Testing Model Utility for Single Index Models Under High Dimension

Qian Lin, Zhigen Zhao, and Jun S. Liu

1 Introduction

Testing whether a quantitative response is dependent/independent of a subset of covariates is one of the central problems in statistical analyses. Most existing literature focuses on linear relationships. For instance, Arias-Castro et al. (2011b) considered the linear model

$$y = X\beta + \epsilon, \tag{1}$$

Lin's research was supported in part by National Key R&D Program of China (2020AAA0105200), the National Natural Science Foundation of China (Grant 11971257), Beijing Natural Science Foundation (Grant Z190001) and Beijing Academy of Artificial Intelligence.

Zhao's research was supported in part by the NSF Grant IIS-1633283. Liu's research was supported in part by the NSF Grants NSF DMS-1712714 and NSF DMS-2015411.

The authors "Qian Lin" and "Zhigen Zhao" contributed equally.

Q. Lin
Center for Statistical Science and Department of Industrial Engineering, Tsinghua University, Beijing, China
e-mail: qianlin@mail.tsinghua.edu.cn

Z. Zhao
Department of Statistical Science, Temple University, Philadelphia, PA, USA
e-mail: zhaozhg@temple.edu

J. S. Liu (✉)
Department of Statistics, Harvard University, Cambridge, MA, USA

Center for Statistical Science, Tsinghua University, Beijing, China
e-mail: jliu@stat.harvard.edu

E. Bura, B. Li (eds.), *Festschrift in Honor of R. Dennis Cook*,
https://doi.org/10.1007/978-3-030-69009-0_4

where $\boldsymbol{\epsilon} \sim N(0, \sigma^2 \mathbf{I})$, to test whether all the β_i's are zero. This can be formulated as the following null and alternative hypotheses:

$$\begin{cases} H_0 : & \beta_1 = \ldots = \beta_p = 0 \\ H_{s,r} : & \boldsymbol{\beta} \in \Theta_s(r) \triangleq \{\boldsymbol{\beta} \in \mathbb{R}_s^p \mid \|\boldsymbol{\beta}\|_2^2 \geq r^2\} \end{cases} \tag{2}$$

where $\mathbb{R}_s^p$ denotes the set of s-sparse vector in $\mathbb{R}^p$ with the number of nonzero entries being no greater than s. Arias-Castro et al. (2011b) and Ingster et al. (2010) showed that one can detect the signal if and only if $r^2 \succ \frac{s\log(p)}{n} \wedge \frac{p^{1/2}}{n} \wedge \frac{1}{\sqrt{n}}$. The upper bound is guaranteed by an asymptotically most powerful test based on higher criticism (Donoho and Jin, 2004).

The linearity or other functional form assumption is often too restrictive in practice. Theoretical and methodological developments beyond parametric models are important, urgent, and extremely challenging. As a first step toward nonparametric testing of the independence, we here study the single index model $y = f(\boldsymbol{\beta}^\tau \boldsymbol{x}, \epsilon)$, where $f(\cdot)$ is an unknown function. Our goal is to test the global null hypothesis that all the β_i's are zero. The first challenge is to find an appropriate formulation of alternative hypotheses because $\|\boldsymbol{\beta}\|_2^2$ used in (2) is not even identifiable in single index models.

When $rank(var(\mathbb{E}[\boldsymbol{x} \mid y]))$ is nonzero in a single index model, the unique nonzero eigenvalue λ of $var(\mathbb{E}[\boldsymbol{x} \mid y])$ can be viewed as the generalized signal-to-noise ratio (gSNR) (Lin et al., 2019). In Sect. 2, we show that for the linear regression model, this λ is almost proportional to $\|\boldsymbol{\beta}\|_2$ when it is small. The alternative hypotheses in (2) can be rewritten as $gSNR > r^2$. Because of this connection, we can treat λ as the separation quantity for the single index model and consider the following contrasting hypotheses:

$$\begin{cases} H_0 : & \text{gSNR} = 0, \\ H_a : & \text{gSNR} \geq \lambda_0. \end{cases}$$

We show that, under certain regularity conditions, one can detect a nonzero gSNR if and only if $\lambda_0 \succ \frac{s\log(p)}{n} \wedge \frac{p^{1/2}}{n} \wedge \frac{1}{\sqrt{n}}$ for the single index model with additive noise.

This is a strong and surprising result because this detection boundary is the same as that for the linear model. Using the idea from the sliced inverse regression (SIR) (Li, 1991), we show that this boundary can be achieved by the proposed **s**pectral test **s**tatistics using **S**IR (SSS) and SSS with ANOVA test assisted (SSSa). Although SIR has been advocated as an effective alternative to linear multivariate analysis (Chen and Li, 1998), the existing literature has not provided satisfactory theoretical foundations for high dimensions until recently (Lin et al., 2018a,b, 2019). We believe that the results in this paper provide further supporting evidence to the speculation that "SIR can be used to take the same role as linear regression in model building, residual analysis, regression diagnoses, etc" (Chen and Li, 1998).

In Sect. 2, after briefly reviewing the SIR and related results in linear regression, we state the optimal detection problem and a lower bound for single index models.

In Sect. 3, we first show that the correlation-based higher criticism (Cor-HC) developed for linear models fails for single index models and then propose a test to achieve the lower bound stated in Sect. 2. Some numerical studies are included in Sect. 4. We list several interesting implications and future directions in Sect. 5. Additional proofs and lemmas are included in appendices.

2 Generalized SNR for Single Index Models

2.1 Notation

The following notations are adopted throughout the paper. For a matrix $\boldsymbol{V}$, we call the space generated by its column vectors the column space and denote it by $col(\boldsymbol{V})$. The i-th row and j-th column of the matrix are denoted by $\boldsymbol{V}_{i,*}$ and $\boldsymbol{V}_{*,j}$, respectively. For vectors $\boldsymbol{x}$ and $\boldsymbol{\beta} \in \mathbb{R}^p$, we denote their inner product $\langle \boldsymbol{x}, \boldsymbol{\beta} \rangle$ by $\boldsymbol{x}(\boldsymbol{\beta})$, and the k-th entry of $\boldsymbol{x}$ by $\boldsymbol{x}(k)$. For two positive numbers a, b, we use $a \vee b$ and $a \wedge b$ to denote $\max\{a, b\}$ and $\min\{a, b\}$, respectively. Throughout the paper, we use C, C', C_1, and C_2 to denote generic absolute constants, though the actual value may vary from case to case. For two sequences $\{a_n\}$ and $\{b_n\}$, we denote $a_n \succeq b_n$ (resp. $a_n \preceq b_n$) if there exists positive constant C (resp. C') such that $a_n \geq Cb_n$ (resp. $a_n \leq C'b_n$). We denote $a_n \asymp b_n$ if both $a_n \succeq b_n$ and $a_n \preceq b_n$ hold. We denote $a_n \prec b_n$ (resp. $a_n \succ b_n$) if $a_n = o(b_n)$ (resp. $b_n = o(a_n)$). The $(1, \infty)$ norm and (∞, ∞) norm of matrix A are defined as $\|A\|_{1,\infty} = \max_{1\leq j\leq p} \sum_{i=1}^{p} |A_{i,j}|$ and $\max_{1\leq i,j\leq n} \|A_{i,j}\|$, respectively. For a finite subset S, we denote by $|S|$ its cardinality. We also write $A_{S,T}$ for the $|S| \times |T|$ submatrix with elements $(A_{i,j})_{i\in S, j\in T}$ and A_S for $A_{S,S}$. For any squared matrix A, we define $\lambda_{min}(A)$ and $\lambda_{\max}(A)$ as the smallest and largest eigenvalues of A, respectively. When y and $\boldsymbol{x}$ are independent, it is denoted as $y \perp\!\!\!\perp \boldsymbol{x}$.

2.2 A Brief Review of the Sliced Inverse Regression (SIR)

SIR was first proposed by Li (1991) to estimate the central space spanned by $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d$ based on n i.i.d. observations $(y_i, \boldsymbol{x}_i)$, $i = 1, \cdots, n$, from the multiple index model $y = f(\boldsymbol{\beta}_1^\tau \boldsymbol{x}, \ldots, \boldsymbol{\beta}_d^\tau \boldsymbol{x}, \epsilon)$, under the assumption that $\boldsymbol{x}$ follows an elliptical distribution and ϵ is Gaussian. SIR starts by dividing the data into H equal-sized slices according to the order statistics $y_{(i)}$. To ease notations and arguments, we assume that $n = cH$ and $\mathbb{E}[\boldsymbol{x}] = 0$ and re-express the data as $y_{h,j}$ and $\boldsymbol{x}_{h,j}$, where h refers to the slice number and j refers to the order number within the slice, i.e., $y_{h,j} \leftarrow y_{(c(h-1)+j)}$, $\boldsymbol{x}_{h,j} \leftarrow \boldsymbol{x}_{(c(h-1)+j)}$. Here $\boldsymbol{x}_{(k)}$ is the concomitant of $y_{(k)}$. Let the sample mean in the h-th slice be denoted by $\overline{\boldsymbol{x}}_{h,\cdot}$; then $\boldsymbol{\Lambda} \triangleq var(\mathbb{E}[\boldsymbol{x}|y])$ can be estimated by

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

where $\boldsymbol{X}_H$ denotes the $p \times H$ matrix formed by the H sample means, i.e., $\boldsymbol{X}_H = (\overline{\boldsymbol{x}}_{1,\cdot}, \ldots, \overline{\boldsymbol{x}}_{H,\cdot})$.

Thus, $col(\boldsymbol{\Lambda})$ is estimated by $col(\widehat{\boldsymbol{V}}_H)$, where $\widehat{\boldsymbol{V}}_H$ is the matrix formed by the d eigenvectors associated with the largest d eigenvalues of $\widehat{\boldsymbol{\Lambda}}_H$. The $col(\widehat{\boldsymbol{V}}_H)$ is a consistent estimator of $col(\boldsymbol{\Lambda})$ under certain technical conditions (Duan and Li, 1991; Hsing and Carroll, 1992; Li, 1991; Lin et al., 2018b; Zhu et al., 2006). It is shown in Lin et al. (2018a,b) that, for single index models ($d = 1$), H can be chosen as a fixed number not depending on $\lambda(\boldsymbol{\Lambda})$, n, and p for the asymptotic results to hold. Throughout this paper, we assume the following mild conditions:

(**A1**) $\boldsymbol{x} \sim N(0, \boldsymbol{\Sigma})$, and there exist two positive constants $C_{\min} < C_{\max}$, such that $C_{\min} < \lambda_{\min}(\boldsymbol{\Sigma}) \le \lambda_{\max}(\boldsymbol{\Sigma}) < C_{\max}$.

(**A2**) Sliced stable condition. For $0 < \mathfrak{a}_1 < 1 < \mathfrak{a}_2$, let $\mathcal{A}_H(\mathfrak{a}_1, \mathfrak{a}_2)$ denote all partitions $\{-\infty = a_0 \le a_2 \le \ldots \le a_H = +\infty\}$ of $\mathbb{R}$ satisfying that

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

A curve $m(y)$ is ϑ-sliced stable with respect to y, if there exist positive constants $\mathfrak{a}_1, \mathfrak{a}_2, \mathfrak{a}_3$ and large enough H_0 such that for any $H > H_0$, for any partition in $\mathcal{A}_H(\mathfrak{a}_1, \mathfrak{a}_2)$ and any $\gamma \in \mathbb{R}^p$, one has

$$\frac{1}{H}\sum_{h=1}^{H} var\left(\gamma^{\tau} m(y) \middle| a_{h-1} \le y < a_h\right) \le \frac{\mathfrak{a}_3}{H^{\vartheta}} var\left(\gamma^{\tau} m(y)\right). \tag{4}$$

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

The sliced stable condition is introduced in Lin et al. (2018b) to study the phase transition of SIR. The sliced stable condition is a mild condition. Neykov et al. (2016) derived the sliced stable condition from a modification of the regularity condition proposed in Hsing and Carroll (1992). For this paper, we modified it for single index models.

2.3 Generalized Signal-to-Noise Ratio of Single Index Models

We consider the following single index model:

$$y = f(\boldsymbol{\beta}^{\tau}\boldsymbol{x}, \epsilon),\ \boldsymbol{x} \sim N(0, \boldsymbol{\Sigma}),\ \epsilon \sim N(0, \sigma^2), \tag{5}$$

where $f(\cdot)$ is an unknown function. What we want to know is whether the coefficient vector $\boldsymbol{\beta}$, when viewed as a whole, is zero. This can be formulated as a global testing problem as

$$H_0 : \boldsymbol{\beta} = 0 \ \text{ versus } \quad H_a : \boldsymbol{\beta} \neq 0.$$

When assuming the linear model $\boldsymbol{y} = \boldsymbol{\beta}^\tau \boldsymbol{x} + \boldsymbol{\epsilon}$, whether we can separate the null and alternative depends on the interplay between σ^2 and the norm of $\boldsymbol{\beta}$. More precisely, it depends on the signal-to-noise ratio (SNR) defined as

$$SNR = \frac{E[(\boldsymbol{\beta}^\tau \boldsymbol{x})^2]}{\mathbb{E}[y^2]} = \frac{\|\boldsymbol{\beta}\|_2^2 \boldsymbol{\beta}_0^\tau \boldsymbol{\Sigma} \boldsymbol{\beta}_0}{\sigma^2 + \|\boldsymbol{\beta}\|_2^2 \boldsymbol{\beta}_0^\tau \boldsymbol{\Sigma} \boldsymbol{\beta}_0}$$

when $\boldsymbol{\beta} \neq 0$ and $\boldsymbol{\beta}_0 = \boldsymbol{\beta}/\|\boldsymbol{\beta}\|_2$ (Janson et al., 2017). Here $||\boldsymbol{\beta}||_2$ is useful for benchmarking prediction accuracy for various model selection techniques such as AIC, BIC, or the Lasso. However, since there is an unknown link function $f(\cdot)$ in the single index model, the norm $||\boldsymbol{\beta}||_2$ becomes non-identifiable. Without loss of generality, we restrict $||\boldsymbol{\beta}||_2 = 1$ and have to find another quantity to describe the separability.

For the single index model (5), to simplify the notation, use λ to denote $\lambda_{\max}(var(\mathbb{E}[\boldsymbol{x}|y]))$. For linear models, we can easily show that

$$var(\mathbb{E}[\boldsymbol{x}|y]) = \frac{\boldsymbol{\Sigma}\boldsymbol{\beta}\boldsymbol{\beta}^\tau \boldsymbol{\Sigma}}{\boldsymbol{\beta}_0^\tau \boldsymbol{\Sigma} \boldsymbol{\beta}_0 \|\boldsymbol{\beta}\|_2^2 + \sigma^2} \text{ and } \lambda = \frac{\boldsymbol{\beta}_0^\tau \boldsymbol{\Sigma}\boldsymbol{\Sigma}\boldsymbol{\beta}_0 \|\boldsymbol{\beta}\|_2^2}{\boldsymbol{\beta}_0^\tau \boldsymbol{\Sigma} \boldsymbol{\beta}_0 \|\boldsymbol{\beta}\|_2^2 + \sigma^2}.$$

Consequently, $\lambda/SNR = \frac{\boldsymbol{\beta}_0^\tau \boldsymbol{\Sigma}\boldsymbol{\Sigma}\boldsymbol{\beta}_0}{\boldsymbol{\beta}_0^\tau \boldsymbol{\Sigma}\boldsymbol{\beta}_0}$. When assuming condition (A2), such a ratio is bounded by two finite limits. Thus, λ can be treated as an equivalent quantity to the SNR for linear models and is therefore named as the generalized signal-to-noise ratio (gSNR) for single index models.

Remark 1 To the best of our knowledge, although SIR uses the estimation of λ to determine the structural dimension (Li, 1991), few investigations have been made toward theoretical properties of this procedure in high dimensions. The only work that uses λ as a parameter to quantify the estimation error when estimating the direction of $\boldsymbol{\beta}$ is Lin et al. (2018a), which, however, does not indicate explicitly what role λ plays. The aforementioned observation about λ for single index models provides a useful intuition: λ is a generalized notion of the SNR, and condition (**A2**) merely requires that gSNR is nonzero.

2.4 Global Testing for Single Index Models

As we have discussed, Arias-Castro et al. (2011b) and Ingster et al. (2010) considered the testing problem (2), which can be viewed as the determination of the detection boundary of gSNR. Through the whole paper, we consider the following testing problem:

$$\begin{cases} H_0: & gSNR = 0, \\ H_a: & \lambda(= gSNR) \quad \text{is nonzero}, \end{cases} \tag{6}$$

based on i.i.d. samples $\{(y_i, \boldsymbol{x}_i), i = 1, \ldots, n\}$. Two models are considered: (i) the general single index model (SIM) defined in (5) and (ii) the single index model with additive noise (SIMa) defined as

$$y = f(\boldsymbol{\beta}^\tau \boldsymbol{x}) + \epsilon,\ \boldsymbol{x} \sim N(0, \boldsymbol{\Sigma}),\ \epsilon \sim N(0, \sigma^2). \tag{7}$$

We assume that conditions (**A1**) and (**A2**) hold for both models.

3 The Optimal Test for Single Index Models

3.1 The Detection Boundary of Linear Regression

To set the goal and scope, we briefly review some related results on the detection boundary for linear models (Arias-Castro et al., 2011b; Ingster et al., 2010).

Proposition 1 *Assume that $\boldsymbol{x}_i \sim N(0, \mathbf{I}_p)$, $i = 1, \cdots, n$, and that $\boldsymbol{\beta}$ has at most s nonzero entries. There is a test with both type I and II errors converging to zero for the testing problem in* (2) *if and only if*

$$r^2 \succ \frac{s\log(p)}{n} \wedge \frac{p^{1/2}}{n} \wedge \frac{1}{\sqrt{n}}. \tag{8}$$

Assuming $\boldsymbol{x} \sim N(0, \mathbf{I}_p)$ and the variance of the noise is known, Ingster et al. (2010) obtained the sharp detection boundary (i.e., with exact asymptotic constant) for the above problem. Since linear models are special cases of SIMa, which is a special subset of SIM, the following statement about the lower bound of detectability is a direct corollary of Proposition 1.

Corollary 1

i) If $s^2\log^2(p) \wedge p \prec n$, then any test fails to separate the null and the alternative hypothesis asymptotically for SIM when

$$\lambda \prec \frac{s\log(p)}{n} \wedge \frac{p^{1/2}}{n}. \tag{9}$$

ii) Any test fails to separate the null and the alternative hypothesis asymptotically for SIMa when

$$\lambda \prec \frac{s\log(p)}{n} \wedge \frac{p^{1/2}}{n} \wedge \frac{1}{\sqrt{n}}. \tag{10}$$

3.2 Single Index Models

Moving from linear models to single index models is a big step. A natural and reasonable start is to consider tests based on the marginal correlation used for linear models (Arias-Castro et al., 2011b; Ingster et al., 2010). However, the following example shows that the marginal correlation fails for the single index models, indicating that we need to look for some other statistics to approximate the gSNR.

Example 1 Suppose that $\boldsymbol{x} \sim N(0, \mathbf{I}_p)$, $\epsilon \sim N(0, 1)$, and we have n samples from the following model:

$$y = (x_1 + \ldots + x_l) - (x_1 + \ldots + x_l)^3/3l + \epsilon. \tag{11}$$

Simple calculation shows that $\mathbb{E}[\boldsymbol{x}y] = 0$. Thus, correlation-based methods do not work for this simple model. On the other hand, since the link function $f(t) = t - t^3/3l$ is monotone when $|t|$ is sufficiently large, we know that $\mathbb{E}[\boldsymbol{x} \mid y]$ is not a constant and $var(\mathbb{E}[\boldsymbol{x} \mid y]) \neq 0$.

Let λ_0 and λ_0^a be two sequences such that

$$\lambda_0 \succ \frac{s\log(p)}{n} \wedge \frac{p^{1/2}}{n}, \quad \lambda_0^a \succ \frac{s\log(p)}{n} \wedge \frac{p^{1/2}}{n} \wedge \frac{1}{\sqrt{n}}.$$

For a $p \times p$ symmetric matrix A and a positive constant k such that $ks < p$, we define

$$\lambda_{\max}^{(ks)}(A) = \max_{|S|=ks} \lambda_{\max}(A_S). \tag{12}$$

For model $y = f(\boldsymbol{\beta}^\tau \boldsymbol{x}, \epsilon)$, in addition to the condition that $\lambda_0 \prec \lambda$, we further assume that $s^2 \log^2(p) \wedge p \prec n$.

Let $\widehat{\boldsymbol{\Lambda}}_H$ be the estimate of $var(\mathbb{E}[\boldsymbol{x}|y])$ based on SIR. Let τ_n, τ_n', and τ_n'' be three quantities satisfying

$$\frac{\sqrt{p}}{n} \prec \tau_n \prec \lambda_0, \quad \frac{s\log(p)}{n} \prec \tau_n' \prec \lambda_0, \quad \frac{1}{\sqrt{n}} \prec \tau_n'' \prec \lambda_0^a. \tag{13}$$

We introduce the following two assistance tests:

1. Define

$$\psi_1(\tau_n) = \mathbf{1}(\lambda_{\max}(\widehat{\mathbf{\Lambda}}_H) > \frac{tr(\mathbf{\Sigma})}{n} + \tau_n).$$

2. Define

$$\psi_2(\tau_n') = \mathbf{1}(\lambda_{\max}^{(ks)}(\widehat{\mathbf{\Lambda}}_H) > \tau_n').$$

Finally, the **s**pectral test **s**tatistic based on **SIR**, abbreviated as **SSS**, is defined as

$$\Psi_{SSS} = \max\{\psi_1(\tau_n), \psi_2(\tau_n')\}. \tag{14}$$

To show the theoretical properties of SSS, we impose the following condition on the covariance matrix $\mathbf{\Sigma}$:

(**A3**) There are at most k nonzero entries in each row of $\mathbf{\Sigma}$.

This assumption is first explicitly proposed in Lin et al. (2018b), which is partially motivated by the separable after screening (SAS) properties in Ji and Jin (2012). In this paper, we assume such a relative strong condition and focus on establishing the detection boundary. This condition can be possibly relaxed by considering a larger class of covariance matrices

$$\mathcal{S}(\gamma, \Delta) = \left\{ |\mathbf{\Sigma}_{jl}| \leq 1 - (\log(p))^{-1}, \quad |\{l \mid \mathbf{\Sigma}_{jl} > \gamma\}| \leq \Delta \right\},$$

which is used in Arias-Castro et al. (2011a) for analyzing linear models. Our condition contains $\mathcal{S}(0, \Delta)$ for some positive constant Δ, and we could relax our constraint to some $\mathcal{S}(\gamma, \Delta)$. However, the technical details will be much more involved, which masks the importance of the main results. We thus leave it for a future investigation.

Theorem 1 *Assume that* $s^2\log^2(p) \wedge p \prec n$, $\lambda \succ \lambda_0$, *and conditions* (**A1**)$-$(**A3**) *hold. Two sequences* τ_n *and* τ_n' *satisfy the conditions in (13). Then, type I and type II errors of the test* $\Psi_{SSS}(\tau_n, \tau_n')$ *converge to zero for the testing problem under SIM.*

Comparing with the test proposed in Ingster et al. (2010), our test statistics is a spectral statistics and depends on the first eigenvalue of $\widehat{\mathbf{\Lambda}}_H$. It is adaptive in the moderate-sparsity scenario. In the high-sparsity scenario when $s^2\log^2(p) \prec p$, the SSS relies on $\psi_2(\tau_n')$, which depends on the sparsity s of the vector $\boldsymbol{\beta}$. Therefore, SSS is not adaptive to the sparsity level. Both Arias-Castro et al. (2011a) and Ingster et al. (2010) introduced an (adaptive) asymptotically powerful test based on the higher criticism (HC) for the testing problem under linear models. It is an interesting research problem to develop an adaptive test using the idea of higher criticism for (6).

3.3 Optimal Test for SIMa

When the noise is assumed additive as in SIMa (7), the detection boundary can be further improved. In addition to conditions **(A1)–(A3)**, f is further assumed to satisfy the following condition:

(B) $f(z)$ is sub-Gaussian, $\mathbb{E}[f(z)] = 0$, and $var(f(z)) > Cvar(\mathbb{E}[z \mid f(z) + \epsilon])$ for some constant C, where $z, \epsilon \overset{iid}{\sim} N(0, 1)$.

Note that for any fixed function f such that $var(\mathbb{E}[z \mid f(z) + \epsilon]) \neq 0$, there exists a positive constant C such that

$$\frac{var(f(z))}{var(\mathbb{E}[z \mid f(z) + \epsilon])} > C. \tag{15}$$

By continuity, we know that (15) holds in a small neighborhood of f, i.e., if C is sufficiently small, condition (***B***) holds for a large class of functions.

First, we adopt the test $\Psi_{SSS}(\tau_n, \tau_n')$ described in the previous subsection. Since the noise is additive, we include the ANOVA test:

$$\psi_3(\tau_n'') = \mathbf{1}(t > \tau_n'')$$

where $t = \frac{1}{n}\sum_{j=1}^{n}(y_j^2 - 1)$ and τ_n'' is a sequence satisfying the condition (13). Combing this test with the test $\Psi_{SSS}(\tau_n, \tau_n')$, we can introduce SSS assisted by ANOVA test (SSSa) as

$$\Psi_{SSSa}(\tau_n, \tau_n', \tau_n'') = \max\{\Psi_{SSS}(\tau_n, \tau_n'), \psi_3(\tau_n'')\}. \tag{16}$$

We then have the following result.

Theorem 2 *Assume that* $\lambda \succ \lambda_0^a$ *and the conditions* (**A1**)$-$(**A3**) *and* (**B**) *hold. Assume that the sequences* τ_n, τ_n', *and* τ_n'' *satisfy condition (13); then type I and type II errors of the test* $\Psi_{SSSa}(\tau_n, \tau_n', \tau_n'')$ *converge to zero for the testing problem under SIMa.*

Example Continued. For the example in (11), we calculated the test statistic ψ_{SSS} defined by (14) under both the null and alternative hypotheses. Figure 1 shows the histograms of such a statistic under both hypotheses, demonstrating a perfect separation between the null and alternative. For this example, $\lambda_{\max}^{ks}(\widehat{\mathbf{\Lambda}}_H)$ has more discrimination power than $\lambda_{\max}(\widehat{\mathbf{\Lambda}}_H)$.

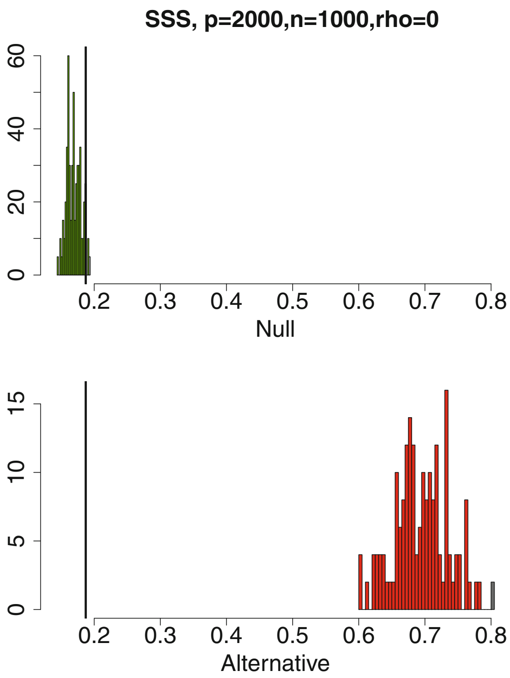


Fig. 1 The histograms of $\lambda_{\max}^{ks}(\hat{\boldsymbol{\Lambda}}_H)$ for the model (11). The top panel corresponds to the scores under the null, and the bottom one corresponds to the scores under the alternative. The "black" vertical line is the 95% quantile under the null

3.4 Computationally Efficient Test

Although the test Ψ_{SSS} (and Ψ_{SSSa}) is rate optimal, it is computationally inefficient. Here we propose an efficient algorithm to approximate $\lambda_{\max}^{(ks)}(\widehat{\boldsymbol{\Lambda}}_H)$ via a convex relaxation, which is similar to the convex relaxation method for estimating the top eigenvector of a semi-definite matrix (Adamczak et al., 2008; Berthet and Rigollet, 2013b; d'Aspremont et al., 2005, 2014). To be precise, given the SIR estimate $\widehat{\boldsymbol{\Lambda}}_H$ of $var(\mathbb{E}[\boldsymbol{x} \mid y])$, consider the following semi-definite programming (SDP) problem:

$$\begin{aligned} \widetilde{\lambda}_{\max}^{(ks)}(\widehat{\boldsymbol{\Lambda}}_H) \triangleq \max \quad & tr(\widehat{\boldsymbol{\Lambda}}_H \boldsymbol{M}), \\ \text{subject to} \quad & tr(\boldsymbol{M}) = 1, \quad |\boldsymbol{M}|_1 \leq ks, \\ & \boldsymbol{M} \text{ is semi-definite positive.} \end{aligned} \tag{17}$$

With $\widetilde{\lambda}_{\max}^{(ks)}(\widehat{\boldsymbol{\Lambda}}_H)$, for a sequence τ_n' satisfying the condition in (13), i.e., $\frac{s\log(p)}{n} \prec \tau_n' \prec \lambda_0$, a computationally feasible test is

$$\widetilde{\psi}_2(\tau_n') = \mathbf{1}(\widetilde{\lambda}_{\max}^{(ks)}(\widehat{\boldsymbol{\Lambda}}_H) > \tau_n').$$

Then, for any sequence τ_n satisfying the inequality in (13), we define the following computationally feasible alternative of Ψ_{SSS}:

$$\widetilde{\Psi}_{SSS} = \max\{\psi_1(\tau_n), \widetilde{\psi}_2(\tau_n')\}. \tag{18}$$

Theorem 3 *Assume that* $s^2\log^2(p) \wedge p \prec n$, $\lambda \succ \lambda_0$, *and conditions* (**A1**)−(**A3**) *hold. Then, type I and type II errors of the test* $\widetilde{\Psi}_{SSS}(\tau_n, \tau_n')$ *converge to zero for the testing problem under SIMa.*

Similarly, if we introduce the test

$$\widetilde{\Psi}_{SSSa}(\tau_n, \tau_n', \tau_n'') = \max\{\widetilde{\Psi}_{SSS}, \psi_3(\tau_n'')\}, \tag{19}$$

for three sequences τ_n, τ_n', and τ_n'', then we have:

Theorem 4 *Assume that* $\lambda \succ \lambda_0^a$ *and conditions* (**A1**)−(**A3**) *and* (**B**) *hold. The test* $\widetilde{\Psi}_{SSSa}(\tau_n, \tau_n', \tau_n'')$ *is asymptotically powerful for the testing problem under SIMa.*

Theorems 2 and 4 not only establish the detection boundary of gSNR for single index models but also open a door of thorough understanding of semi-parametric regression with a Gaussian design. It is shown in Lin et al. (2018a) that if we denoted the single index models satisfying conditions **(A1)**, **(A3)**, and $rank(var(x|y)) > 0$, one has

$$\sup_{\widehat{\beta}} \inf_{\mathfrak{m}\in\mathfrak{M}} \mathbb{E}_{\mathfrak{m}} \| P_{\widehat{\beta}} - P_{\beta}\|_F^2 \asymp 1 \wedge \frac{s\log(ep/s)}{n\lambda}, \tag{20}$$

where $P_{\hat{\beta}} = \hat{\beta}(\hat{\beta}^T\hat{\beta})^{-1}\hat{\beta}^T$ and $P_{\beta} = \beta(\beta^T\beta)^{-1}\beta^T$ are the projection operators with respect to $\hat{\beta}$ and β, respectively, and the space $\mathfrak{M}$ is defined in Equation (14) of Lin et al. (2018a). This implies that the necessary and sufficient condition for obtaining a consistent estimate of the projection operator $P_{\boldsymbol{\beta}}$ is $\frac{s\log(ep/s)}{n} \prec \lambda$. On the other hand, Theorems 2 and 4 state that, for single index models with additive noise, if $\frac{s\log(p)}{n} \wedge \frac{p^{1/2}}{n} \wedge \frac{1}{\sqrt{n}} \prec \lambda$, then one can detect the existence of gSNR (aka nontrivial direction $\boldsymbol{\beta}$). Our results thus imply for SIMa that, if $\frac{p^{1/2}}{n} \wedge \frac{1}{\sqrt{n}} \prec \lambda \prec \frac{s\log(p)}{n}$, one can detect the existence of nonzero $\boldsymbol{\beta}$, but cannot provide a consistent estimation of its direction. To estimate the location of nonzero coefficient especially when focusing on the almost recovery region (Ji and Jin, 2012), we must tolerate a certain error rate such as the false discovery rate (Benjamini and Hochberg, 1995). For example, the knockoff procedure (Barber and Candès, 2015), SLOPE (Su and Candes, 2016), and UPT (Ji and Zhao, 2014) might be extended to single index models.

3.5 Practical Issues

In practice, we do not know whether the noise is additive or not. Therefore, we only consider the test statistic $\widetilde{\Psi}_{SSS}$. Condition (13) provides us a theoretical basis for choosing the sequences τ_n and τ_n'. In practice, however, we determine these thresholds by simulating the null distribution of $\lambda_{\max}(\widehat{\mathbf{\Lambda}}_H)$ and $\tilde{\lambda}_{\max}^{(ks)}(\widehat{\mathbf{\Lambda}}_H)$. Our final algorithm is as follows.

Algorithm 1 Spectral test statistic based on SIR (SSS) algorithm

1. Calculate $\lambda_{\max}(\widehat{\mathbf{\Lambda}}_H)$ and $\widetilde{\lambda}_{\max}^{(ks)}(\widehat{\mathbf{\Lambda}}_H)$ for the given input $(\boldsymbol{x}, \boldsymbol{y})$
2. Generate $\boldsymbol{z} = (z_1, \cdots, z_n)$, where $z_i \overset{\text{iid}}{\sim} N(0, 1)$
3. Calculate $\lambda_{\max}(\widehat{\mathbf{\Lambda}}_H)$ and $\widetilde{\lambda}_{\max}^{(ks)}(\widehat{\mathbf{\Lambda}}_H)$ based on $(\boldsymbol{x}, \boldsymbol{z})$
4. Repeat Steps 2 and 3 $N(= 100)$ times to get two sequences of $\lambda_{\max}$ and $\widetilde{\lambda}_{\max}^{(ks)}$. Let τ_n and τ_n' be the 95% quantile of these two simulated sequences
5. Reject the null if $\lambda_{\max}(\widehat{\mathbf{\Lambda}}_H) > \tau_n$ and/or $\tilde{\lambda}_{\max}^{(ks)}(\widehat{\mathbf{\Lambda}}_H) > \tau_n'$

4 Numerical Studies

Let $\boldsymbol{\beta}$ be the vector of coefficients, and let $\mathcal{S}$ be the active set, $\mathcal{S} = \{i : \beta_i \neq 0\}$, for which we simulated $\beta_i \overset{iid}{\sim} N(0, 1)$. Let $\boldsymbol{x}$ be the random design matrix with each row following $N(0, \mathbf{\Sigma})$. We consider two types of covariance matrices: (i) $\mathbf{\Sigma} = (\sigma_{ij})$ with $\sigma_{ii} = 1$ and $\sigma_{ij} = \rho^{|i-j|}$ and (ii) $\sigma_{ii} = 1, \sigma_{ij} = \rho$ when $i, j \in \mathcal{S}$ or $i, j \in \mathcal{S}^c$ and $\sigma_{ij} = \sigma_{ji} = 0.1$ when $i \in \mathcal{S}, j \in \mathcal{S}^c$. The first one represents a covariance matrix which is essentially sparse, and we choose ρ among 0, 0.3, 0.5, and 0.8. The second one represents a dense covariance matrix with ρ chosen as 0.2. In all the simulations, $n = 1000$, p varies among 100, 500, 1000, and 2000 and the number of replication is 100. The random error $\boldsymbol{\epsilon}$ follows $N(0, \mathbf{I}_n)$. We consider the following models:

I. $\boldsymbol{y} = 0.02 * (16\boldsymbol{x\beta} - \exp(\boldsymbol{x\beta})) + \boldsymbol{\epsilon}$, where $|\mathcal{S}| = 7$;
II. $\boldsymbol{y} = 0.2 * \sin(\boldsymbol{x\beta}/2) * \exp(\boldsymbol{x\beta}/2) + \epsilon$, where $|\mathcal{S}| = 10$;
III. $\boldsymbol{y} = 0.8 * \left(\boldsymbol{x\beta} - (\boldsymbol{x\beta})^3/15\right) + \epsilon$, where $|\mathcal{S}| = 5$;
IV. $\boldsymbol{y} = \sin(\boldsymbol{x\beta}) * \exp(\boldsymbol{x\beta}/10) * \epsilon$, where $|\mathcal{S}| = 10$.

We choose $H = 20$ in the Algorithm 1 and assume the oracle information of the sparsity in the numerical studies because the goal of the numerical investigation is to demonstrate the theoretical detection boundary. A data-driven choice of such a tuning parameter is challenging to get and unnecessarily obscures the theoretical pattern. If we calculate $N(= 100)$ test statistics for each replication, it will take an extremely long time. Therefore, in the simulation, we calculate τ_n and τ_n' slightly different from Algorithm 1. For each generated data set, we simulated only one vector z where $z \sim N(0, \mathbf{I}_n)$ and calculate the statistic $\lambda_{\max}(\widehat{\mathbf{\Lambda}}_H)$ and $\widetilde{\lambda}^{(ks)}_{\max}(\widehat{\mathbf{\Lambda}}_H)$. The τ_n and τ_n' are chosen as 95% quantile from the corresponding sequence for all the replications.

For each generated data, we also calculated Cor-HC scores according to Arias-Castro et al. (2012). The threshold c_{hc} is chosen according to the same scheme as choosing the thresholds τ_n and τ_n'. Namely, we calculated the Cor-HC scores based on z where $z \sim N(0, \mathbf{I}_n)$. The threshold c_{hc} is the 95% quantile of these simulated scores. The hypothesis is rejected if the Cor-HC score is greater than c_{hc}. The power for both methods is calculated as the average number of rejections out of 100 replications. These numbers are reported in Table 1.

It is clearly seen that the power of SSS decreases when the dimension p increases. Nevertheless, the power of SSS is better than the one based on Cor-HC except for one case. In Fig. 2, we plot the histogram of the statistic $\widetilde{\lambda}^{(ks)}_{\max}(\widehat{\mathbf{\Lambda}}_H)$ under the null in the top-left panel and the histogram of this statistic under the alternative in the bottom-left panel for Model III when $p = 500$ and $\rho = 0.3$ for type (i) covariance matrix. It is clearly seen that the test statistic Ψ_{SSS} is well separated under the null and alternative. However, Cor-HC fails to distinguish between the null and alternative as shown in the two panels on the right side.

To see how the performance of Cor-HC varies, we consider the following model:

V. $\boldsymbol{y} = \kappa \boldsymbol{x}\boldsymbol{\beta} - \exp(\boldsymbol{x}\boldsymbol{\beta}) + \boldsymbol{\epsilon}$, where $|\mathcal{S}| = 7, \kappa = 1, 3, 5, \cdots, 19$.

Set $n = 1000$, $p = 1000$, and $\rho = 0.3$ for type (i) covariance matrix, and the power of both methods are displayed in Fig. 3. The coefficient κ determines the magnitude of the marginal correlation between the active predictors and the response. It is seen that when κ is close to 16, representing the case of diminishing marginal correlation, the power of Cor-HC dropped to the lowest. Under all the models, SSS is more powerful in detecting the existence of the signal.

To observe the influence of the signal-to-noise ratio on the power of the tests, we consider the following two models:

VI. $\boldsymbol{y} = (15\boldsymbol{x}\boldsymbol{\beta} - \exp(\boldsymbol{x}\boldsymbol{\beta})) * \kappa + 4\boldsymbol{\epsilon}$, where $|\mathcal{S}| = 7$;
VII. $\boldsymbol{y} = \sin(\boldsymbol{x}\boldsymbol{\beta}) * \exp(10\boldsymbol{x}\boldsymbol{\beta}\kappa) * \epsilon$, where $|\mathcal{S}| = 10$.

Here $\kappa = 0.01, 0.02, \ldots, 0.10$.

Set $n = 1000$, $p = 1000$, and $\rho = 0.3$; we plot the power of both methods against the coefficient κ in Fig. 4. It is clearly seen that for both examples there is a sharp "phase transition" for the power of SSS as the signal strength increases, validating our theory about the detection boundary. In both examples SSS is much more powerful than Cor-HC.

Table 1 Power comparison of SSS and HC for four models I–IV for different parameter settings. Symbol "$*$" indicates the type (ii) covariance matrix

Model	Dim	ρ	SSS	HC	Model	Dim	ρ	SSS	HC
I	100	0	1.00	0.16	II	100	0	0.98	0.12
		0.3	1.00	0.29			0.3	0.97	0.16
		0.5	0.99	0.54			0.5	0.96	0.24
		0.8	1.00	0.93			0.8	1.00	0.37
		0.2*	0.90	0.35			0.2*	0.96	0.56
	500	0	0.98	0.16		500	0	0.87	0.06
		0.3	0.99	0.18			0.3	0.80	0.09
		0.5	0.97	0.34			0.5	0.82	0.13
		0.8	0.98	0.71			0.8	0.83	0.14
		0.2*	0.52	0.25			0.2*	0.77	0.32
	1000	0	0.89	0.19		1000	0	0.81	0.09
		0.3	0.88	0.16			0.3	0.74	0.06
		0.5	0.91	0.33			0.5	0.77	0.08
		0.8	0.96	0.53			0.8	0.84	0.11
		0.2*	0.37	0.30			0.2*	0.69	0.25
	2000	0	0.92	0.18		2000	0	0.75	0.11
		0.3	0.86	0.25			0.3	0.68	0.12
		0.5	0.83	0.43			0.5	0.68	0.13
		0.8	0.90	0.60			0.8	0.81	0.10
		0.2*	0.43	0.17			0.2*	0.63	0.41
III	100	0	1.00	0.21	IV	100	0	0.89	0.01
		0.3	1.00	0.25			0.3	0.91	0.03
		0.5	1.00	0.63			0.5	0.89	0.04
		0.8	1.00	1.00			0.8	1.00	0.10
		0.2*	0.98	0.78			0.2*	0.94	0.07
	500	0	0.99	0.11		500	0	0.70	0.03
		0.3	1.00	0.12			0.3	0.57	0.04
		0.5	0.98	0.11			0.5	0.57	0.07
		0.8	0.99	0.22			0.8	0.69	0.09
		0.2*	0.62	0.72			0.2*	0.45	0.08
	1000	0	0.99	0.11		1000	0	0.55	0.07
		0.3	0.97	0.06			0.3	0.56	0.04
		0.5	0.97	0.18			0.5	0.51	0.09
		0.8	0.92	0.10			0.8	0.73	0.06
		0.2*	0.60	0.59			0.2*	0.44	0.08
	2000	0	0.96	0.16		2000	0	0.58	0.07
		0.3	0.97	0.19			0.3	0.47	0.07
		0.5	0.93	0.15			0.5	0.45	0.09
		0.8	0.88	0.10			0.8	0.61	0.02
		0.2*	0.59	0.58			0.2*	0.40	0.08

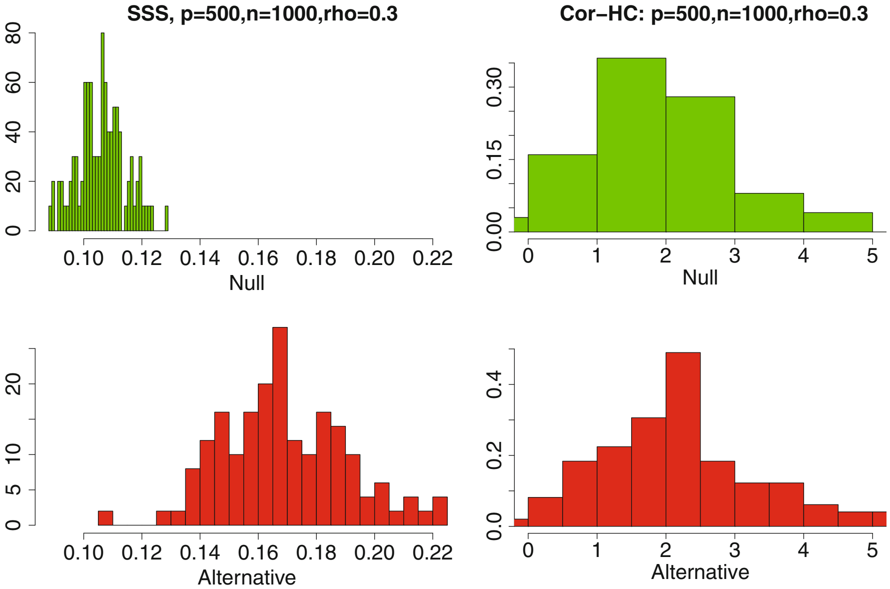
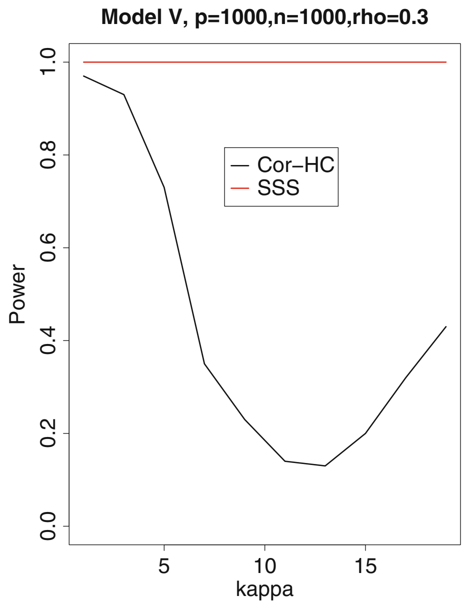


Fig. 2 Model III, $n = 1000$, $p = 500$, type (i) covariance matrix, $\rho = 0.3$

Fig. 3 Power: Model V, $n = 1000$, $p = 1000$, $\rho = 0.3$ for type (i) covariance matrix



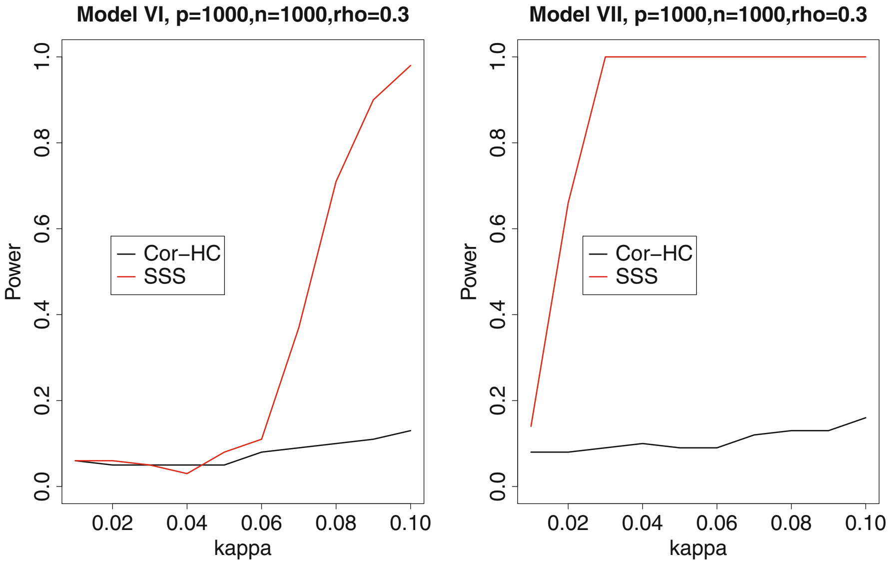


Fig. 4 Power: Models VI and VII, $n = 1000$, $p = 1000$, $\rho = 0.3$ for the type (i) covariance matrix

5 Discussion

Assuming that $var(\mathbb{E}[\boldsymbol{x} \mid y])$ is nonvanishing, we show in this paper that λ, the unique nonzero eigenvalue of $var(\mathbb{E}[\boldsymbol{x} \mid y])$ associated with the single index model, is a generalization of the SNR. We demonstrate a surprising similarity between linear regression and single index models with Gaussian design: the detection boundary of gSNR for the testing problem (6) under SIMa matches that of SNR for linear models (2). This similarity provides an additional support to the speculation that "the rich theories developed for linear regression can be extended to the single/multiple index models" (Chen and Li, 1998; Lin et al., 2019).

Besides the gap we explicitly depicted between detection and estimation boundaries, we provide here several other directions which might be of interests to researchers. First, although this paper only deals with single index models, the results obtained here are very likely extendable to multiple index models. Assume that the noise is additive, and let $0 < \lambda_d \leq \ldots \leq \lambda_1$ be the nonzero eigenvalues associated with the matrix $var(\mathbb{E}[\boldsymbol{x}|y])$ of a multiple index model. Similar arguments can show that the i-th direction is detectable if $\lambda_i \succ \frac{\sqrt{p}}{n} \wedge \frac{s\log(p)}{n} \wedge \frac{1}{\sqrt{n}}$. New thoughts and technical preparations might be needed for a rigorous argument for determining the lower bound of the detection boundary. Second, the framework can be extended to study theoretical properties of other sufficient dimension reduction algorithms such as SAVE and directional regression (Lin et al., 2018a,b, 2019).

Acknowledgments We thank Dr. Zhisu Zhu for his generous help with SDP.

Appendix: Proofs

Assisting Lemmas

Since our approaches are based on the technical tools developed in Lin et al. (2018a,b, 2019), we briefly recollect the necessary (modified) statements without proofs below.

Lemma 1 *Let* $z_j \overset{\text{iid}}{\sim} N(0,1)$, $j=1,\ldots,p$. *Let* $\sigma_1,\ldots,\sigma_p$ *be* p *positive constants satisfying* $\sigma_1 \le \ldots \le \sigma_p$. *Then for any* $0<\alpha \le \frac{1}{\sigma_p^2}\sum_j \sigma_j^4$, *we have*

$$\mathbb{P}\left(\sum_j \sigma_j^2\left(z_j^2-1\right) > \alpha\right) \le \exp\left(-\frac{\alpha^2}{4\sum \sigma_j^4}\right). \tag{21}$$

Lemma 2 *Suppose that a* $p\times H$ *matrix* $\boldsymbol{X}$ *formed by* H *i.i.d.* p *dimensional vector* $\boldsymbol{x}_j \sim N(0,\boldsymbol{\Sigma})$, $j=1,\ldots,H$ *where* $0<C_1\le \lambda_{\min}(\boldsymbol{\Sigma}) \le \lambda_{\max}(\boldsymbol{\Sigma}) \le C_2$ *for some constants* C_1 *and* C_2. *We have*

$$\left\|\frac{1}{p}\boldsymbol{X}^\tau \boldsymbol{X} - \frac{tr(\boldsymbol{\Sigma})}{p}\mathbf{I}_H\right\|_F > \alpha \tag{22}$$

with probability at most $4H^2\exp\left(-\frac{Cp\alpha^2}{H^2}\right)$ *for some positive constant* C. *In particular, we know that*

$$\lambda_{\max}\left(\boldsymbol{X}\boldsymbol{X}^\tau/p\right) = \lambda_{\max}\left(\boldsymbol{X}^\tau\boldsymbol{X}/p\right) \le tr(\boldsymbol{\Sigma})/p+\alpha \tag{23}$$

happens with probability at least $1-4H^2\exp\left(-\frac{Cp\alpha^2}{H^2}\right)$.

Lemma 3 *Assume that* $p^{1/2} \prec n\lambda$. *Let* $\boldsymbol{M} = \begin{pmatrix} B_1 & 0 \\ B_2 & B_3 \\ 0 & B_4 \end{pmatrix}$ *be a* $p\times H$ *matrix, where* B_1 *and* B_2 *are scalar,* B_3 *is a* $1\times(H-1)$ *vector, and* B_4 *is a* $(p-2)\times(H-1)$ *matrix satisfying*

$$\begin{aligned} &\left(1-\frac{1}{2\nu}\right)\lambda \leq B_1^2 \leq \left(1+\frac{1}{2\nu}\right)\lambda \\ &\left\| \begin{pmatrix} B_2^2 & B_2 B_3 \\ B_3^\tau B_2 & B_3^\tau B_3 + B_4^\tau B_4 \end{pmatrix} - \frac{A}{n}\mathbf{I}_H \right\|_F \leq \frac{\sqrt{p}\alpha}{n} \end{aligned} . \tag{24}$$

for a constant $\nu > 1$ *where* $\alpha \prec \frac{n\lambda}{p^{1/2}}$. *Then we have*

$$\lambda_{\max}\left(\boldsymbol{M}\boldsymbol{M}^\tau\right) > \frac{A}{n} - \frac{\sqrt{p}\alpha}{n} + \left(1-\frac{1}{2\nu}\right)\lambda. \tag{25}$$

Sliced Approximation Inequality The next result is referred to as "key lemma" in Lin et al. (2018a,b, 2019), which depends on the following sliced stable condition stated as Assumption A2.

Lemma 4 *Assume that Condition* (**A1**) *and the sliced stable condition A2 (for some* $\vartheta > 0$*) hold in the single index model* $y = f(\boldsymbol{\beta}^\tau \boldsymbol{x}, \epsilon)$. *Further assume that* $rank(var(x|y)) > 0$. *Let* $\widehat{\boldsymbol{\Lambda}}_H$ *be the SIR estimate of* $\boldsymbol{\Lambda} = var(\mathbb{E}[\boldsymbol{x} \mid y])$, *and let* $P_{\boldsymbol{\Lambda}}$ *be the projection matrix associated with the column space of* $\boldsymbol{\Lambda}$. *For any vector* $\beta \in \mathbb{R}^p$ *and any* $\nu > 1$, *let* $E_\beta(\nu) = \left\{ \left| \beta^\tau \left(P_{\boldsymbol{\Lambda}} \widehat{\boldsymbol{\Lambda}}_H P_{\boldsymbol{\Lambda}} - \boldsymbol{\Lambda}\right)\beta \right| \leq \frac{1}{2\nu}\beta^\tau \boldsymbol{\Lambda}\beta \right\}$. *There exist positive constants* C_1, C_2, C_3, *and* C_4 *such that for any* $\nu > 1$ *and* H *satisfying that* $H^\vartheta > C_4\nu$, *one has*

$$\mathbb{P}\left(\bigcap_\beta E_\beta\right) \geq 1 - C_1 \exp\left(-C_2 \frac{n\lambda_{\max}(\boldsymbol{\Lambda})}{H\nu^2} + C_3 \log(H)\right). \tag{26}$$

Proof of Theorems

Proof of Theorem 1 Theorem 1 follows from Lemmas 5 and 6. □

Lemma 5 *Assume that* $p^{1/2} \prec n\lambda_0$, *and let* τ_n *be a sequence such that* $\frac{\sqrt{p}}{n} \prec \tau_n \prec \lambda_0$. *Then, as* $n \to \infty$, *we have:*

i) Under H_0, *i.e., if* $y \perp\!\!\!\perp \boldsymbol{x}$, *then* $\lambda_{\max}(\widehat{\boldsymbol{\Lambda}}_H) < \frac{tr(\boldsymbol{\Sigma})}{n} + \tau_n$ *with probability converging to 1;*

ii) Under H_1, *if* $\lambda \succ \lambda_0$, *then* $\lambda_{\max}(\widehat{\boldsymbol{\Lambda}}_H) > \frac{tr(\boldsymbol{\Sigma})}{n} + \tau_n$ *with probability converging to 1.*

Proof

i) If $y \perp\!\!\!\perp \boldsymbol{x}$, we know that $\frac{1}{\sqrt{H}}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}_H$ is a $p \times H$ matrix with entries $i.i.d.$ to $N(0, \frac{1}{n})$. From Lemma 2, we know that

$$\lambda_{\max}\left(\frac{1}{H}\boldsymbol{X}_H^{\tau}\boldsymbol{X}_H\right) \leq \frac{tr(\boldsymbol{\Sigma})}{n} + \tau_n \tag{27}$$

with probability at least $1 - 4H^2 \exp\left(-\frac{Cn^2\tau_n^2}{H^2 p}\right)$ which $\to 1$ as $n \to \infty$.

ii) For any event ω, there exist $p \times p$ orthogonal matrix S and $H \times H$ orthogonal matrix T such that

$$S\boldsymbol{X}_H(\omega)T = \begin{pmatrix} Z_1 & 0 \\ Z_2 & Z_3 \\ 0 & Z_4 \end{pmatrix} \tag{28}$$

where Z_1, Z_2 are two scalars, Z_3 is a $1 \times (H-1)$ vector, and Z_4 is a $(p-2) \times (H-1)$ matrix. Lemmas 4 and 2 imply that there exist a constant A and an events set Ω, such that $\mathbb{P}(\Omega^c) \to 0$ as $n \to \infty$. For any $\omega \in \Omega$, one has

$$\begin{aligned} &\left(1 - \frac{1}{2\nu}\right)\lambda \leq Z_1^2 \leq \left(1 + \frac{1}{2\nu}\right)\lambda, \\ &\left\| \begin{pmatrix} Z_2^{\tau}Z_2 & Z_2^{\tau}Z_3 \\ Z_3^{\tau}Z_2 & Z_3^{\tau}Z_3 + Z_4^{\tau}Z_4 \end{pmatrix} - \frac{tr(\boldsymbol{\Sigma})}{n}\mathbf{I}_H \right\|_F \leq \frac{\sqrt{p}\alpha}{n}. \end{aligned} \tag{29}$$

Lemma 3 implies that

$$\lambda_{\max}\left(\frac{1}{H}\boldsymbol{X}_H\boldsymbol{X}_H\right)^{\tau} \geq \frac{tr(\boldsymbol{\Sigma})}{n} - \frac{\sqrt{p}}{n}\alpha + \left(1 - \frac{1}{2\nu}\right)\lambda \succ \frac{tr(\boldsymbol{\Sigma})}{n} + \tau_n. \tag{30}$$

□

Lemma 6 *Assume that* $\frac{s\log(p)}{n} \prec \lambda_0$. *Let* τ_n *be a sequence such that* $\frac{s\log(p)}{n} \prec \tau_n \prec \lambda_0$. *Then, as* $n \to \infty$, *we have:*

i) If $y \perp\!\!\!\perp \boldsymbol{x}$, *then* $\lambda_{\max}^{(ks)}\left(\widehat{\boldsymbol{\Lambda}}_H\right) < \tau_n$ *with probability converging to 1;*

ii) If $\lambda \succ \lambda_0$, *then* $\lambda_{\max}^{(ks)}\left(\widehat{\boldsymbol{\Lambda}}_H\right) > \tau_n$ *with probability converging to 1.*

Proof

i) If $y \perp\!\!\!\perp \boldsymbol{x}$, we know that $\frac{1}{\sqrt{H}}\boldsymbol{E}_H = \frac{1}{\sqrt{H}}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}_H$ is a $p \times H$ matrix with entries $i.i.d.$ to $N(0, \frac{1}{n})$. Thus

$$\lambda_{\max}^{(ks)}\left(\widehat{\boldsymbol{\Lambda}}_H\right) = \lambda_{\max}^{(ks)}\left(\frac{1}{H}\boldsymbol{\Sigma}^{1/2}\boldsymbol{E}_H\boldsymbol{E}_H^{\tau}\boldsymbol{\Sigma}^{1/2}\right) \text{ and } \lambda_{\max}^{(ks)}\left(\boldsymbol{D}^{1/2}\boldsymbol{E}_H\boldsymbol{E}_H^{\tau}\boldsymbol{D}^{1/2}\right)$$

are identically distributed where $\boldsymbol{D}$ is diagonal matrix consisting of the eigenvalues of $\boldsymbol{\Sigma}$. For any subset $S \subset [p]$, let $\boldsymbol{X}_{H,S} = \boldsymbol{D}_S^{1/2}\boldsymbol{E}_{S,H}$ where $\boldsymbol{E}_{S,H}$ is a submatrix of $\boldsymbol{E}_H$ consisting of the rows in S. Note that

$$\lambda_{\max}\left(\left(\boldsymbol{D}^{1/2}\boldsymbol{E}_H\boldsymbol{E}_H^{\tau}\boldsymbol{D}^{1/2}\right)_S\right) = \lambda_{\max}\left(\frac{1}{H}\boldsymbol{X}_{H,S}\boldsymbol{X}_{H,S}^{\tau}\right). \tag{31}$$

Thus, by Lemma 3, we have

$$\lambda_{\max}\left(\frac{1}{H}\boldsymbol{X}_{H,S}\boldsymbol{X}_{H,S}^{\tau}\right) < tr(\boldsymbol{D}_S)/n + \alpha \le \frac{ks\lambda_{\max}(\boldsymbol{\Sigma})}{n} + \alpha \tag{32}$$

with probability at least $1 - 4H^2\exp\left(-\frac{Cn^2\alpha^2}{H^2 s}\right)$. Let $\alpha = C\frac{s\log(p)}{n}$ for some sufficiently large constant C. Since $\binom{p}{ks} \le \left(\frac{ep}{ks}\right)^{ks}$, we know that $\lambda_{\max}^{(ks)}(\widehat{\boldsymbol{\Lambda}}_H) \le C\frac{s\log(p)}{n} \prec \tau_n$ with probability converging to 1.

ii) Let $\boldsymbol{\eta}$ be the eigenvector associated with the largest eigenvalue of $\boldsymbol{\Lambda}$. Thus $|supp(\boldsymbol{\eta})| = ks$. From Lemma 4, we know that

$$\widehat{\lambda}_{\max}^{(ks)}(\widehat{\boldsymbol{\Lambda}}_H) \ge \boldsymbol{\eta}^{\tau}\widehat{\boldsymbol{\Lambda}}_H\boldsymbol{\eta} \ge \left(1 - \frac{1}{2\nu}\right)\lambda \tag{33}$$

with probability converging to 1. Thus, $\lambda_{\max}^{(ks)}\left(\widehat{\boldsymbol{\Lambda}}_H\right) > \tau_n$ with probability converging to 1.

□

Proof of Theorem 2 Theorem 2 follows from the Theorem 1 and the following Lemma 7. □

Lemma 7 *Assume that* $\frac{1}{\sqrt{n}} \prec \lambda_0^a$. *Let* τ_n *be a sequence such that* $\frac{1}{\sqrt{n}} \prec \tau_n \prec \lambda_0^a$, $\tau_n \to 0$. *Then we have:*

i) *If* $y \perp\!\!\!\perp \boldsymbol{x}$, *then* $t < \tau_n$ *with probability converging to 1.*
ii) *If* $\lambda \succ \lambda_0^a$, *then* $t > \tau_n$ *with probability converging to 1.*

Proof

(i) Since $y \perp\!\!\!\perp \boldsymbol{x}$,we know that $\mathbb{E}[t] = 0$. Let $z_j = y_j^2 - 1$; then we have

$$\mathbb{P}\left(\frac{1}{n}\sum_j z_j > \tau_n\right) \le \exp\left(-Cn\tau_n^2\right) \tag{34}$$

for some constant C. In other words, the probability of $t > \tau_n$ converges to 0 as $n \to \infty$.

(ii) If $\lambda \succ \lambda_0^a$, we have $var(f(z)) \geq C\lambda$ and $\mathbb{E}[y^2 - 1] \geq C\lambda$ for some constant C. Let $z_j = y_j^2 - 1$, $j = 1, \ldots, n$. Since $f(x_j)$, $j = 1, \ldots, n$ are sub-Gaussian, we know that

$$\mathbb{P}\left(\frac{1}{n}\sum_j z_j > \mathbb{E}[y^2 - 1] + \delta\right) \leq \exp\left(-Cn\delta^2\right). \tag{35}$$

By choosing $\delta = C\mathbb{E}[y^2 - 1]$ for some constant C, we know that the probability of $t \geq (C + 1)\lambda \succ \tau_n$ converges to 1.

□

Proof of Theorem 3 Theorems 3 and 4 follow from the following lemma, the Theorems 1 and 2. □

Lemma 8 *Assume that $\frac{s\log(p)}{n} \prec \lambda_0$. Let τ_n be a sequence such that $\frac{s\log(p)}{n} \prec \tau_n \prec \lambda_0$. Then we have:*

$i)$ If $y \perp\!\!\!\perp \boldsymbol{x}$, then $\widetilde{\lambda}_{\max}^{(ks)}\left(\widehat{\boldsymbol{\Lambda}}_H\right) < \tau_n$ with probability converging to 1;
$ii)$ If $\lambda \succ \lambda_0$, then $\widetilde{\lambda}_{\max}^{(ks)}\left(\widehat{\boldsymbol{\Lambda}}_H\right) > \tau_n$ with probability converging to 1.

Proof

$i)$ Under H_0, i.e., $y \perp\!\!\!\perp \boldsymbol{x}$, the entries of $\frac{1}{\sqrt{H}}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}_H$ are identically distributed as $N(0, \frac{1}{n})$. Thus, if $1 \prec \alpha \prec \frac{n\tau_n}{s\log(p)}$, we have

$$\max_{(i,j)} \left|\widehat{\boldsymbol{\Lambda}}_H(i, j)\right| \leq \frac{\alpha\log(p)}{n} \tag{36}$$

with probability at least $1 - p^2\exp\left(-C\alpha^2\log(p)^2\right)$ for some constant C which converges to 1 as $n \to \infty$. Since (see, e.g., Lemma 6.1 in Berthet and Rigollet (2013a))

$$\widetilde{\lambda}_{\max}^{(ks)}\left(\widehat{\boldsymbol{\Lambda}}_H\right) \leq \lambda_{\max}\left(st_{\frac{\alpha\log(p)}{n}}\left(\widehat{\boldsymbol{\Lambda}}_H\right)\right) + ks\frac{\alpha\log(p)}{n} \prec \tau_n \tag{37}$$

where $st_z(A)_{i,j} = sign(A_{i,j})(A_{i,j} - z)_+$, we know that (i) holds.

$ii)$ Follows from that $\widetilde{\lambda}_{\max}^{(ks)}(\widehat{\boldsymbol{\Lambda}}_H) \geq \lambda_{\max}^{(ks)}(\widehat{\boldsymbol{\Lambda}}_H)$. □

References

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

E. Arias-Castro, E.J. Candès, A. Durand, Detection of an anomalous cluster in a network. Ann. Stat. **39**(1), 278–304 (2011a)

E. Arias-Castro, E.J. Candès, Y. Plan, Global testing under sparse alternatives: ANOVA, multiple comparisons and the higher criticism. Ann. Stat. **39**(5), 2533–2556 (2011b)

E. Arias-Castro, S. Bubeck, G. Lugosi, Detection of correlations. Ann. Stat. **40**(1), 412–435 (2012)

R.F. Barber, E.J. Candès, Controlling the false discovery rate via knockoffs. Ann. Stat. **43**(5), 2055–2085 (2015)

Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B **57**(1), 289–300 (1995)

Q. Berthet, P. Rigollet, Optimal detection of sparse principal components in high dimension. Ann. Stat. **41**(4), 1780–1815 (2013a)

Q. Berthet, P. Rigollet, Optimal detection of sparse principal components in high dimension. Ann. Stat. **41**(4), 1780–1815 (2013b). https://doi.org/10.1214/13-AOS1127

C.H. Chen, K.C. Li, Can SIR be as popular as multiple linear regression? Stat. Sin. **8**(2), 289–316 (1998)

A. d'Aspremont, L.E. Ghaoui, M.I. Jordan, G.R. Lanckriet, A direct formulation for sparse PCA using semidefinite programming, in *Advances in Neural Information Processing Systems* (2005), pp. 41–48

A. d'Aspremont, F. Bach, L. El Ghaoui, Approximation bounds for sparse principal component analysis. Math. Program. **148**(1–2), 89–110 (2014)

D. Donoho, J. Jin, Higher criticism for detecting sparse heterogeneous mixtures. Ann. Stat. **32**(3), 962–994 (2004)

N. Duan, K.C. Li, Slicing regression: a link-free regression method. Ann. Stat. **19**(2), 505–530 (1991)

T. Hsing, R.J. Carroll, An asymptotic theory for sliced inverse regression. Ann. Stat. **20**(2), 1040–1061 (1992)

Y.I. Ingster, A.B. Tsybakov, N. Verzelen, Detection boundary in sparse regression. Electron. J. Stat. **4**, 1476–1526 (2010)

L. Janson, R.F. Barber, E. Candes, Eigenprism: inference for high dimensional signal-to-noise ratios. J. R. Stat. Soc. Ser. B **79**(4), 1037–1065 (2017)

P. Ji, J. Jin, UPS delivers optimal phase diagram in high-dimensional variable selection. Ann. Stat. **40**(1), 73–103 (2012)

P. Ji, Z. Zhao, Rate optimal multiple testing procedure in high-dimensional regression. arXiv preprint arXiv:1404.2961 (2014)

K.C. Li, Sliced inverse regression for dimension reduction. J. Am. Stat. Assoc. **86**(414), 316–327 (1991)

Q. Lin, X. Li, D. Huang, J.S. Liu, On the optimality of sliced inverse regression for high dimension (2018a)

Q. Lin, Z. Zhao, J.S. Liu, On consistency and sparsity for sliced inverse regression in high dimensions. Ann. Stat. **46**(2), 580–610 (2018b)

Q. Lin, Z. Zhao, J.S. Liu, Sparse sliced inverse regression via lasso. J. Am. Stat. Assoc. **114**(528), 1726–1739 (2019). PMID: 32952234. https://doi.org/10.1080/01621459.2018.1520115

M. Neykov, Q. Lin, J.S. Liu, Signed support recovery for single index models in high-dimensions. Ann. Math. Sci. Appl. **1**(2), 379–426 (2016)

W. Su, E.J. Candes, SLOPE is adaptive to unknown sparsity and asymptotically minimax. Ann. Stat. **44**(3), 1038–1068 (2016)

L. Zhu, B. Miao, H. Peng, On sliced inverse regression with high-dimensional covariates. J. Am. Stat. Assoc. **101**(474), 640–643 (2006)