp\bigg(-C_2\frac{c\operatorname{var}(\boldsymbol{m}(\boldsymbol{\beta}))}{H\nu} + C_3\log(H)\bigg)\end{aligned}$$

for some positive constant C_1, C_2 and C_3. It is easy to see $E(N_4,\nu)\subset E(N_4,\nu^2)$.

For I. Conditioning on the event $E(\epsilon)^c\cap E_1(N_3,\nu)^c$, combining with (28), we have

$$\begin{aligned}I &\le \frac{1}{H}\sum_h \eta(\boldsymbol{\beta})\big(\eta(\boldsymbol{\beta}) + 2|\mu_h(\boldsymbol{\beta})|\big)\\ &\le \eta(\boldsymbol{\beta})^2 + \frac{2\eta(\boldsymbol{\beta})}{H}\sum_h|\mu_h(\boldsymbol{\beta})|\\ &\le \bigg(\bigg(\frac{1}{N_3\nu}\bigg)^2 + \frac{2}{N_3\nu}\bigg(1+\frac{4\boldsymbol{\gamma}_3}{H^{\vartheta}}\bigg)^{1/2}\bigg)\operatorname{var}(\boldsymbol{m}(\boldsymbol{\beta}))\\ &\le \frac{1}{32\nu}\operatorname{var}(\boldsymbol{m}(\boldsymbol{\beta}))\end{aligned}$$

if N_3 is sufficiently large.

REMARK 6. From above, conditioning on the event $E(\epsilon)^c\cap E_1(N_3,\nu)^c$, we have

$$\frac{1}{H}\sum_{h=1}^{H}\overline{\boldsymbol{m}}'(\boldsymbol{\beta})^2 \le \frac{1+32\nu}{32\nu}\operatorname{var}(\boldsymbol{m}(\boldsymbol{\beta})). \tag{32}$$

For II. Conditioning on $E_2(N_4,\nu)^c$, we have $II \le \frac{\operatorname{var}(\boldsymbol{m}(y))}{N_4\nu}$.

For III. When the event $E(\epsilon)^c \cap E_1(N_3, \nu)^c \cap E_2(N_4, \nu^2)^c$ occurs, according to equation (32),

$$III \le \frac{2(c-1)}{c}\sqrt{\frac{1+32\nu}{32\nu}}\frac{1}{\sqrt{N_4}\nu}\operatorname{var}(\boldsymbol{m}(\boldsymbol{\beta})) < \frac{1}{16\nu}\operatorname{var}(\boldsymbol{m}(\boldsymbol{\beta})),$$

if N_4 is sufficiently large.

For IV. When the event $E(\epsilon)^c \cap E_1(N_3, \nu)^c \cap E_2(N_4, \nu)^c$ occurs, from (26), we know

$$IV = \frac{2}{Hc}\sum_h \mu_h(\boldsymbol{\beta})^2 \le \frac{9}{4c}\operatorname{var}(\boldsymbol{m}(\boldsymbol{\beta})) < \frac{1}{16\nu}\operatorname{var}(\boldsymbol{m}(\boldsymbol{\beta})).$$

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

$$A_2 \le I + II + III + IV \le \frac{1}{8\nu}\operatorname{var}(\boldsymbol{m}(\boldsymbol{\beta}))$$

holds on the event $E(\epsilon)^c \cap E_1(N_3, \nu)^c \cap E_2(N_4, \nu^2)^c$, which is with probability at least

$$1 - C_1 \exp\left(-C_2\frac{c\operatorname{var}(\boldsymbol{m}(\boldsymbol{\beta}))}{H\nu^2} + C_3\log(H)\right)$$

for some positive constants C_1, C_2 and C_3.

A.2.1.3. *Proof of* (iii). Similar to the proof of Lemma 1(i), we have

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

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

$$A_3 \le \frac{1}{16\nu}\operatorname{var}(\boldsymbol{m}(\boldsymbol{\beta}))$$

with probability at least

$$1 - C_1 \exp\left(-C_2\frac{c\operatorname{var}(\boldsymbol{m}(\boldsymbol{\beta}))}{H\nu^2} + C_3\log(H)\right)$$

for some positive constant C_1, C_2 and C_3.

A.2.1.4. *Proof of* (iv). Let

$$D_1 \triangleq \frac{1}{H}\sum_h \overline{\boldsymbol{m}}_{h,\cdot}(\boldsymbol{\beta})^2,$$

$$D_2 \triangleq A_3 = \frac{1}{H}\sum_h \bar{\boldsymbol{\epsilon}}_{h,\cdot}(\boldsymbol{\beta})^2.$$

Consequently,

$$\begin{aligned}(33)\qquad &\mathbb{P}\Big(D_1^{1/2}D_2^{1/2} > \frac{1}{16\nu}\operatorname{var}(\boldsymbol{m}(\boldsymbol{\beta}))\Big)\\ &\quad\le \mathbb{P}\Big(|D_1| > \frac{2\nu+1}{2\nu}\operatorname{var}(\boldsymbol{m}(\boldsymbol{\beta}))\Big) + \mathbb{P}\Big(D_2 > \frac{\operatorname{var}(\boldsymbol{m}(\boldsymbol{\beta}))}{(2\nu+1)16\nu}\Big).\end{aligned}$$

Note that

$$\big|D_1 - \operatorname{var}(\boldsymbol{m}(\boldsymbol{\beta}))\big| \le A_2 + A_1.$$

According to (i) and (ii), the right-hand side of (33) is bounded by

$$C_1\exp\Big(-C_2\frac{c\operatorname{var}(\boldsymbol{m}(\boldsymbol{\beta}))}{H\nu^2} + C_3\log(H)\Big)$$

for some positive constants C_1, C_2 and C_3.

APPENDIX B: PROOFS OF THEOREMS IN SECTION 3

B.1. Proof of Theorem 1. We have the decomposition

$$\begin{aligned}(34)\qquad \boldsymbol{x} &= \boldsymbol{P}_{\operatorname{col}(\boldsymbol{\Lambda})}\boldsymbol{x} + \boldsymbol{P}_{\operatorname{col}(\boldsymbol{\Lambda})^\perp}\boldsymbol{x} \triangleq z + \boldsymbol{w}\\ &= \mathbb{E}[z|y] + z - \mathbb{E}[z|y] + \boldsymbol{w} \triangleq \boldsymbol{m} + \boldsymbol{v} + \boldsymbol{w},\end{aligned}$$

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

$$(35)\qquad \boldsymbol{m}_{h,j}, \overline{\boldsymbol{m}}_{h,\cdot}, \overline{\overline{\boldsymbol{m}}}, \qquad z_{h,j}, \overline{z}_{h,\cdot}, \overline{\overline{z}} \quad\text{and}\quad \boldsymbol{w}_{h,j}, \overline{\boldsymbol{w}}_{h,\cdot}, \overline{\overline{\boldsymbol{w}}}$$

similar to the definition of $\boldsymbol{x}_{h,j}$, $\overline{\boldsymbol{x}}_{h,\cdot}$ and $\overline{\overline{\boldsymbol{x}}}$. Consequently, we can define $\widehat{\boldsymbol{\Lambda}}_z$ and have the following decomposition:

$$(36)\qquad \widehat{\boldsymbol{\Lambda}}_H \equiv \frac{1}{H}\sum_h \overline{\boldsymbol{x}}_{h,\cdot}\overline{\boldsymbol{x}}_{h,\cdot}^\tau = \widehat{\boldsymbol{\Lambda}}_z + \mathcal{Z}\mathcal{W}^\tau + \mathcal{W}\mathcal{Z}^\tau + \mathcal{W}\mathcal{W}^\tau,$$

where

$$\mathcal{Z} = \frac{1}{\sqrt{H}}(\overline{z}_{1,\cdot}, \ldots, \overline{z}_{H,\cdot}) \quad\text{and}\quad \mathcal{W} = \frac{1}{\sqrt{H}}(\overline{\boldsymbol{w}}_{1,\cdot}, \ldots, \overline{\boldsymbol{w}}_{H,\cdot}).$$

We need to bound $\|\widehat{\boldsymbol{\Lambda}}_z - \boldsymbol{\Lambda}\|_2$ and $\|\mathcal{W}\mathcal{W}^\tau\|_2$.

LEMMA 3.

$$\|\mathcal{W}\mathcal{W}^\tau\|_2 \le O_P\Big(\frac{H^2 p}{n}\Big). \tag{37}$$

PROOF. For any unit vector $\boldsymbol{\beta} \perp \operatorname{col}(\boldsymbol{\Lambda})$, we have $\operatorname{var}(\boldsymbol{m}(\boldsymbol{\beta})) = 0$ and $\widehat{\operatorname{var}}_H(\boldsymbol{\beta}^\tau \boldsymbol{x}) = \boldsymbol{\beta}\widehat{\boldsymbol{\Lambda}}_H\boldsymbol{\beta} = \boldsymbol{\beta}^\tau\mathcal{W}\mathcal{W}^\tau\boldsymbol{\beta}$. From Lemma 1, we know

$$\mathbb{P}\Big(\boldsymbol{\beta}^\tau\mathcal{W}\mathcal{W}^\tau\boldsymbol{\beta} > C\frac{H^2 p}{n}\Big) \le C_1\exp(-C_2 p + \log(H)) \tag{38}$$

for some positive constants C_1 and C_2. Then the ε-net argument [see, e.g., Vershynin (2012)] implies that $\|\mathcal{W}\mathcal{W}^\tau\| \le O_P(\frac{H^2 p}{n})$. □

LEMMA 4.

$$\|\widehat{\boldsymbol{\Lambda}}_z - \boldsymbol{\Lambda}\| \le O_P\Big(\frac{1}{H^\vartheta}\Big). \tag{39}$$

As a direct corollary, we have $\|\widehat{\boldsymbol{\Lambda}}_z\| \le O_P(1)$.

PROOF. From Lemma 1, we have

$$\begin{aligned}&\mathbb{P}\Big(|\boldsymbol{\beta}^\tau(\widehat{\boldsymbol{\Lambda}}_z - \boldsymbol{\Lambda})\boldsymbol{\beta}| > \frac{C}{H^\vartheta}\|\boldsymbol{\Lambda}\|_2\Big)\\ &\qquad \le C_1\exp\Big(-C_2\frac{c\operatorname{var}(\boldsymbol{m}(\boldsymbol{\beta}))}{H^{1+2\vartheta}} + C_3\log(H)\Big).\end{aligned}$$

Note that we only need to verify it for $\boldsymbol{\beta} \in \operatorname{col}(\boldsymbol{\Lambda})$, which is a d-dimensional space. Then the ε-net argument implies that $\|\widehat{\boldsymbol{\Lambda}}_z - \boldsymbol{\Lambda}\|_2 \le O_P(\frac{1}{H^\vartheta})$. □

Theorem 1 follows from Lemma 4 and Lemma 3. In fact,

$$\begin{aligned}\|\widehat{\boldsymbol{\Lambda}}_H - \boldsymbol{\Lambda}\| &\le \|\widehat{\boldsymbol{\Lambda}}_z - \boldsymbol{\Lambda}\| + \|\mathcal{Z}\mathcal{W}^\tau + \mathcal{W}\mathcal{Z}^\tau\|_2 + \|\mathcal{W}\mathcal{W}^\tau\|_2\\ &\le O_P\Big(\frac{1}{H^\vartheta} + \sqrt{\frac{H^2 p}{n}} + \frac{H^2 p}{n}\Big).\end{aligned}$$

B.2. Proof of Theorem 2. Theorem 2 is a direct corollary of Theorem 1 and Lemma 13. In fact, we have

$$\|\widehat{\boldsymbol{\Sigma}}_X^{-1}\widehat{\boldsymbol{\Lambda}}_H - \boldsymbol{\Sigma}_X^{-1}\boldsymbol{\Lambda}\|_2 \le \|\widehat{\boldsymbol{\Sigma}}_X^{-1} - \boldsymbol{\Sigma}_X^{-1}\|_2\|\widehat{\boldsymbol{\Lambda}}_H\|_2 + \|\boldsymbol{\Sigma}_X^{-1}\|_2\|\widehat{\boldsymbol{\Lambda}}_H - \boldsymbol{\Lambda}\|_2,$$

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

B.3. Proof of Theorem 3. (i) The proof for part (i) is similar to the proof of Theorem 1 and the standard Gaussian assumption on $\boldsymbol{x}$ simplifies the argument and improves the results. 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 $\boldsymbol{\Sigma} = \operatorname{cov}(\boldsymbol{w})$. Using the decomposition (36), we may write

$$1 - 1 - \mathcal{W} = \frac{1}{\sqrt{Hc}}\boldsymbol{\Sigma}^{1/2}\boldsymbol{E}_{p\times H}, \tag{40}$$

where $\boldsymbol{E}_{p,H}$ is a $p \times H$ matrix with i.i.d. standard normal entries. Corollary 4 implies that

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

Lemma 4 implies

$$\|\widehat{\boldsymbol{\Lambda}}_z\|_2 \le \|\boldsymbol{\Lambda}\|_2 + O_P\left(\frac{1}{H^\vartheta}\right).$$

By the Cauchy inequality, we have

$$\|\mathcal{Z}\mathcal{W}^\tau\|_2^2 \le \|\widehat{\boldsymbol{\Lambda}}_z\|_2\|\mathcal{W}\mathcal{W}^\tau\|_2 \le O_P\left(\frac{p}{n}\right).$$

Thus,

$$\|\widehat{\boldsymbol{\Lambda}}_H - \boldsymbol{\Lambda}\|_2 \le O_P\left(\frac{1}{H^\vartheta} + \frac{p}{n} + \sqrt{\frac{p}{n}}\right).$$

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

(ii) The proof for part (ii) is similar to the proof of Theorem 2 in Johnstone and Lu (2009) but is technically more challenging. Let $D = \mathcal{Z}\mathcal{Z}^\tau + \mathcal{W}\mathcal{W}^\tau$ and $B = \mathcal{Z}\mathcal{W}^\tau + \mathcal{W}\mathcal{Z}^\tau$, then

$$\widehat{\boldsymbol{\Lambda}}_H = D + B.$$

Since we are working on single index model with $\boldsymbol{x}$ is standard normal, $\boldsymbol{z} = P_{\boldsymbol{\beta}}\boldsymbol{x} = \boldsymbol{\beta}z(y)$ for some scalar function $z(y)$ and $\boldsymbol{w} = P_{\boldsymbol{\beta}^\perp}\boldsymbol{x}$ are independent normal random variables. Let $\boldsymbol{\Sigma} = \operatorname{var}(\boldsymbol{w})$, then we can write

$$\mathcal{W} = \frac{1}{\sqrt{Hc}}\boldsymbol{\Sigma}^{1/2}\mathbf{E},$$

where $\mathbf{E}$ is a $p \times H$ matrix with i.i.d. standard normal entries.

Since $\boldsymbol{z} = \boldsymbol{\beta}z(y)$, we have $\mathcal{Z} = \frac{1}{\sqrt{H}}\boldsymbol{\beta}(\overline{z}_{1,\cdot}, \overline{z}_{2,\cdot}, \ldots, \overline{z}_{H,\cdot})$. To ease notation, let $\boldsymbol{\theta}^\tau = (\overline{z}_{1,\cdot}, \overline{z}_{2,\cdot}, \ldots, \overline{z}_{H,\cdot})$, then

$$\begin{aligned} D &= \frac{1}{H}\|\boldsymbol{\theta}\|^2\boldsymbol{\beta}\boldsymbol{\beta}^\tau + \frac{1}{n}\boldsymbol{\Sigma}^{1/2}\boldsymbol{E}\boldsymbol{E}^\tau\boldsymbol{\Sigma}^{1/2}, \\ B &= \boldsymbol{\beta}\boldsymbol{u}^\tau + \boldsymbol{u}\boldsymbol{\beta}^\tau \qquad \text{where } \boldsymbol{u} = \frac{1}{H\sqrt{c}}\boldsymbol{\Sigma}^{1/2}\boldsymbol{E}\boldsymbol{\theta}. \end{aligned} \tag{41}$$

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

$$N_\alpha = \{\boldsymbol{x} \in \mathbb{R}^p : \angle(\boldsymbol{x}, \boldsymbol{\beta}) \le \alpha \text{ and } \|x\| = 1\} \tag{42}$$

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 $\boldsymbol{y}$. In order to proceed, we need the following lemma.

LEMMA 5. *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 $\widehat{\boldsymbol{\beta}} \in N_\alpha$, that is, $\angle(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}) \le \alpha$, we have*

$$\angle(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\beta}}_-) \ge \frac{1}{3}\omega(\alpha) \tag{43}$$

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

PROOF. The proof is presented in Lin, Zhao and Liu (2018). □

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 $\{\angle(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}) \le \alpha\} \cup \{\angle(\widehat{\boldsymbol{\beta}}_-, \boldsymbol{\beta}) \le \alpha\}$, then

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

SUPPLEMENTARY MATERIAL

Supplement to "On the consistency and sparsity for sliced inverse regression for high dimensions" (DOI: 10.1214/17-AOS1561SUPP; .pdf). In the supplement, we prove the rest of the results stated in the paper.

REFERENCES

BICKEL, P. J. and LEVINA, E. (2008). Covariance regularization by thresholding. *Ann. Statist.* **36** 2577–2604. MR2485008

CAI, T. T., ZHANG, C.-H. and ZHOU, H. H. (2010). Optimal rates of convergence for covariance matrix estimation. *Ann. Statist.* **38** 2118–2144. MR2676885

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

COOK, R. D. (1996). Graphics for regressions with a binary response. *J. Amer. Statist. Assoc.* **91** 983–992. MR1424601

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

CUI, H., LI, R. and ZHONG, W. (2015). Model-free feature screening for ultrahigh dimensional discriminant analysis. *J. Amer. Statist. Assoc.* **110** 630–641. MR3367253

FAN, J. and LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 849–911. MR2530322

HSING, T. and CARROLL, R. J. (1992). An asymptotic theory for sliced inverse regression. *Ann. Statist.* **20** 1040–1061. MR1165605

JIANG, B. and LIU, J. S. (2014). Variable selection for general index models via sliced inverse regression. *Ann. Statist.* **42** 1751–1786. MR3262467

JOHNSTONE, I. M. and LU, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *J. Amer. Statist. Assoc.* **104** 682–693. MR2751448

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

LI, L. (2007). Sparse sufficient dimension reduction. *Biometrika* **94** 603–613. MR2410011

LI, L. and NACHTSHEIM, C. J. (2006). Sparse sliced inverse regression. *Technometrics* **48** 503–510. MR2328619

LIN, Q., ZHAO, Z. and LIU, J. S. (2018). Supplement to "On consistency and sparsity for sliced inverse regression in high dimensions." DOI:10.1214/17-AOS1561SUPP.

LUO, X., STEFANSKI, L. A. and BOOS, D. D. (2006). Tuning variable selection procedures by adding noise. *Technometrics* **48** 165–175. MR2277672

NEYKOV, M., LIN, Q. and LIU, J. S. (2015). Signed support recovery for single index models in high-dimensions. *Ann. Math. Sci. Appl.* **1** 379–426. DOI:10.4310/AMSA.2016.v1.n2.a5.

SZÉKELY, G. J., RIZZO, M. L. and BAKIROV, N. K. (2007). Measuring and testing dependence by correlation of distances. *Ann. Statist.* **35** 2769–2794. MR2382665

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

VERSHYNIN, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing* 210–268. Cambridge Univ. Press, Cambridge. MR2963170

WU, Y., BOOS, D. D. and STEFANSKI, L. A. (2007). Controlling variable selection by the addition of pseudovariables. *J. Amer. Statist. Assoc.* **102** 235–243. MR2345541

YU, Z., DONG, Y. and ZHU, L.-X. (2016). Trace pursuit: A general framework for model-free variable selection. *J. Amer. Statist. Assoc.* **111** 813–821. MR3538707

YU, Z., ZHU, L., PENG, H. and ZHU, L. (2013). Dimension reduction and predictor selection in semiparametric models. *Biometrika* **100** 641–654. MR3094442

ZHONG, W., ZHANG, T., ZHU, Y. and LIU, J. S. (2012). Correlation pursuit: Forward stepwise variable selection for index models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **74** 849–870. MR2988909

ZHU, L.-X. and FANG, K.-T. (1996). Asymptotics for kernel estimate of sliced inverse regression. *Ann. Statist.* **24** 1053–1068. MR1401836

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

ZHU, L. X. and NG, K. W. (1995). Asymptotics of sliced inverse regression. *Statist. Sinica* **5** 727–736. MR1347616

ZHU, L.-P., LI, L., LI, R. and ZHU, L.-X. (2011). Model-free feature screening for ultrahigh-dimensional data. *J. Amer. Statist. Assoc.* **106** 1464–1475. MR2896849

ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 301–320. MR2137327

Q. LIN
CENTER FOR STATISTICAL SCIENCE
DEPARTMENT OF INDUSTRIAL ENGINEERING
TSINGHUA UNIVERSITY
BEIJING
CHINA
AND
DEPARTMENT OF STATISTICS
HARVARD UNIVERSITY
1 OXFORD STREET
CAMBRIDGE, MASSACHUSETTS 02138
USA
E-MAIL: qianlin88@gmail.com

Z. ZHAO
DEPARTMENT OF STATISTICAL SCIENCE
TEMPLE UNIVERSITY
342 SPEAKMAN HALL
1801 N. 13TH STREET
PHILADELPHIA, PENNSYLVANIA 19122
USA
E-MAIL: zhaozhg@temple.edu

J. S. LIU
DEPARTMENT OF STATISTICS
HARVARD UNIVERSITY
1 OXFORD STREET
CAMBRIDGE, MASSACHUSETTS 02138
USA
AND
CENTER FOR STATISTICAL SCIENCE
DEPARTMENT OF INDUSTRIAL ENGINEERING
TSINGHUA UNIVERSITY
BEIJING
CHINA
E-MAIL: jliu@stat.harvard.edu