Signed Support Recovery for Single Index Models in High-Dimensions

Matey Neykov* Qian Lin† Jun S. Liu‡

Abstract

In this paper we study the support recovery problem for single index models $Y = f(\boldsymbol{X}^\intercal \boldsymbol{\beta}, \varepsilon)$, where f is an unknown link function, $\boldsymbol{X} \sim N_p(0, \mathbb{I}_p)$ and $\boldsymbol{\beta}$ is an s-sparse unit vector such that $\boldsymbol{\beta}_i \in \{\pm\frac{1}{\sqrt{s}}, 0\}$. In particular, we look into the performance of two computationally inexpensive algorithms: (a) the diagonal thresholding sliced inverse regression (DT-SIR) introduced by Lin et al. (2015); and (b) a semi-definite programming (SDP) approach inspired by Amini & Wainwright (2008). When $s = O(p^{1-\delta})$ for some $\delta > 0$, we demonstrate that both procedures can succeed in recovering the support of $\boldsymbol{\beta}$ as long as the *rescaled sample size* $\Gamma = \frac{n}{s\log(p-s)}$ is larger than a certain critical threshold. On the other hand, when Γ is smaller than a critical value, any algorithm fails to recover the support with probability at least $\frac{1}{2}$ asymptotically. In other words, we demonstrate that both DT-SIR and the SDP approach are optimal (up to a scalar) for recovering the support of $\boldsymbol{\beta}$ in terms of sample size. We provide extensive simulations, as well as a real dataset application to help verify our theoretical observations.

Keywords: Single index models, Sliced inverse regression, Sparsity, Support recovery, High-dimensional statistics, Semidefinite programming

1 Introduction

Due to the recent advances in technology, collecting data becomes a routine. The '*small n, large p*' characteristic of modern data brings new challenges in interpreting and processing it. Dimension reduction and variable selection procedures become an indispensable step in data exploration. Regrettably, the majority of the classical algorithms were developed to work in the regime $p \ll n$, and hence one needs to exercise caution when applying existing methods in a high dimensional setting. A firm understanding of the limitations of classical dimension reduction procedures in the modern $p \gg n$ regime, will help facilitate their appropriate application.

For example, the archetypical unsupervised dimension reduction procedure — principal component analysis (PCA), has been widely and successfully applied in a range of scientific problems (Price et al. 2010, Brenner et al. 2000, e.g.). The behavior of PCA in high dimensions has been well studied in the recent years. In Johnstone & Lu (2004, 2009), Paul (2007), it was shown that in the spiked covariance model, PCA succeeds if and only if $\lim \frac{p}{n} \neq 0$. This result stimulated the statistical community to discuss the minimax rate of estimating the principal space under sparsity

*Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544
†Center of Mathematical Sciences and Applications, Harvard University, Cambridge, MA 02138
‡Department of Statistics, Harvard University, Cambridge, MA 02138

constraints (see e.g., Cai et al. (2013), Vu & Lei (2012) and references therein) and the tradeoff between the statistical and computational efficiency (see e.g., Berthet & Rigollet (2013)).

Another line of dimension reduction research, studies the so-called sufficient dimension reduction (SDR). The goal of SDR is to find the minimal subspace $\mathcal{S}$ such that $Y \perp\!\!\!\perp \boldsymbol{X} \big| P_{\mathcal{S}}\boldsymbol{X}$, where $\boldsymbol{X} \in \mathbb{R}^p$ is the predictor vector and $Y \in \mathbb{R}$ is the response (Cook et al. 2004, Li 1991, Cook & Ni 2005, e.g.). Unlike its unsupervised counterpart, much less attention has been paid to how SDR algorithms behave in a high dimensional setting. The optimal estimation rates of SDR algorithms in terms of sparsity (s), dimensionality (p), and sample size (n) are unclear.

Sliced inverse regression, proposed by Li (1991), is one of the most popular SDR methods for estimating the space $\mathcal{S}$. When the dimensionality p is larger than or comparable to the sample size n, sparsity assumptions are often imposed on the loading vector $\boldsymbol{\beta}$ (Li & Nachtsheim 2006, e.g.). Lin et al. (2015) proved that in fact $\mathbb{E}[\angle(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta})] > 0$ if $\rho = \lim \frac{p}{n} \neq 0$ and $\sin\big(\angle(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta})\big) = 0$ when $\rho = 0$ where $\widehat{\boldsymbol{\beta}}$ is the SIR estimator of $\boldsymbol{\beta}$. In other words, the SIR estimator $\widehat{\boldsymbol{\beta}}$ is consistent (up to a sign) if and only if $\rho = \lim \frac{p}{n} = 0$. One implication of this result is that in order to estimate $\boldsymbol{\beta}$, structural assumptions such as sparsity are necessary in the high dimensional setting.

In the present paper, inspired by Amini & Wainwright (2008), we investigate the support recovery problem for single index models (SIM) (1), under the sparsity assumption $\|\boldsymbol{\beta}\|_0 = s$ in the regime $s = O(p^{1-\delta})$ for some $\delta > 0$. More formally we study the models:

$$Y = f(\boldsymbol{X}^\intercal \boldsymbol{\beta}, \varepsilon) \text{ with } \boldsymbol{X} \sim N_p(0, \mathbb{I}_p), \tag{1}$$

where the noise ε is independent of $\boldsymbol{X}$ and $f, \varepsilon, \boldsymbol{\beta}$ belong to the class:

$$\begin{aligned} \mathcal{F}_A = \Big\{ (f, \varepsilon, \boldsymbol{\beta}) : \mathrm{Var}(\mathbb{E}[Z|f(Z,\varepsilon)]) \geq A \text{ where } Z \sim N(0,1), \\ \boldsymbol{\beta}_i \in \Big\{ \pm \frac{1}{\sqrt{s}}, 0 \Big\} \text{ and } (f, \varepsilon) \text{ is } \textit{sliced stable} \text{ (see Definition 1)} \Big\}. \end{aligned} \tag{2}$$

Notice that no generality is lost in assuming that the vector $\boldsymbol{\beta}$ is a unit vector since otherwise model (1) is not identifiable. Model class (2), further assumes the idealized setting where all non-zero coordinates of $\boldsymbol{\beta}$ have a signal strength of exactly the same magnitude, which simplifies our presentation while preserving the inherent complexity of the support recovery problem. We believe that studying support recovery in SIM, would bring us insightful understanding of SIR and other SDR algorithms.

In this paper, we study two procedures for signed support recover of SIM (1): the DT-SIR introduced by Lin et al. (2015) and the SDP approach inspired by Amini & Wainwright (2008). We let $\Gamma = \frac{n}{s \log(p-s)}$ be the *rescaled sample size*. Our main contribution is to establish the existence of constants $\omega > 0$ and $\Omega > 0$ such that when $\Gamma > \Omega$ both DT-SIR and SDP approaches recover the signed support of $\boldsymbol{\beta}$ correctly, with probability converging to 1 asymptotically. Conversely, we show that when $\Gamma < \omega$ any algorithm fails to recover the support of $\boldsymbol{\beta}$ with probability at least 1/2. In other words, we show that the optimal sample size of the support recovery problem of model (1) is of the order $s \log(p-s)$. To the best of our knowledge, this optimality result, regarding the sample size of SIM (1), has not been previously discussed in the literature. Our second contribution is, to establish a sliced stability conjecture formulated by Lin et al. (2015), under the SIM case. We demonstrate that classical conditions of Hsing & Carroll (1992) imply sliced stability in Section 2.1. This technical result might be of independent interest, especially when one wants to discuss the optimal rate problem for other SDR algorithms such as SAVE. We further develop a novel tool

along the way of our analysis, which may represent further interest — a concentration inequality stated under Lemma 7.

1.1 Related Work

In the fixed p setting, the first asymptotic results on SIR appeared in the seminal papers (Duan & Li 1991, Hsing & Carroll 1992). Later on Zhu et al. (2006) allowed p to diverge slowly with n and established asymptotics in the regime $p = o(n^{1/2})$. In the super high-dimensional setting where $p \gg n$, several algorithms, hinging on regularization such as LASSO (Tibshirani et al. 1997) and Dantzig Selector (Candes & Tao 2007) were proposed by Li & Nachtsheim (2006), Yu et al. (2013), but these algorithms are not concerned with support recovery. Moreover, the algorithm suggested by Li & Nachtsheim (2006) did not come with theoretical guarantees, and in Yu et al. (2013) it is not allowed for s to scale with p and n. A generic variable selection procedure was suggested in Zhong et al. (2012), with guarantees of support recovery, in a more general setting than our present paper, but with a much more restrictive relationship ($p = o(n^{1/2})$) than the one we consider.

In the parallel line of research on sparse PCA there have been more developments. In Johnstone & Lu (2004) the algorithm Diagonal Thresholding (DT) was suggested, to deal with the spiked-covariance model. It was later analyzed by Amini & Wainwright (2008), who showed that support recovery is achieved by DT in the sparse spiked covariance model, provided that $n \gtrsim s^2 \log(p)$. Amini & Wainwright (2008) further showed an information theoretic obstruction, in that no algorithm can recover the support of the principal eigenvector if $n \lesssim s \log(p)$. A computationally inefficient algorithm that succeeds in support recovery with high probability as long as $n \gtrsim s \log(p)$ is exhaustively scanning through all $\binom{p}{s}$ subsets of the coordinates of the principal eigenvector. In order to find feasible procedures, Amini & Wainwright (2008) studied a semidefinite programming (SDP) estimator originally suggested in d'Aspremont et al. (2008) — and showed that if $n \gtrsim s \log(p)$ and the SDP has a rank 1 solution, this solution can recover the signed support with high probability. Surprisingly however, Krauthgamer et al. (2013) showed that the rank 1 condition, does not hold if $s^2 \log(p) \gtrsim n \gtrsim s \log(p)$. In contrast to the PCA case, our current paper argues that if one is concerned with the support recovery for SIM in the class $\mathcal{F}_A$, no computational and statistical tradeoff exists in the regime $s = O(p^{1-\delta})$, since the computationally tractable algorithms DT-SIR and SDP algorithms solve the support recovery problem of SIM with optimal sample size.

1.2 Preliminaries and Notation

We first briefly recall the SIR procedure for SIM. Suppose we observe $n = Hm$ independent and identically distributed (i.i.d.) samples $(Y_i, \boldsymbol{X}_i)$ from model (1). SIR proceeds to sort and divide the data into H slices of equal size, according to the order statistics $Y_{(i)}$ of Y_i. Let the concomitant of $Y_{(i)}$ be $X_{(i)}$. We will also use the double subscript notation $X_{h,i}$ for $X_{((h-1)m+i)}$. Let $S_h = (Y_{((h-1)m)}, Y_{(hm)}], 1 \le h < H$, and $S_H = (Y_{((H-1)m)}, +\infty)$ denote the random intervals partitioned by the points $Y_{(jm)}$(with $Y_{(0)} = -\infty$), $j \le H-1$. Furthermore let $\mathfrak{m}_j(Y) = \mathbb{E}[\boldsymbol{X}^j|Y]$ denote the j^{th} coordinate of the centered inverse regression curve $\mathbb{E}[\boldsymbol{X}|Y]$, and $\boldsymbol{\mu}_h^j$ denotes $\mathbb{E}[\boldsymbol{X}^j|Y \in S_h]$. Note that conditionally on the values $Y_{((h-1)m)}$ and $Y_{(hm)}$ the quantities S_h and $\boldsymbol{\mu}_h^j$ become deterministic. For any integer $k \in \mathbb{N}$ put $[k] = \{1, \dots, k\}$ for brevity. Let $\overline{\boldsymbol{X}}_{h,S}^j = \frac{1}{|S|}\sum_{i\in S} \boldsymbol{X}_{h,i}^j$, where $S \subset [m]$, $j \in [p]$. If $S = [m]$ we omit it from the notation, i.e., $\overline{\boldsymbol{X}}_h^j = \frac{1}{m}\sum_{i=1}^m \boldsymbol{X}_{h,i}^j$. Put $\overline{\boldsymbol{X}}^j = \frac{1}{H}\sum_{h=1}^H \overline{\boldsymbol{X}}_h^j$ for the global average. In terms of this notation, the SIR estimator $\mathbf{V}$ of the conditional covariance

matrix is given by

$$\mathbf{V}^{jk} = \frac{1}{H}\sum_{h=1}^{H} \overline{\boldsymbol{X}}_h^j \overline{\boldsymbol{X}}_h^k, \tag{3}$$

entrywise. The SIR estimator $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is defined as the principal eigenvector of $\mathbf{V}$. We will denote the support of the vector $\boldsymbol{\beta}$ by $\mathcal{S}_{\boldsymbol{\beta}}$, i.e. we set $\mathcal{S}_{\boldsymbol{\beta}} := \mathrm{supp}(\boldsymbol{\beta}) = \{j : \boldsymbol{\beta}_j \neq 0\}$.

Finally, we need some standard notations. For a vector $\mathbf{v}$, let $\|\mathbf{v}\|_p$ denote the usual ℓ_p norm for $1 \leq p \leq \infty$ and $\|\mathbf{v}\|_0 = |\mathrm{supp}(\mathbf{v})|$. For a real random variable X, let $\|X\|_{\psi_2} = \sup_{p\geq 1} p^{-1/2}(\mathbb{E}|X|^p)^{1/p}$, $\|X\|_{\psi_1} = \sup_{p\geq 1} p^{-1}(\mathbb{E}|X|^p)^{1/p}$. Recall that a random variable is called *sub-Gaussian* if $\|X\|_{\psi_2} < \infty$ and *sub-exponential* if $\|X\|_{\psi_1} < \infty$. For a matrix $\mathbf{M} \in \mathbb{R}^{d\times d}$ and two sets $S_1, S_2 \subset [d]$, by double indexing $\mathbf{M}_{S_1,S_2}$ we denote the sub-matrix of $\mathbf{M}$ with entries $\mathbf{M}_{ij}$ for $i \in S_1, j \in S_2$. Furthermore, for a $d \times d$ matrix $\mathbf{M}_{d\times d}$, let $\|\mathbf{M}\|_{\max} = \max_{jk} |\mathbf{M}_{jk}|$ and $\|\mathbf{M}\|_{p,q} = \sup_{\|\mathbf{v}\|_p=1} \|\mathbf{M}\mathbf{v}\|_q$. In particular, we have $\|\mathbf{M}\|_{2,2} = \max_{i\in[d]}\{\sigma_i(\mathbf{M})\}$, where $\sigma_i(\mathbf{M})$ is the i^{th} singular value of $\mathbf{M}$, and $\|\mathbf{M}\|_{\infty,\infty} = \max_{i\in[d]} \sum_{j=1}^{d} |\mathbf{M}_{ij}|$. Let $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(Y_i \leq x)$ denote the empirical distribution of the Y sample, and Φ (*resp.* ϕ) denote the cdf (*resp.* pdf) of a standard normal random variable. We will also occasionally use the abbreviation WLOG to stand for "without loss of generality".

1.3 Organization

The paper is organized as follows. The main results, and confirmatory numerical studies and a real data analysis are presented in Section 2 and Section 3 respectively. Proofs of our main results are included in Sections 4 and 5 while technical lemmas are deferred to Appendix A. A brief discussion on the potential directions is included in Section 6. In Appendix B we provide several concrete examples of SIM.

2 Main Results

In this section, we first prove a conjecture regarding the coordinate-wise sliced stability conditions introduced by Lin et al. (2015). Second, we establish the optimal rate of support recovery in terms of the sample size. Throughout the remainder of the paper, we assume that Y is a continuously distributed random variable.

2.1 Sliced Stability

Our proofs of signed support recovery rely on the following property of the inverse regression curve of a SIM.

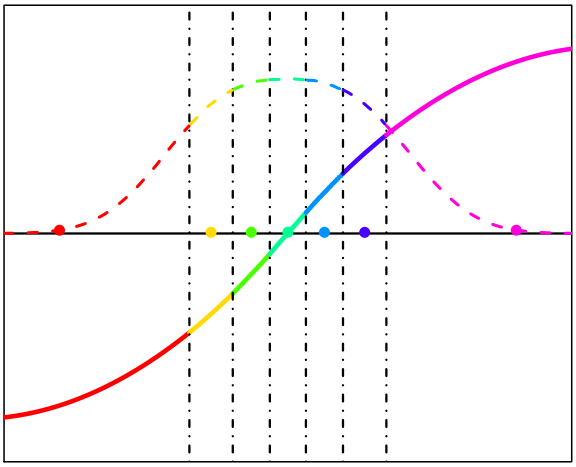
Definition 1 (Sliced Stability). *We call the pair (f, ε) sliced stable iff there exist constants $0 < l < 1, 1 < K, 0 < M$, such that for any $H \in \mathbb{N}, H > M$, and all partitions of $\mathbb{R} = \{a_1 = -\infty, \ldots, a_{H+1} = +\infty\}$ with $\frac{l}{H} \leq \mathbb{P}(a_h < Y \leq a_{h+1}) \leq \frac{K}{H}$ there exist two constants $0 \leq$*

$\kappa(l, K, M) < 1$, $C(l, K, M) > 0$ *such that for all* $j \in \mathcal{S}_{\boldsymbol{\beta}}$[1]*:*

$$\sum_{h=1}^{H} \mathrm{Var}[\mathfrak{m}_j(Y)|a_h < Y \leq a_{h+1}] \leq C(l, K, M) H^{\kappa(l,K,M)} \mathrm{Var}[\mathfrak{m}_j(Y)]. \tag{4}$$

The sliced stability assumption, is an implicit assumption on the function f and the error distribution ε. If $\kappa = 0$, the condition means that the cumulative relative variability of the inverse regression curve is bounded for all slicing schemes with sufficiently small slices. If $\kappa > 0$ the cumulative relative variability of the inverse regression curve is allowed to scale sub-linearly with the number of slices (see also Figure 1). Intuitively, sliced stability allows us to ensure that the estimates $\mathbf{V}^{jj}$ of $\mathrm{Var}[\mathfrak{m}_j(Y)]$ become increasingly accurate the more slices we introduce (see Lemma 3 for a more rigorous treatment). Relying on the subsequent developments of this section, in Example 2 and Remark 3 of Appendix B we demonstrate that models of the form $Y = G(h(\boldsymbol{X}^\intercal \boldsymbol{\beta}) + \varepsilon)$, where G, h are continuous and monotone and ε is a log-concave random variable, satisfy the sliced stability assumption.

Figure 1: Below, with a solid line we plot the standardized inverse regression curve $\mathfrak{m}(y) := \frac{\mathrm{sign}(\beta_j)\mathfrak{m}_j(y)}{\sqrt{\mathrm{Var}(\mathfrak{m}_j(Y))}}, j \in \mathcal{S}_{\boldsymbol{\beta}}$, for the model $Y = 2\,\mathrm{atan}(\boldsymbol{X}^\intercal \boldsymbol{\beta}) + N(0, 1)$ (see also (14)). The different colored parts of the dashed curve represent the conditional densities of $Y|Y \in S_h$, where S_h for $h \in [H]$ (with H = 7) are the slices illustrated by punctured vertical black lines. Finally the seven points' vertical-axis values represent the variances $\mathrm{Var}[\mathfrak{m}(Y)|Y \in S_h]$. Sliced stability ensures that the average value of the variances $\mathrm{Var}[\mathfrak{m}(Y)|Y \in S_h]$ has a decay rate of the order $H^{\kappa-1}$.



[1]Observe that (4) automatically holds for $j \notin \mathcal{S}_{\boldsymbol{\beta}}$ in our case, since both the LHS and the RHS of (4) are 0 in this case.

Definition 1 is the sliced stability definition from Lin et al. (2015) restated in terms of the SIM. Lin et al. (2015) conjectured that the sliced stability condition could be implied from the well accepted conditions proposed by Hsing & Carroll (1992), which we state below with a slight modification. The first contribution of this paper is proving this conjecture in the case of SIM.

Let $\mathcal{A}_H(l, K)$, with $1 < K, 0 < l < 1$, denote all partitions of $\mathbb{R}$ of the sort $\{-\infty = a_1 \leq a_2 \leq \ldots \leq a_{H+1} = +\infty\}$, such that $\frac{l}{H} \leq \mathbb{P}(a_h \leq Y \leq a_{h+1}) \leq \frac{K}{H}$. Moreover, for any fixed $B \in \mathbb{R}$, let $\Pi_r(B)$ denote all possible partitions of the closed interval $[-B, B]$ into r points $-B \leq b_1 \leq b_2 \leq \ldots \leq b_r \leq B$. Define the normalized version of the centered inverse regression curve $\mathfrak{m}(y) := \frac{\text{sign}(\beta_j)\mathfrak{m}_j(y)}{\sqrt{\text{Var}(\mathfrak{m}_j(Y))}}, j \in [p]$[2], and let $\mathfrak{m}$ satisfy the following smoothness condition:

$$\lim_{r\to\infty} \sup_{b\in\Pi_r(B)} r^{-1/(2+\xi)} \sum_{i=2}^{r} |\mathfrak{m}(b_i) - \mathfrak{m}(b_{i-1})| = 0, \tag{5}$$

for any $B > 0$ for some fixed $\xi > 0$. Note that as mentioned in Hsing & Carroll (1992), assumption (5) is weaker than assuming that $\mathfrak{m}$ is of bounded variation, and furthermore the bigger the ξ the more stringent this assumption becomes. In addition, assume that there exists $B_0 > 0$ and a non-decreasing function $\widetilde{\mathfrak{m}} : (B_0, \infty) \mapsto \mathbb{R}$, such that:

$$|\mathfrak{m}(x) - \mathfrak{m}(y)| \leq |\widetilde{\mathfrak{m}}(|x|) - \widetilde{\mathfrak{m}}(|y|)|, \text{ for } x, y \in (-\infty, -B_0) \text{ or } (B_0, +\infty), \tag{6}$$

and moreover, $\mathbb{E}[|\widetilde{\mathfrak{m}}(|Y|)|^{(2+\xi)}] < \infty$ (where in the expectation we set $\widetilde{\mathfrak{m}}(y) = 0$ for $|y| \leq B_0$). We are now in a position to formulate the following:

Proposition 1. *Assume that the standardized centered inverse regression curve satisfies properties (5) and (6) for some $\xi > 0$. Then we have that for any fixed $0 < l < 1 < K$:*

$$\lim_{H\to\infty} \sup_{a\in\mathcal{A}_H(l,K)} \frac{1}{H^{2/(2+\xi)}} \sum_{h=1}^{H} \text{Var}[\mathfrak{m}(Y)|a_h < Y \leq a_{h+1}] \to 0. \tag{7}$$

We defer the proof of Proposition 1 to Appendix A. It is clear however that (7) implies the existence of constants M, $C(l, K, M)$ such that (4) holds, with $\kappa = \frac{2}{2+\xi} < 1$.

2.2 Optimal sample size for support recovery

Recall that $\Gamma = \frac{n}{s\log(p-s)}$ is the rescaled sample size. First we establish an information theoretic barrier of the support recovery problem, i.e. we show that there exists a positive constant ω such that, when $\Gamma < \omega$ every algorithm fails with probability at least $1/2$. Then we prove that two algorithms — DT-SIR and an SDP based procedure achieve this bound, i.e. we demonstrate that there exists a positive constant Ω, such that, when $\Gamma > \Omega$, these two algorithms successfully recover the signed support with probability 1 as $n \to \infty$.

2.2.1 Information theoretic barrier

Intuitively, when the sample size is small, no algorithm is expected to be able to recover the support successfully. In fact, let us consider the simple linear regression model:

$$Y = \boldsymbol{X}^\intercal\boldsymbol{\beta} + \varepsilon \text{ where } \boldsymbol{X} \sim N_p(0, \mathbb{I}_p), \varepsilon \sim N(0, \sigma^2), \tag{8}$$

[2] By symmetry $\text{sign}(\beta_j)\mathfrak{m}_j(y) = \text{sign}(\beta_i)\mathfrak{m}_i(y)$ for $i, j \in \mathcal{S}_{\boldsymbol{\beta}}$

where $\boldsymbol{\beta}$ is a unit vector such that $\boldsymbol{\beta}_i \in \{\pm\frac{1}{\sqrt{s}}, 0\}$. This model belongs to the class $\mathcal{F}_{(\sigma^2+1)^{-1}}$ (to verify sliced stability refer to Example 2 and Remark 3 of Appendix B). The following result, which is obtained by a slight modification of the arguments in (Wainwright 2009), shows that support recovery is infeasible if Γ is small, by effectively arguing that no algorithm works even for model (8).

Proposition 2. *Suppose there are n observations from the model (8). Then there exists a positive constant ω, such that if*

$$\Gamma = \frac{n}{s\log(p-s)} < \omega,$$

any algorithm for support support recovery will have errors with probability at least $\frac{1}{2}$ asymptotically.

2.2.2 Optimality of DT-SIR

Recall that we are working with the class of models $\mathcal{F}_A$ (2). One key observation is the following signal strength result.

Lemma 1. *For any $j \in \mathcal{S}_{\boldsymbol{\beta}} = supp(\boldsymbol{\beta})$, one has:*

$$\frac{A}{s} \leq \mathrm{Var}(\mathfrak{m}_j(Y)) \leq \frac{1}{s}, \tag{9}$$

and $\mathrm{Var}(\mathfrak{m}_j(Y)) = 0$ for $j \notin \mathcal{S}_{\boldsymbol{\beta}}$.

By Lemma 1 WLOG we can assume that there exists a $C_V > 0$ such that $\mathrm{Var}[\mathfrak{m}_j(Y)] = \frac{C_V}{s}$ for any $j \in \mathcal{S}_{\boldsymbol{\beta}}$. Now we are ready to discuss the properties of the DT algorithm which we formulate below. In this paper, we assume that the sparsity s is known.

Algorithm 1: DT algorithm

input: $(Y_i, \boldsymbol{X}_i)_{i=1}^n$: data, H: number of slices, s: the sparsity of $\boldsymbol{\beta}$

1. Calcluate $\mathbf{V}^{jj}, j \in [p]$ – according to formula (3);
2. Collect the s highest $\mathbf{V}^{jj}$ into the set $\widehat{S}$;
3. Output the set $\{j : \mathbf{V}^{jj} \in \widehat{S}\}$.

Recall the definitions (see (3)) of $\mathbf{V}$ and $\mathbf{V}^{jj}$ the estimates of $\mathrm{Cov}[\mathbb{E}[\boldsymbol{X}|Y]]$ and $\mathrm{Var}[\mathfrak{m}_j(Y)]$ respectively. To obtain the result on sample size optimality of Algorithm 1, a key point is to establish a concentration inequality for $\mathbf{V}^{jj}$. When $j \notin \mathcal{S}_{\boldsymbol{\beta}}$, a standard deviation inequality for the χ^2 distribution is applicable. For $j \in \mathcal{S}_{\boldsymbol{\beta}}$, we need to pay extra effort to obtain the appropriate deviation inequality, which is also the main technical contribution of this paper. Once we have established these deviation inequalities, we can show the following theorem.

Theorem 1. *Suppose $s = O(p^{1-\delta})$ for some $\delta > 0$. There exists a positive constant Ω such that, for any*

$$\Gamma = \frac{n}{s\log(p-s)} \geq \Omega, \tag{10}$$

the support S is recovered by the DT algorithm (i.e., $\widehat{S} = S$) with probability converging to 1 as n increases. Additionally, the number of slices can be held large enough but fixed (depending solely on C, l, K, M, κ, C_V).

Remark 1 (Choice of H)**.** An interesting by-product of our analysis is that we can choose the number of slices H large enough but finite. When dimension p is fixed, this has already been observed in the literature (Li 1991), however in high dimensional setting to the best of our knowledge this property has not been discussed.

Clearly the DT algorithm does not recover the signed support of $\boldsymbol{\beta}$ standalone. One can apply DT first to select the variables and then apply the SIR procedure to obtain an estimate of the principal eigenvector. To obtain the signed support of $\boldsymbol{\beta}$, take the sign of the principal eigenvector. We summarize this algorithm in the following:

Algorithm 2: DT-SIR

input: $(Y_i, \boldsymbol{X}_i)_{i=1}^n$: data, H: number of slices, s: the sparsity of $\boldsymbol{\beta}$

1. Perform Algorithm 1 to obtain the set $\{j : \mathbf{V}^{jj} \in \widehat{S}\}$.
2. Evaluate $\mathbf{v}$ — the principal eigenvector of the matrix $\mathbf{V}_{\widehat{S},\widehat{S}}$.
3. Output sign($\mathbf{v}$).

We note that this algorithm recovers the signed support, up to multiplication by ± 1. The following result is a direct corollary of Theorem 1, whose proof is omitted.

Proposition 3. *Under the assumptions of Theorem 1, with a potentially bigger value of* Ω*, applying Algorithm 2 restores the signed support (up to a sign) with asymptotic probability converging to* 1*. In fact Algorithm 2 works when* $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \mathbb{I}_p)$ *with* $\mathbf{V}$ *substituted with* $\widehat{\mathbf{V}}$ *where the* jk^{th} *entry of* $\widehat{\mathbf{V}}$ *is defined as* $\widehat{\mathbf{V}}^{jk} = \frac{1}{H}\sum_{h=1}^{H}(\overline{\boldsymbol{X}}_h^j - \overline{\boldsymbol{X}}^j)(\overline{\boldsymbol{X}}_h^k - \overline{\boldsymbol{X}}^k)$.

2.2.3 Optimality of SDP

Algorithm 2 is a two-step procedure, selecting variables by applying DT and then obtaining the SIR estimate of $\boldsymbol{\beta}$. Recent advances of optimization theory provide us with a more sophisticated one-step approach to obtain a sparse principal eigenvector. It is well known that the principal eigenvector of a symmetric positive definite matrix $\mathbf{A}$ is given by:

$$\widehat{\mathbf{z}} = \underset{\mathbf{z}\in\mathbb{R}:\|\mathbf{z}\|_2=1}{\operatorname{argmax}} \ \mathbf{z}^\intercal \mathbf{A}\mathbf{z}$$

and the principal eigenvalue is $\widehat{\mathbf{z}}^\intercal\mathbf{A}\mathbf{z}$. When $\mathbf{z}$ is sparse, i.e., $\|\mathbf{z}\|_0 \leq s$, the above optimization problem with an additional sparsity constraint is computationally expensive. To remedy this difficulty, d'Aspremont et al. (2007) proposed an SDP approach to solve the sparse optimization problem. They suggested to solve the following convex program:

$$\widehat{\mathbf{Z}} = \underset{\operatorname{tr}(\mathbf{Z})=1,\mathbf{Z}\in\mathbb{S}_+^p}{\operatorname{argmax}} \operatorname{tr}(\mathbf{A}\mathbf{Z}) - \lambda_n \sum_{i,j=1}^{p} |\mathbf{Z}_{ij}|, \tag{11}$$

where $\mathbb{S}_+^p$ is the set of all the $p \times p$ positive semi-definite matrices. If the solution happens to be a rank 1 solution, then it is of the form $\widehat{\mathbf{Z}} = \widehat{\mathbf{z}}\widehat{\mathbf{z}}^\intercal$, and hence we can easily obtain an estimate of the principal eigenvector.

In the following algorithm we summarize the SDP approach, tailored for the signed support recovery of SIM. We remind the reader, that this algorithm recovers the signed support, up to multiplication by a global constant equal to ± 1.

Algorithm 3: SDP algorithm for SIR

input: $(Y_i, \boldsymbol{X}_i)_{i=1}^n$: data, H: number of slices, s: the sparsity of $\boldsymbol{\beta}$

1. Calcluate the matrix $\mathbf{V}$ — as given in (3);
2. Obtain the matrix $\widehat{\mathbf{Z}}$ by solving (11), with $\mathbf{A} = \mathbf{V}$;
3. Find the principal eigenvector $\widehat{\mathbf{z}}$ of $\widehat{\mathbf{Z}}$;
4. Output $\operatorname{sign}(\widehat{\mathbf{z}})$.

The performance of this algorithm is guaranteed by the following theorem.

Theorem 2. *Suppose $s = O(p^{1-\delta})$ for some $\delta > 0$, then there is a positive constant Ω such that, for any*

$$\Gamma = \frac{n}{s \log(p-s)} \geq \Omega, \tag{12}$$

with a properly chosen the tuning parameter λ_n , Algorithm 3 recovers the signed support with probability converging to 1, i.e. $\mathbb{P}(\operatorname{sign}(\widehat{\mathbf{z}}) = \operatorname{sign}(\boldsymbol{\beta})) \to 1$[3].

Remark 2. In fact Algorithm 3 works when $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \mathbb{I}_p)$ with $\mathbf{V}$ substituted with $\widehat{\mathbf{V}}$ where the jk^{th} entry of $\widehat{\mathbf{V}}$ is defined as $\widehat{\mathbf{V}}^{jk} = \frac{1}{H}\sum_{h=1}^H (\overline{\boldsymbol{X}}_h^j - \overline{\boldsymbol{X}}^j)(\overline{\boldsymbol{X}}_h^k - \overline{\boldsymbol{X}}^k)$. The proof of this fact is trivial and is omitted.

3 Numerical Experiments and Data Analysis

We open this section with extensive numerical studies and in Section 3.2 we apply our algorithms to a real dataset.

3.1 Simulations

In this section we compare Algorithms 2 and 3 in terms of signed support recovery. We consider the following scenarios:

$$Y = \boldsymbol{X}^\intercal \boldsymbol{\beta} + \sin(\boldsymbol{X}^\intercal \boldsymbol{\beta}) + N(0,1), \tag{13}$$
$$Y = 2\operatorname{atan}(\boldsymbol{X}^\intercal \boldsymbol{\beta}) + N(0,1), \tag{14}$$
$$Y = (\boldsymbol{X}^\intercal \boldsymbol{\beta})^3 + N(0,1), \tag{15}$$
$$Y = \sinh(\boldsymbol{X}^\intercal \boldsymbol{\beta}) + N(0,1). \tag{16}$$

In each simulation we create a sample of size n with dimensionality p and sparsity levels $s = \sqrt{p}$ ($s = \log(p)$ *resp.*) and we vary the rescaled sample size $\Gamma \in [0, 30]$ for the DT-SIR and $\Gamma \in [0, 40]$

[3] We understand that $\widehat{\mathbf{z}}$ is selected so that $\widehat{\mathbf{z}}^\intercal \boldsymbol{\beta} \geq 0$ by convention.

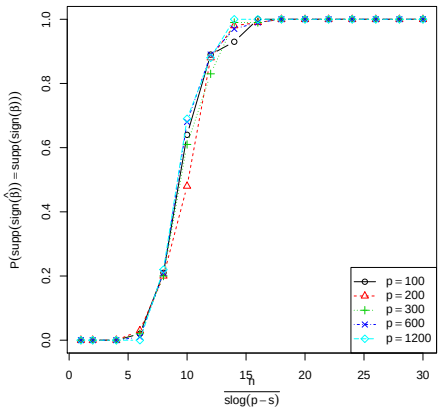
for the SDP approach. The vector $\boldsymbol{\beta}$ is selected in the following manner:

$$\beta_j = \frac{1}{\sqrt{s}},\ j \in [s-1], \quad \beta_s = -\frac{1}{\sqrt{s}},\ \text{and } \beta_j = 0,\ \text{for } j > s.$$

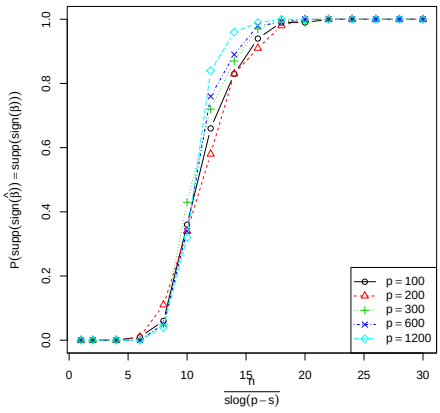
Each simulation is repeated 500 times, and we report the proportion of correctly recovered signed supports of the vector $\boldsymbol{\beta}$. We remark that in general, it should not be expected that the phase transition described in Theorems 1 and 2 to occur at the same places for these 4 models.

We first explore the predictions of Proposition 3 and Algorithm 2. Even though we provide theoretical values of the constants H and m, we ran all simulations with $H = 10$ slices. We believe this scenario, is still reflective of the true nature of the DT-SIR algorithm, as the theoretical value of H we provide is not optimized in any fashion. In Figure 2, we present DT-SIR results from plots for different p values in the regime $s = \sqrt{p}$.

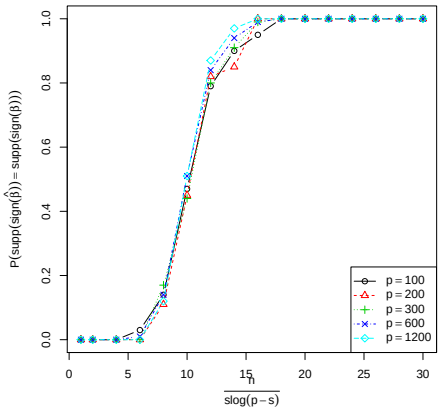
Figure 2: Efficiency Curves for DT-SIR, $s = \sqrt{p},\ \Gamma = \frac{n}{s\log(p-s)} \in [0, 30]$



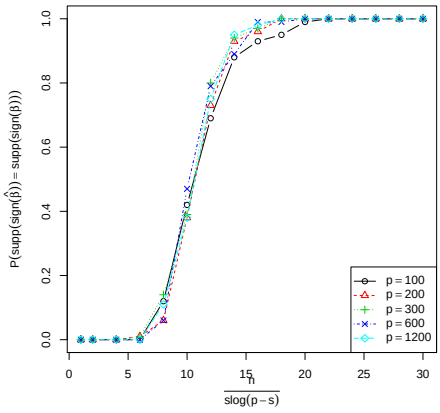
(a) Model (13)



(b) Model (14)



(c) Model (15)

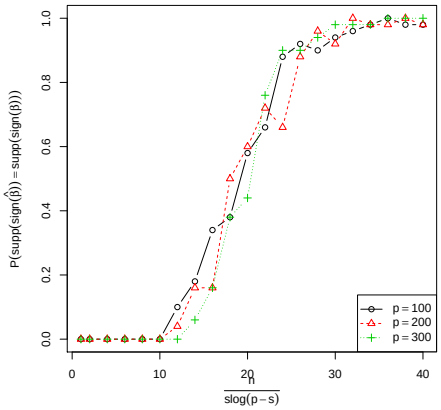


(d) Model (16)

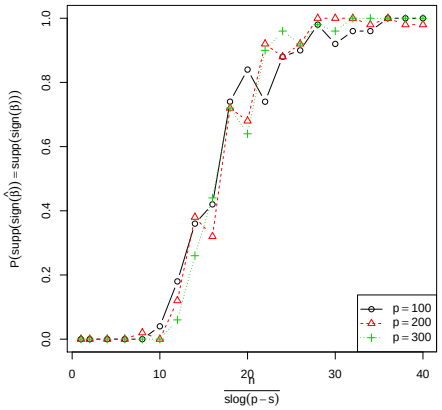
On the X-axis we have plotted the rescaled sample size $\frac{n}{s\log(p-s)}$ and on the Y-axis is the estimated probability of successful signed support recovery. We would refer to these curves as efficiency curves (EC). The EC plots in the case $s = \sqrt{p}$, are very similar to the corresponding plots in the regime $s = \log(p)$, which can be seen in Figure 4 in Appendix C. Both EC plots are in concordance with the predictions from our theoretical results. We can distinctly see the phase transition occurring in approximately the same place regardless of the values of the dimension p.

Next, we present the corresponding ECs for Algorithm 3. We used the code from an efficient implementation of program (11), as suggested in Zhang & Ghaoui (2011). The code was kindly provided to us by the authors of Zhang & Ghaoui (2011). In Figure 3 we plot the ECs for the four models in the case when $s = \sqrt{p}$. Due to running time limitations we only show scenarios where $p \in \{100, 200, 300\}$.

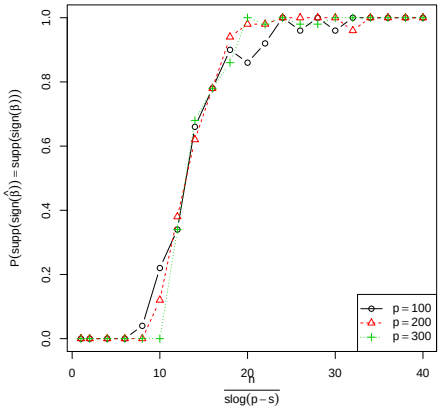
Figure 3: Efficiency Curves for SDP, $s = \sqrt{p}$, $\Gamma = \frac{n}{s\log(p-s)} \in [0, 40]$



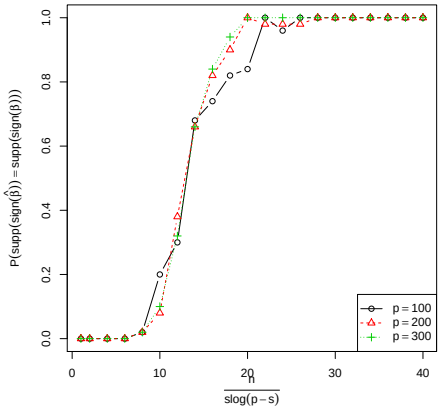
(a) Model (13)



(b) Model (14)



(c) Model (15)



(d) Model (16)

Here we have again used $H = 10$ in all scenarios, for simplicity. We observe that phase transitions are occurring in all of the models, and the signed support is being correctly recovered for large enough values of $\Gamma = \frac{n}{s\log(p-s)}$. Plots for the setting $s = \log(p)$ are provided in Figure 5. An important empirical observation is that the constant Ω needs to be higher when running the SDP algorithm compared to when running the DT-SIR. This fact coupled with the much slower run-times of the SDP algorithm speak in favor of using the DT-SIR algorithm in practice.

In addition to the above simulations, we also performed numerical studies over the same set of models (13), (14), (15) and (16), where the coefficients of the vector $\boldsymbol{\beta}$ were not of equal magnitude. In our second set of simulations we chose $\boldsymbol{\beta} = \frac{\widetilde{\boldsymbol{\beta}}}{\|\widetilde{\boldsymbol{\beta}}\|_2}$, where the entries of $\widetilde{\boldsymbol{\beta}}$ were drawn in each simulation as

$$\tilde{\beta}_j = U_j \text{ for } j \leq \lfloor s/2 \rfloor, \quad \tilde{\beta}_j = -U_j \text{ for } s/2 < j \leq s, \text{ and } \tilde{\beta}_j = 0 \text{ otherwise,} \tag{17}$$

where U_j for $j \in [s]$ are i.i.d. uniform $U(\frac{1}{2}, 1)$ random variables. The results are similar to the ones reported above, although the efficiency curves required somewhat higher values of the rescaled sample size Γ, and extra variability can be seen in the curves due to the random choice of $\boldsymbol{\beta}$. This is to be expected, since under such a generation scheme, the lowest signal will be of lower magnitude compared to the uniform signal case. We show the results for the case $s = \sqrt{p}$ in Appendix C under Figures 6 and 7. Due to space considerations, we do not show results for the case $s = \log p$, as the results are comparable to the results plotted in Figures 4 and 5.

3.2 Real Data Example

In this sub-section we will illustrate the practical benefits of our algorithm on a real dataset. Our dataset is the mouse embryonic stem cells (ESCs) dataset, and has been previously analyzed by Zhong et al. (2012), Jiang et al. (2014).

The outcome variable is the expression levels of $12,408$ genes, and is derived using RNA-seq technology in mouse ESCs (Cloonan et al. 2008). The predictors correspond to 312 transcription factors (TFs). From the predictors, 12 TFs are known to be associated with different roles in the ES-cell biology as constituents of important signaling pathways, self renewal regulators, and key reprogramming factors, and the remaining 300 are putative mouse TFs compiled from the TRANSFAC database. For each gene and TF, the predictor matrix contains a score which is the *transcription factor association strength* score (Chen et al. 2008, Ouyang et al. 2009) for the 12 known TFs, and motif matching scores for the remaining 300 genes which were introduced by Zhong et al. (2012). Hence, the design matrix $\mathbb{X} = (\boldsymbol{X}_i^\intercal)_{i\in[n]}^\intercal$ is a $n \times p$ matrix, where $n = 12,408$ and $p = 312$, each entry of which represents the score of a TF for the corresponding gene. The EFCs dataset is not truly a high-dimensional dataset in the sense $n < p$, but serves to illustrate that our algorithms can successfully perform variable selection. Since $\mathbb{X}$ is coming from a real dataset with non-homogeneous scores across the predictors, we wouldn't expect it to be centered with identity covariance. To deal with this problem we use the statistics $\widetilde{\mathbf{V}}_{jk}$ in place of $\mathbf{V}_{jk}$ where the matrix $\widetilde{\mathbf{V}}$ is defined as:

$$\widetilde{\mathbf{V}} := \widehat{\boldsymbol{\Sigma}}^{-1/2} \frac{1}{H} \sum_{h=1}^{H} (\overline{\boldsymbol{X}}_h - \overline{\boldsymbol{X}})(\overline{\boldsymbol{X}}_h - \overline{\boldsymbol{X}})^\intercal \widehat{\boldsymbol{\Sigma}}^{-1/2},$$

and $\widehat{\boldsymbol{\Sigma}}^{-1/2} := \left[n^{-1} \sum_{i\in[n]} (\boldsymbol{X}_i - \overline{\boldsymbol{X}})(\boldsymbol{X}_i - \overline{\boldsymbol{X}})^\intercal \right]^{-1/2}$ is the symmetric square root of the covariance

matrix of $\mathbb{X}$. Notice that in this dataset $\widehat{\mathbf{\Sigma}}^{-1/2}$ can be estimated without the application of sparse inverse procedures since $n > p$. As in the simulation section we used $H = 10$ slices.

We ran the DT and SDP procedures (i.e. Algorithms 1 and 3) on the ESCs dataset, and the DT procedure selected 28 TFs which it found associated with the outcome, while the SDP procedure selected 36. In Table 1 we compare the rankings of the 12 known ES-cell TFs within the TFs selected by DT and SDP algorithms to the same rankings produced by the procedures SIRI-AE (Jiang et al. 2014) and COP (Zhong et al. 2012).

Table 1: Rankings of the 12 known TFs among the selected TFs by different algorithms

TFs names	DT	SDP	SIRI-AE	COP
E2f1	1	1	1	1
Zfx	2	3	3	3
Mycn	3	2	4	10
Klf4	7	4	5	19
Myc	4	7	6	–
Esrrb	14	10	8	–
Oct4	20	20	9	11
Tcfcp2l1	8	8	10	36
Nanog	11	17	14	–
Stat3	–	23	17	20
Sox2	24	–	18	–
Smad1	25	25	32	13

The top 10 highly ranked genes of SIRI-AE, DT, SDP and COP contain $8, 6, 7, 3$ known ES-cell TFs respectively. The DT algorithm is the most parsimonious out of all procedures (SIRI-AE and COP have reportedly selected 34 and 42 TFs correspondingly), but nevertheless includes all of the known TFs with the exception of the Stat3. The DT algorithm is additionally able to capture both Nanog and Sox2 which are generally believed to be the master ESC regulators. These TFs are completely missed by COP and the Sox2 TF was omitted by SDP. The SIRI-AE algorithm is the only algorithm which is able to recognize all of the 12 known TFs. We would like to mention that compared to our work the SIRI-AE algorithm is designed to work in a lower dimensional setup, and is able to capture interactions beyond the single-index models, which could help explain its better performance on the ESCs dataset.

All in all, we believe the results of our algorithms are satisfactory, and illustrate that both DT and SDP can be successfully applied in practice. As a remark, in a truly high-dimensional setting when $n < p$, instead of the estimate $\widehat{\mathbf{\Sigma}}^{-1/2}$, in practice one might resort to sparse estimators produced by procedures such as CLIME (Cai et al. 2011), and use the corrected statistics $\widetilde{\mathbf{V}}$ in place of $\mathbf{V}$.

4 Proof of Theorem 1

We will prove this theorem under slightly more general conditions:

i. Let $\boldsymbol{X}^j$, $j \in [p]$ be centered, sub-Gaussian random variables with $\max_{j\in[p]} \|\boldsymbol{X}^j\|_{\psi_2} \leq \mathcal{K}$.

ii. For $j \in \mathcal{S}_{\boldsymbol{\beta}}$, we assume that $\text{Var}[\mathfrak{m}_j(Y)] \geq \frac{C_V}{s}$.

iii. For $j \in \mathcal{S}_{\boldsymbol{\beta}}^c$, we assume that $\mathfrak{m}_j(Y) = \mathbb{E}[\boldsymbol{X}^j|Y] = 0$ a.s.

It is easy to see for that for SIM in $\mathcal{F}_A$, the above conditions are satisfied in the case when $\boldsymbol{X} \sim N_p(0, \mathbb{I}_p)$. We start with a high level outline of the proof. As we pointed out, the key argument is to show a deviation inequality for $\mathbf{V}^{jj}$, i.e. we will show that $|\mathbf{V}^{jj} - \text{Var}[\mathfrak{m}_j(Y)]|$ is small (e.g. $< \frac{1}{3}\text{Var}[\mathfrak{m}_j(Y)]$) with high probability provided that $\Gamma = \frac{n}{s\log(p-s)}$ is large enough. Note that:

$$|\mathbf{V}^{jj} - \text{Var}[\mathfrak{m}_j(Y)]| = \left| \frac{1}{H}\sum_{h=1}^{H} \left(\overline{\boldsymbol{X}}_h^j\right)^2 - \int \mathfrak{m}_j^2(y) p_Y(y) dy \right|. \tag{18}$$

We will show below (see (22)), that the above expression is well approximated by:

$$\left| \frac{1}{H}\sum_{h=1}^{H} \left(\overline{\boldsymbol{X}}_h^j\right)^2 - \sum_{h=1}^{H} (\boldsymbol{\mu}_h^j)^2 \mathbb{P}(Y \in S_h) \right|, \tag{19}$$

under sliced stability (where $Y_{(0)} = -\infty$). Intuitively, we have 1) $\mathbb{P}(Y \in S_h) \approx \frac{1}{H}$ and 2) $\overline{\boldsymbol{X}}_h^j \approx \boldsymbol{\mu}_h^j$, which shows that (19) should be close to 0. To rigorously validate this intuition, we use the Dvoretzky-Kiefer-Wolfowitz (DKW) inequality (Massart 1990) for 1) and concentration inequalities for 2).

Note that the probability $\mathbb{P}(Y \in S_h)$ is a random variable, where the randomness comes from the two endpoints — $Y_{(m(h-1))}$ and $Y_{(mh)}$ of the interval S_h. Recall that F_n is the empirical distribution function of Y, based on the sample Y_i. By the DKW inequality, we have that $\mathbb{P}(\sup_y |F_n(y) - F(y)| > \epsilon) \leq 2\exp(-2n\epsilon^2)$, which in addition to the fact that Y has a continuous distribution, implies that for all h we have:

$$\frac{1}{H} - 2\epsilon \leq \mathbb{P}\left(\frac{h-1}{H} < F_n(Y) < \frac{h}{H}\right) \leq \mathbb{P}\left(\frac{h-1}{H} \leq F_n(Y) \leq \frac{h}{H}\right) \leq \frac{1}{H} + 2\epsilon, \tag{20}$$

on an event with probability at least $1 - 2\exp(-2n\epsilon^2)$. Denote the event where (20) holds with E.

Next, notice that for $h \in [H-2]$ and a random permutation $\pi : [m-1] \mapsto [m-1]$, the collection random variables $\boldsymbol{X}_{h,\pi_i}^j, i \in [m-1]$ are conditionally i.i.d. given $(Y_{(m(h-1))}, Y_{(mh)})$, and for $h = H-1$ and a random permutation $\pi : [m] \mapsto [m]$ the random variables $\boldsymbol{X}_{h,\pi_i}^j, i \in [m]$ are conditionally i.i.d. given $Y_{(m(H-1))}$, as they can be generated via simple rejection sampling. Hence conditionally on $(Y_{(mh)})_{h\in[H-1]}$ the sample means $\overline{\boldsymbol{X}}_{h,1:(m-1)}^j$ have corresponding means $\boldsymbol{\mu}_h^j$ when $h \in [H-1]$ and $\overline{\boldsymbol{X}}_H^j$ has a mean of $\boldsymbol{\mu}_H^j$. For consistency of notation we will discard a point at random in the H^{th} slice, and WLOG (upon re-labeling the points) we assume that the discarded point is $\boldsymbol{X}_{H,m}$. Next, we formulate the following concentration result for the sliced means, which we show in Appendix A:

Lemma 2. *Let $G \subseteq [p]$. On the event E, for $\eta > 0$ we have the following:*

$$\mathbb{P}\left(\max_{j\in G, h\in[H]} \left|\overline{\boldsymbol{X}}_{h,1:(m-1)}^j - \boldsymbol{\mu}_h^j\right| > \eta\right) \leq 2|G|H\exp\left(-\frac{\eta^2(m-1)}{C_1\mathcal{K}^2 q^{-1} + C_2\mathcal{K}\eta}\right), \tag{21}$$

where $q = \frac{1}{H} - 2\epsilon$, for some absolute constants $C_1, C_2 > 0$.

Set $G = [p]$ and denote with $\widetilde{E}$ the event on which we have

$$\max_{j\in[p],h\in[H]} \left|\overline{\boldsymbol{X}}^j_{h,1:(m-1)} - \boldsymbol{\mu}^j_h\right| \leq \eta.$$

By (21), (20) and the union bound we have that:

$$\mathbb{P}(\widetilde{E}) \geq 1 - 2pH\exp\left(-\frac{\eta^2(m-1)}{C_1\mathcal{K}^2q^{-1} + C_2\mathcal{K}\eta}\right) - 2\exp(-2n\epsilon^2).$$

Next we move on, to show that (18) is close to (19) on the event E, as well as we collect two straightforward inequalities in the following helpful:

Lemma 3. *Assume that the sliced stability condition (4) holds. Then we have the following inequalities holding on the event E, for large enough H, and small enough ϵ:*

$$|\mathbf{V}^{jj} - \mathrm{Var}[\mathfrak{m}_j(Y)]| \leq \left|\frac{1}{H}\sum_{h=1}^{H}\left(\overline{\boldsymbol{X}}^j_h\right)^2 - \sum_{h=1}^{H}(\boldsymbol{\mu}^j_h)^2\mathbb{P}(Y\in S_h)\right| + \underbrace{\frac{CH^\kappa}{s}\left(\frac{1}{H}+2\epsilon\right)}_{B_1}, \tag{22}$$

$$\sum_{h=1}^{H}(\boldsymbol{\mu}^j_h)^2 \leq \underbrace{\frac{\frac{C_V}{s}+B_1}{\left(\frac{1}{H}-2\epsilon\right)}}_{B_2}, \tag{23}$$

$$\sum_{h=1}^{H}|\boldsymbol{\mu}^j_h| \leq \underbrace{\frac{\sqrt{\frac{C_V}{s}+B_1}}{\left(\frac{1}{H}-2\epsilon\right)}}_{B_3}. \tag{24}$$

Note. *We refer to the constants from (4) as C and κ, dropping the dependence on K and M for brevity, and in fact $C = C(l,K,M)C_V$.*

Note that by an elementary calculation — using (20) and Lemma 3, on the event $\widetilde{E}$ we get:

$$|\mathbf{V}^{jj} - \mathrm{Var}[\mathfrak{m}_j(Y)]| \leq \frac{1}{H}\sum_{h=1}^{H}\left|\left(\overline{\boldsymbol{X}}^j_h\right)^2 - \frac{(m-1)^2}{m^2}(\boldsymbol{\mu}^j_h)^2\right| \tag{25}$$

$$+ \underbrace{\left(2\epsilon + \frac{1}{H} - \frac{(m-1)^2}{Hm^2}\right)B_2}_{I_1} + \underbrace{B_1}_{I_2}, \tag{26}$$

where we used (22), the triangle inequality and (23). Consider the following:

Lemma 4. *There exists a subset* $\widetilde{\widetilde{E}} \subset \widetilde{E}$ *such that* $\mathbb{P}(\widetilde{E} \setminus \widetilde{\widetilde{E}}) \leq p\exp(-c\frac{\tau^2 n}{4\mathcal{K}^4})$*, for some fixed* $\tau \in [0, 2\mathcal{K}^2)$ *on which we have the following bound for all* $j \in [p]$*:*

$$\frac{1}{H}\sum_{h=1}^{H}\left|\left(\overline{\boldsymbol{X}}_h^j\right)^2 - \frac{(m-1)^2}{m^2}(\boldsymbol{\mu}_h^j)^2\right| \leq \underbrace{\frac{(2\mathcal{K}^2+\tau)}{m}}_{I_3} + \underbrace{\frac{2\sqrt{2\mathcal{K}^2+\tau}}{\sqrt{m}}\sqrt{2\eta^2+2\frac{B_2}{H}}}_{I_4} + \underbrace{\eta^2}_{I_5} + \underbrace{2\eta\frac{B_3}{H}}_{I_6}. \tag{27}$$

We defer the proof of this lemma to the appendix. Next, we provide exact constants, such that $|\mathbf{V}^{jj} - \mathrm{Var}[\mathfrak{m}_j(Y)]| \leq \frac{C_V}{\rho s}$ for some constant $\rho > 0$ and all $j \in [p]$, so that the probability of the event $\widetilde{\widetilde{E}}$ still converges to 1. The remarkable phenomenon here is that the number of slices H, can be selected so that it is a constant, which might seem counterintuitive. Select the constants in the following manner:

$$H = \max\left\{M, \left(\frac{\gamma C K}{C_V}\right)^{\frac{1}{1-\kappa}}\right\}, \tag{28}$$

$$\epsilon = \min\left\{\frac{K-1}{2H}, \frac{1-l}{2H}, \frac{l}{4(\gamma+1)H}\right\}, \tag{29}$$

$$\eta = \frac{l\sqrt{C_V}}{2\sqrt{\gamma(\gamma+1)}\sqrt{s}}, \tag{30}$$

$$m \geq \max\left\{\frac{8\gamma(\gamma+1+l)(2\mathcal{K}^2+\tau)s}{lC_V}, \frac{4(\gamma+1)}{l}, \frac{2s\log(p-s)}{\Upsilon}+1, \frac{8\mathcal{K}^4\log(p-s)}{\tau^2 cH}\right\} \tag{31}$$

where $\gamma \geq 12$ is a positive constant and $\Upsilon = \frac{l^2 C_V}{4(\gamma(\gamma+1))\left(\mathcal{K}^2 C_1 H l^{-1} + \frac{C_2 \mathcal{K} l\sqrt{C_V}}{2\sqrt{\gamma(\gamma+1)}}\right)}$. Simple algebra shows that selecting these constants ensures the following inequality:

$$\max(I_1, I_2, I_3, I_4, I_5, I_6) \leq \frac{C_V}{\gamma s}.$$

By combining (25) and (27) we arrive at:

$$|\mathbf{V}^{jj} - \mathrm{Var}[\mathfrak{m}_j(Y)]| \leq \frac{6C_V}{\gamma s}, \tag{32}$$

as promised for $\rho = \frac{\gamma}{6}$. Next we proceed to show that the probability of the event $\widetilde{\widetilde{E}}$ converges to 0. To achieve this, note that in (29) we chose ϵ to be a constant, so evidently $\exp(-n\epsilon^2) \to 0$. Moreover $p\exp(-c\frac{\tau^2 mH}{4\mathcal{K}^4}) \leq \frac{p}{(p-s)^2} \to 0$ by (31). Finally to show that $\mathbb{P}(\widetilde{\widetilde{E}}) \to 1$, we need:

$$\frac{\eta^2(m-1)}{C_1(H^{-1}-2\epsilon)^{-1}\mathcal{K}^2 + C_2\mathcal{K}\eta} - \log(p) - \log(H) \to +\infty.$$

Noting that $H^{-1} - 2\epsilon \geq lH^{-1}$, $\eta \leq \frac{l\sqrt{C_V}}{2\sqrt{\gamma(\gamma+1)}}$ and that by (28) H is fixed, the above expression is implied by:

$$\Upsilon \frac{(m-1)}{s} - \log(p) \geq 2\log(p-s) - \log(p) \to +\infty,$$

where we used (31) and the fact that $2\log(p-s) \geq \log(p)$ asymptotically (since $s = O(p^{1-\delta})$ for $\delta > 0$). Hence we have guaranteed that $\mathbb{P}(\widetilde{\widetilde{E}}) \to 1$.

Now, note that for variables $j \notin \mathcal{S}_{\boldsymbol{\beta}}$ we have $\mathfrak{m}_j(Y) = \mathbb{E}[\boldsymbol{X}^j] = 0$, which implies that for all $h \in [H], j \in \mathcal{S}_{\boldsymbol{\beta}}^c$ we have $\boldsymbol{\mu}_h^j = 0$. Using (32) and selecting $\gamma = 18$, we conclude that:

$$\mathbf{V}^{jj} \geq \frac{2}{3}\frac{C_V}{s}, j \in \mathcal{S}_{\boldsymbol{\beta}} \text{ and } \mathbf{V}^{jj} \leq \frac{C_V}{3s}, j \in \mathcal{S}_{\boldsymbol{\beta}}^c,$$

and hence separation of the signals is asymptotically possible, as we claimed.

5 Proof of Theorem 2

In this section we show that the SDP relaxation will have a rank 1 solution with high probability and moreover this solution will recover the signed support of the vector $\boldsymbol{\beta}$. In contrast to Section 4, here we make full usage of the fact that $\boldsymbol{X} \sim N_p(0, \mathbb{I}_p)$. One simple implication of this, is for example the fact that $\max_j \|\boldsymbol{X}^j\|_{\psi_2} < 1$, and hence we can set $\mathcal{K} = 1$. For the analysis of the algorithm we set the regularization parameter $\lambda_n = \frac{C_V}{2s}$.

To this end, we restate Lemma 5 from Amini & Wainwright (2008), which provides a sufficient condition for a global solution of the SDP problem:

Lemma 5. *Suppose there exists a matrix* $\mathbf{U}$ *satisfying:*

$$\mathbf{U}_{ij} = \begin{cases} \operatorname{sign}(\widehat{\mathbf{z}}_i)\operatorname{sign}(\widehat{\mathbf{z}}_j), & \text{if } \widehat{\mathbf{z}}_i\widehat{\mathbf{z}}_j \neq 0; \\ \in [-1,1], & \text{otherwise.} \end{cases} \tag{33}$$

Then if $\widehat{\mathbf{z}}$ *is the principal eigenvector of the matrix* $\mathbf{A} - \lambda_n\mathbf{U}$, $\widehat{\mathbf{z}}\widehat{\mathbf{z}}^\intercal$ *is the optimal solution to problem (11).*

Recall that the SIR estimate of the variance-covariance matrix has entries:

$$\mathbf{V}^{jk} = \frac{1}{H}\sum_{h=1}^{H} \overline{\boldsymbol{X}}_h^j \overline{\boldsymbol{X}}_h^k.$$

Denote with $\widetilde{\mathbf{V}} = \mathbf{V} - \lambda_n\mathbf{U}$, where $\mathbf{U}$ is a to be defined sign matrix from Lemma 5. We furthermore consider the decomposition of $\widetilde{\mathbf{V}}$ into blocks — $\widetilde{\mathbf{V}}_{\mathcal{S}_{\boldsymbol{\beta}},\mathcal{S}_{\boldsymbol{\beta}}}$, $\widetilde{\mathbf{V}}_{\mathcal{S}_{\boldsymbol{\beta}}^c,\mathcal{S}_{\boldsymbol{\beta}}}$, $\widetilde{\mathbf{V}}_{\mathcal{S}_{\boldsymbol{\beta}}^c,\mathcal{S}_{\boldsymbol{\beta}}^c}$. Here, these three matrices are sub-matrices of the matrix $\widetilde{\mathbf{V}}$ restricted to entries with indexes in the sets $\mathcal{S}_{\boldsymbol{\beta}}$ or $\mathcal{S}_{\boldsymbol{\beta}}^c$ correspondingly. We observe that $\mathbf{U}_{\mathcal{S}_{\boldsymbol{\beta}},\mathcal{S}_{\boldsymbol{\beta}}} = \operatorname{sign}(\boldsymbol{\beta}_{\mathcal{S}_{\boldsymbol{\beta}}})\operatorname{sign}(\boldsymbol{\beta}_{\mathcal{S}_{\boldsymbol{\beta}}})^\intercal$.

We first focus on the $\mathbf{V}_{\mathcal{S}_{\boldsymbol{\beta}},\mathcal{S}_{\boldsymbol{\beta}}}$ matrix. We calculate the value of the covariance of two coordinates $j, k \in \mathcal{S}_{\boldsymbol{\beta}}$:

$$\begin{aligned} \operatorname{Cov}[\mathfrak{m}_j(Y), \mathfrak{m}_k(Y)] &= \mathbb{E}[\mathfrak{m}_j(Y)\mathfrak{m}_k(Y)] = \operatorname{sign}(\boldsymbol{\beta}_j)\operatorname{sign}(\boldsymbol{\beta}_k)\mathbb{E}[\mathfrak{m}_k^2(Y)] \\ &= \boldsymbol{\beta}_j\boldsymbol{\beta}_k C_V, \end{aligned} \tag{34}$$

where we used that $\text{sign}(\boldsymbol{\beta}_j)\mathfrak{m}_j(Y) = \text{sign}(\boldsymbol{\beta}_k)\mathfrak{m}_k(Y)$, which follows by noticing that the distribution of $\boldsymbol{X}^j|Y$ is the same as the distribution of $\boldsymbol{X}^k|Y$ except the potential difference in the signs of the coefficients, because of the symmetry in the problem.

Next, let $G = \mathcal{S}_{\boldsymbol{\beta}}$ (hence $|G| = s$) in Lemma 2 to obtain that:

$$\max_{j\in\mathcal{S}_{\boldsymbol{\beta}},h\in[H]} \left|\overline{\boldsymbol{X}}^j_{h,1:(m-1)} - \boldsymbol{\mu}^j_h\right| \leq \eta, \tag{35}$$

with probability at least $1 - 2sH\exp\left(-\frac{\eta^2(m-1)}{C_1q^{-1}+C_2\eta}\right) - 2\exp(-2n\epsilon^2)$. Let $\widetilde{E}_{\mathcal{S}_{\boldsymbol{\beta}}} \subset E$, be the event where (35) holds. We proceed with formulating a bound similar to Lemma 4, but for the covariance:

Lemma 6. *There exists an event $\widetilde{\widetilde{E}}_{\mathcal{S}_{\boldsymbol{\beta}}}$ with $\mathbb{P}(\widetilde{E}_{\mathcal{S}_{\boldsymbol{\beta}}} \setminus \widetilde{\widetilde{E}}_{\mathcal{S}_{\boldsymbol{\beta}}}) \leq s\exp(-c\frac{\tau^2 n}{4})$ ($c > 0$ is an absolute constant), for some fixed $\tau \in (0, 2]$, such that for all $j, k \in \mathcal{S}_{\boldsymbol{\beta}}$, we have the following inequality:*

$$\begin{aligned}\left|\mathbf{V}^{jk} - \text{Cov}(\mathfrak{m}_j(Y), \mathfrak{m}_k(Y))\right| \leq &\left(2\epsilon + \frac{1}{H} - \frac{(m-1)^2}{Hm^2}\right)B_2 + B_1 + 4\eta\frac{B_3}{H} \\ &+ \frac{4(2+\tau)}{m} + \frac{4\sqrt{2+\tau}}{\sqrt{m}}\sqrt{2\eta^2 + 2\frac{B_2}{H}} + 4\eta^2.\end{aligned} \tag{36}$$

Let H', ϵ', η' are constants selected according to (28) — (30), with $\mathcal{K} = 1$, correspondingly. Set $H = H', \epsilon = \epsilon', \eta = \frac{\eta'}{2}$, and take

$$m \geq \max\left\{32\frac{8\gamma(\gamma+1+l)(2+\tau)s}{lC_V}, \frac{8s\log(p-s)}{\Upsilon} + 1, \frac{8\log(p-s)}{\tau^2 cH}\right\}, \tag{37}$$

where Υ is specified as before with $\mathcal{K} = 1$. Similarly to Section 4, we can show that the following inequality:

$$\sup_{j,k\in\mathcal{S}_{\boldsymbol{\beta}}} \left|\mathbf{V}^{jk} - \text{Cov}(\mathfrak{m}_j(Y), \mathfrak{m}_k(Y))\right| \leq \frac{6C_V}{\gamma s}, \tag{38}$$

holds on $\widetilde{\widetilde{E}}_{\mathcal{S}_{\boldsymbol{\beta}}}$, with the probability of $\widetilde{\widetilde{E}}_{\mathcal{S}_{\boldsymbol{\beta}}}$ tending to 1. To get the bound in (38), one can observe that with the choices of constants as above all 6 terms in (36) are guaranteed to be smaller than $\frac{C_V}{\gamma s}$. Here similarly to Section 4, H is large enough but fixed. Having in mind the above inequality we consider the matrix $\widetilde{\mathbf{V}}_{\mathcal{S}_{\boldsymbol{\beta}},\mathcal{S}_{\boldsymbol{\beta}}}$:

$$\widetilde{\mathbf{V}}_{\mathcal{S}_{\boldsymbol{\beta}},\mathcal{S}_{\boldsymbol{\beta}}} = \frac{C_V}{2}\boldsymbol{\beta}_{\mathcal{S}_{\boldsymbol{\beta}}}\boldsymbol{\beta}^{\intercal}_{\mathcal{S}_{\boldsymbol{\beta}}} + \mathbf{N},$$

where $\mathbf{N}$ is some symmetric noise matrix. Note that by (34), (38) gives a bound on $\|\mathbf{N}\|_{\max}$. Next we make usage of Lemma 8, which is a restatement of Lemma 6 in Amini & Wainwright (2008) and can be found in the Appendix for the reader's convenience. We start by verifying that condition (53) indeed holds for the matrix $\mathbf{N}$. We have that:

$$\|\mathbf{N}\|_{\max} = \|\mathbf{V}_{\mathcal{S}_{\boldsymbol{\beta}},\mathcal{S}_{\boldsymbol{\beta}}} - C_V\boldsymbol{\beta}_{\mathcal{S}_{\boldsymbol{\beta}}}\boldsymbol{\beta}^{\intercal}_{\mathcal{S}_{\boldsymbol{\beta}}}\|_{\max} \leq \frac{6C_V}{\gamma s}, \tag{39}$$

with the last inequality following from (38). Selecting $\gamma = 240$ bounds $\|\mathbf{N}\|_{\max}$ by $\frac{C_V}{40s}$, as required in (53). Thus, by Lemma 8 we conclude that:

(a) For $\gamma_1 := \lambda_{\max}(\widetilde{\mathbf{V}})$ and the second largest in magnitude eigenvalue of $\widetilde{\mathbf{V}}$ we have $\gamma_1 > |\gamma_2|$.

(b) The corresponding principal eigenvector of $\widetilde{\mathbf{V}}$ — $\widetilde{\mathbf{z}}_{\mathcal{S}_{\boldsymbol{\beta}}}$ satisfies the following inequality:

$$\left\|\widetilde{\mathbf{z}}_{\mathcal{S}_{\boldsymbol{\beta}}} - \boldsymbol{\beta}_{\mathcal{S}_{\boldsymbol{\beta}}}\right\|_\infty \leq \frac{1}{2\sqrt{s}}.$$

Next we show that the rest of the sign matrix $\mathbf{U}$, i.e. $\mathbf{U}_{\mathcal{S}^c_{\boldsymbol{\beta}},\mathcal{S}_{\boldsymbol{\beta}}}$ and $\mathbf{U}_{\mathcal{S}^c_{\boldsymbol{\beta}},\mathcal{S}^c_{\boldsymbol{\beta}}}$ can be selected in such a way, so that the blocks $\widetilde{\mathbf{V}}_{\mathcal{S}^c_{\boldsymbol{\beta}},\mathcal{S}_{\boldsymbol{\beta}}}$ and $\widetilde{\mathbf{V}}_{\mathcal{S}^c_{\boldsymbol{\beta}},\mathcal{S}^c_{\boldsymbol{\beta}}}$ are 0. For this purpose we select $\mathbf{U}_{\mathcal{S}^c_{\boldsymbol{\beta}},\mathcal{S}_{\boldsymbol{\beta}}} = \frac{1}{\lambda_n}\mathbf{V}_{\mathcal{S}^c_{\boldsymbol{\beta}},\mathcal{S}_{\boldsymbol{\beta}}}$ and $\mathbf{U}_{\mathcal{S}^c_{\boldsymbol{\beta}},\mathcal{S}^c_{\boldsymbol{\beta}}} = \frac{1}{\lambda_n}\mathbf{V}_{\mathcal{S}^c_{\boldsymbol{\beta}},\mathcal{S}^c_{\boldsymbol{\beta}}}$. Since it is clear that the vector $(\widetilde{\mathbf{z}}_{\mathcal{S}_{\boldsymbol{\beta}}}, 0_{\mathcal{S}^c_{\boldsymbol{\beta}}})$ is the principal eigenvector of $\widetilde{\mathbf{V}}$, if $\mathbf{U}$ is a sign matrix, Lemma 5 will conclude that — $(\widetilde{\mathbf{z}}^\intercal_{\mathcal{S}_{\boldsymbol{\beta}}}, 0^\intercal_{\mathcal{S}^c_{\boldsymbol{\beta}}})^\intercal(\widetilde{\mathbf{z}}^\intercal_{\mathcal{S}_{\boldsymbol{\beta}}}, 0^\intercal_{\mathcal{S}^c_{\boldsymbol{\beta}}})$ is the optimal solution to the optimization problem, which will in turn conclude our claim.

It remains to show that the specified $\mathbf{U}$ is indeed a sign matrix. Note that by Cauchy-Schwartz for $k \in \mathcal{S}^c_{\boldsymbol{\beta}}$ and any j, we have:

$$\mathbf{V}^{jk} \leq \sqrt{\mathbf{V}^{jj}}\sqrt{\mathbf{V}^{kk}}. \tag{40}$$

From (38) if $j \in \mathcal{S}_{\boldsymbol{\beta}}$, we have that high probability:

$$\mathbf{V}^{jj} \leq \frac{C_V}{s} + \frac{6C_V}{\gamma s} = \frac{(\gamma+6)C_V}{\gamma s}.$$

Hence, it is sufficient to select m, H large enough so that:

$$\mathbf{V}^{kk} \leq \frac{\gamma C_V}{4(\gamma+6)s}.$$

To achieve the above bound, we make usage of the following tail inequality, for χ^2 random variables which we take from Laurent & Massart (2000) (see Lemma 1):

$$\mathbb{P}\left(\frac{\chi^2_H}{H} \geq 1 + 2\sqrt{\frac{x}{H}} + \frac{2x}{H}\right) \leq \exp(-x).$$

Note that, $\mathbf{V}^{kk} \sim \frac{1}{mH}\chi^2_H$ for $k \in \mathcal{S}^c_{\boldsymbol{\beta}}$. Thus applying the bound above we have

$$\frac{1}{mH}\chi^2_H \leq \frac{1}{m} + \frac{2}{m}\sqrt{\frac{x}{H}} + \frac{2x}{mH} \leq \frac{2}{m} + \frac{3x}{mH}, \tag{41}$$

with probability at least $\exp(-x)$. Applying (41) it can be easily seen that by selecting:

$$x = \frac{n\gamma C_V}{24(\gamma+6)s},$$

we can ensure (after using (37)) that $\mathbf{V}^{kk} \leq \frac{\gamma C_V}{4(\gamma+6)s}$ for all $k \in \mathcal{S}^c_{\boldsymbol{\beta}}$. Requiring:

$$\frac{n\gamma C_V}{24(\gamma+6)s} \geq 2\log(p-s), \tag{42}$$

ensures that the probability of the event is asymptotically 1 from the union bound. This combined with (40) shows that the so defined matrix $\mathbf{U}$ is indeed a sign matrix, which concludes the proof.

6 Discussion

Sliced inverse regression has been widely applied in various scientific problems. Though it is a successful tool in terms of data visualization, and provably works when the dimension p is not large, its behavior in the high-dimensional regime $p \gg n$ is much less well understood.

In this paper, we studied the support recovery of SIM within the class of models $\mathcal{F}_A$. We demonstrated that the optimal sample size of this problem is of the order $s \log(p-s)$. Two unforeseeable results of our analysis might be of particular interest for future investigations.

Recall that a central subspace of a pair of random variables $(Y, \boldsymbol{X})$ is the minimal subspace $\mathcal{S}$ such that $Y \perp\!\!\!\perp \boldsymbol{X} \big| P_{\mathcal{S}} \boldsymbol{X}$. The first implication of our results, as we hinted in Remark 1, is that if we focus on estimating the support of the central subspace rather than consistently estimating the intermediate matrix, better convergence rate might be expected. For instance, the results in Lin et al. (2015) show that, under mild conditions, the convergence rate of the SIR estimate $\mathbf{V}$ of $\mathrm{Var}(\mathbb{E}[\boldsymbol{X}|Y])$ is $O_P(\frac{1}{H} + \sqrt{\frac{p}{n}})$. In order to get a consistent estimation of the intermediate matrix $\mathrm{Var}(\mathbb{E}[\boldsymbol{X}|Y])$, the slices number H has to be diverging. The result regarding the choice of H in this paper suggests that it would be possible to get an improved convergence rate of estimating the central subspace which is expected to be $\sqrt{\frac{p}{n}}$.

The second consequence is that estimating the central subspace might be easier than that of sparse PCA in terms of computational cost. At first glance, the unknown nonlinear link function f could bring in difficulties in determining the optimal rate, and it might be reasonable to expect that this difficulty will increase the computational cost in general. The results in this paper however, provide the opposite evidence — as long as $s = O(p^{1-\delta})$ for some $\delta > 0$, the computationally efficient algorithms DT-SIR and SDP approach can solve the support recovery problem. In other words, the tradeoff between computational and statistical efficiency for estimating the central subspace might occur in a more subtle regime, where $s \propto p$ and $n \propto p$. Finally combining our results with the results of Lin et al. (2015), calculating the minimax rate of an estimator for the SDR space in model (1) appears to be a plausible exercise, and is left for future work.

Acknowledgments

This research was conveyed while the first author was a graduate student in the Department of Biostatistics at Harvard University. Lin's research is supported by the Center of Mathematical Sciences and Applications at Harvard University. Liu's research is supported by the NSF Grant DMS-1120368 and NIH Grant R01 GM113242-01. We thank Professor Tianxi Cai, Harvard University and Professor Noureddine El Karoui, University of California, Berkeley for valuable discussions which led to improvements of the present manuscript. In addition, the authors are grateful to Professor Laurent El Ghaoui, University of California, Berkeley and Dr Youwei Zhang for providing the Matlab code used for solving the SDP problem. The authors also express gratitude to the referee and anonymous reviewers for raising important points, which further bettered the manuscript.

A Technical Proofs

Proof of Lemma 1. First take any vector $\mathbf{b} \in \mathbb{R}^p$ such that $\mathbf{b} \perp \boldsymbol{\beta}$. We have:

$$\mathbb{E}[\boldsymbol{X}|Y]^\intercal \mathbf{b} = \mathbb{E}[\boldsymbol{X}^\intercal \mathbf{b}] = 0,$$

since $\boldsymbol{X}^\intercal \mathbf{b}$ is independent of $\boldsymbol{X}^\intercal \boldsymbol{\beta}$ and ε. This implies that

$$\mathbb{E}[\boldsymbol{X}|Y] = c(Y)\boldsymbol{\beta}, \tag{43}$$

for some real valued function c. Since $\boldsymbol{\beta}$ is a unit vector it follows that $c(Y) = \mathbb{E}[\boldsymbol{X}^\intercal \boldsymbol{\beta}|Y]$. Take $j \in \mathcal{S}_{\boldsymbol{\beta}}$ and apply (43) to get:

$$\mathrm{Var}[\mathfrak{m}_j(Y)] = \frac{\mathrm{Var}[\mathbb{E}[\boldsymbol{X}^\intercal \boldsymbol{\beta}|Y]]}{s}. \tag{44}$$

Combining the observation above with the following two inequalities:

$$A \leq \mathrm{Var}(\mathbb{E}[\boldsymbol{X}^\intercal \boldsymbol{\beta}|Y]) \leq \mathrm{Var}(\boldsymbol{X}^\intercal \boldsymbol{\beta}) = 1,$$

gives the desired result. □

Proof of Proposition 1. We first note that without loss of generality we can consider the function $\widetilde{\mathfrak{m}}$ to be non-negative, at the price of potentially shrinking the interval (B_0, ∞) to $(B_0 + \eta, \infty)$ by any $\eta > 0$. To see this fix an $\epsilon > 0$, and define $\widetilde{\mathfrak{m}}'(x) = \widetilde{\mathfrak{m}}(x) - \widetilde{\mathfrak{m}}(B_0 + \eta)$ for $x \in (B_0 + \eta, \infty)$. Then since (6) holds on $(-\infty, -B_0) \cup (B_0, \infty)$, clearly:

$$|\mathfrak{m}(x) - \mathfrak{m}(y)| \leq |\widetilde{\mathfrak{m}}'(|x|) - \widetilde{\mathfrak{m}}'(|y|)|, \text{ for } x, y \in (-\infty, -B_0 - \eta) \text{ or } (B_0 + \eta, +\infty).$$

By the convexity of the map $x \mapsto x^{2+\xi}$ we have $\widetilde{\mathfrak{m}}'(x)^{2+\xi} \leq 2^{1+\xi}(|\widetilde{\mathfrak{m}}(x)|^{2+\xi} + |\widetilde{\mathfrak{m}}(B_0 + \eta)|^{2+\xi})$ and hence $\mathbb{E}[|\widetilde{\mathfrak{m}}'(|Y|)|^{(2+\xi)}] < \infty$. Finally by definition $\widetilde{\mathfrak{m}}'$ is non-negative and non-decreasing on $(B_0 + \eta, \infty)$.

Next note that if Y has a bounded support, this proposition clearly follows from assumption (5) alone. Thus, without loss of generality we assume that Y has unbounded support (from both sides, as if one of them is bounded we can handle it in much the same way as the proof below).

Let $\widetilde{B}_0 = B_0 + \eta$, for some small fixed $\eta > 0$. Fix any partition $a \in \mathcal{A}_H(l, K)$. Let $S_0 = \{h : a_h \in [-\widetilde{B}_0, \widetilde{B}_0]\}$, and let $h_m = \min S_0, h_M = \max S_0$. Note that the following simple inequality holds for any $2 \leq h \leq h_m - 2$ or $h_M + 1 \leq h \leq H - 1$:

$$\begin{aligned}\mathrm{Var}[\mathfrak{m}(Y)|a_h < Y \leq a_{h+1}] &\leq \inf_{t \in (a_h, a_{h+1}]} \mathbb{E}[(\mathfrak{m}(Y) - \mathfrak{m}(t))^2 | a_h < Y \leq a_{h+1}] \\ &\leq \sup_{y,t \in (a_h, a_{h+1}]} (\mathfrak{m}(y) - \mathfrak{m}(t))^2 \leq (\widetilde{\mathfrak{m}}(|a_h|) - \widetilde{\mathfrak{m}}(|a_{h+1}|))^2.\end{aligned}$$

This gives us the following inequality:

$$\begin{aligned}\sum_{h=2}^{h_m - 2} \mathrm{Var}[\mathfrak{m}(Y)|a_h < Y \leq a_{h+1}] &\leq \sum_{h=2}^{h_m - 2} (\widetilde{\mathfrak{m}}(|a_h|) - \widetilde{\mathfrak{m}}(|a_{h+1}|))^2 \\ &\leq (\widetilde{\mathfrak{m}}(|a_2|) - \widetilde{\mathfrak{m}}(|a_{h_m - 1}|))^2,\end{aligned} \tag{45}$$

where the last inequality holds since $\widetilde{\mathfrak{m}}$ is non-decreasing. Similar inequality holds for the other tail as well.

Using a similar technique we obtain the following bound on the interval: $[-\widetilde{B}_0, \widetilde{B}_0]$:

$$\sum_{h=h_m}^{h_M-1} \mathrm{Var}[\mathfrak{m}(Y)|a_h < Y \leq a_{h+1}] \leq \sum_{h=h_m}^{h_M-1} \mathbb{E}[(\mathfrak{m}(Y) - \mathfrak{m}(a_h))^2 | a_h < Y \leq a_{h+1}]$$
$$\leq \sum_{h=h_m}^{h_M-1} \sup_{y \in (a_h, a_{h+1}]} (\mathfrak{m}(y) - \mathfrak{m}(a_h))^2.$$

Notice further that:

$$\mathrm{Var}[\mathfrak{m}(Y)|a_{h_m-1} < Y \leq a_{h_m}] \leq \sup_{y \in (a_{h_m-1}, a_{h_m}]} (\mathfrak{m}(y) - \mathfrak{m}(-\widetilde{B}_0))^2$$
$$\leq \sup_{y \in (a_{h_m-1}, -\widetilde{B}_0]} (\mathfrak{m}(y) - \mathfrak{m}(-\widetilde{B}_0))^2 + \sup_{y \in [-\widetilde{B}_0, a_{h_m}]} (\mathfrak{m}(y) - \mathfrak{m}(-\widetilde{B}_0))^2.$$

And a similar inequality holds for $\mathrm{Var}[\mathfrak{m}(Y)|a_{h_M} < Y \leq a_{h_M+1}]$. Thus:

$$\sum_{h=h_m-1}^{h_M} \mathrm{Var}[\mathfrak{m}(Y)|a_h < Y \leq a_{h+1}] \leq \underbrace{\sum_{h=h_m}^{h_M-1} \sup_{y \in (a_h, a_{h+1}]} (\mathfrak{m}(y) - \mathfrak{m}(a_h))^2}_{I_1}$$
$$+ \underbrace{\sup_{y \in (a_{h_m-1}, -\widetilde{B}_0]} (\mathfrak{m}(y) - \mathfrak{m}(-\widetilde{B}_0))^2}_{I_2} + \underbrace{\sup_{y \in [-\widetilde{B}_0, a_{h_m}]} (\mathfrak{m}(y) - \mathfrak{m}(-\widetilde{B}_0))^2}_{I_3}$$
$$+ \underbrace{\sup_{y \in [\widetilde{B}_0, a_{h_M+1}]} (\mathfrak{m}(y) - \mathfrak{m}(\widetilde{B}_0))^2}_{I_4} + \underbrace{\sup_{y \in (a_{h_M}, \widetilde{B}_0]} (\mathfrak{m}(y) - \mathfrak{m}(\widetilde{B}_0))^2}_{I_5}.$$

We have:

$$I_1 + I_3 + I_5 \leq \sup_{b \in \Pi_{2|S_0|+3}(\widetilde{B}_0)} \sum_{i=2}^{2|S_0|+3} (\mathfrak{m}(b_i) - \mathfrak{m}(b_{i-1}))^2 \tag{46}$$
$$\leq \sup_{b \in \Pi_{2|S_0|+3}(\widetilde{B}_0)} \left(\sum_{i=2}^{2|S_0|+3} |\mathfrak{m}(b_i) - \mathfrak{m}(b_{i-1})| \right)^2. \tag{47}$$

To see this, consider a partition containing the points $b_1 = -\widetilde{B}_0, b_3 = a_{h_m}, \ldots, b_{2|S_0|+1} = a_{h_M}, b_{2|S_0|+3} = \widetilde{B}_0$, and $b_{2k} = \mathrm{argmax}_{y \in (b_{2k-1}, b_{2k+1}]} (\mathfrak{m}(y) - \mathfrak{m}(b_{2k-1}))^2, k \in [|S_0|]$ and $b_{2|S_0|+2} = \mathrm{argmax}_{y \in (b_{2|S_0|+1}, \widetilde{B}_0]} (\mathfrak{m}(y) - \mathfrak{m}(\widetilde{B}_0))^2$ (note that if the max doesn't exist we can take a limit of partitions converging to it).

Next, we control I_2:

$$I_2 = \sup_{y \in (a_{h_m-1}, -\widetilde{B}_0]} (\mathfrak{m}(y) - \mathfrak{m}(-\widetilde{B}_0))^2 \leq (\widetilde{\mathfrak{m}}(|a_{h_m-1}|) - \widetilde{\mathfrak{m}}(\widetilde{B}_0))^2.$$

with the last inequality following from (6). Combining this bound with (45) we get:

$$(\widetilde{\mathfrak{m}}(|a_2|) - \widetilde{\mathfrak{m}}(|a_{h_m-1}|))^2 + I_2 \leq (\widetilde{\mathfrak{m}}(|a_2|) - \widetilde{\mathfrak{m}}(\widetilde{B}_0))^2. \tag{48}$$

Similarly, for I_4 and the other bound in (45) we have:

$$(\widetilde{\mathfrak{m}}(|a_H|) - \widetilde{\mathfrak{m}}(|a_{h_M+1}|))^2 + I_4 \leq (\widetilde{\mathfrak{m}}(|a_H|) - \widetilde{\mathfrak{m}}(\widetilde{B}_0))^2. \tag{49}$$

Finally, we deal with the tail part:

$$\begin{aligned}
\mathrm{Var}[\mathfrak{m}(Y)|Y \leq a_2] &\leq \mathbb{E}[(\mathfrak{m}(Y) - \mathfrak{m}(a_2))^2 | Y \leq a_2] \\
&\leq \mathbb{E}[(\widetilde{\mathfrak{m}}(|Y|) - \widetilde{\mathfrak{m}}(|a_2|))^2 | Y \leq a_2] \\
&\leq 4\mathbb{E}[(\widetilde{\mathfrak{m}}(|Y|))^2 | Y \leq a_2] \leq 4(\mathbb{E}[|\widetilde{\mathfrak{m}}(|Y|)|^{2+\xi} | Y \leq a_2])^{2/(2+\xi)} \\
&= 4\left(\int_{-\infty}^{a_2} |\widetilde{\mathfrak{m}}(|y|)|^{2+\xi} d\mathbb{P}(Y \leq y)\mathbb{P}(Y \leq a_2)^{-1}\right)^{2/(2+\xi)} \\
&= o(1)\mathbb{P}(Y \leq a_2)^{-2/(2+\xi)}.
\end{aligned} \tag{50}$$

where we used the fact that $\mathbb{E}[|\widetilde{\mathfrak{m}}(|Y|)|^{2+\xi}]$ is bounded by assumption, and the $o(1)$ is in the sense of $|a_2| \to \infty$. We can show a similar inequality for the other tail — $\mathrm{Var}[\mathfrak{m}(Y)|Y \geq a_H]$.

Combining (45), (48), (49), (46) and (50) we have:

$$\begin{aligned}
&\sum_{h=1}^{H} \mathrm{Var}[\mathfrak{m}(Y)|a_h < Y \leq a_{h+1}] \leq \sup_{b \in \Pi_{2|S_0|+3}(\widetilde{B}_0)} \left(\sum_{i=2}^{2|S_0|+3} |\mathfrak{m}(b_i) - \mathfrak{m}(b_{i-1})|\right)^2 \\
&+ o(1)\mathbb{P}(Y \geq a_H)^{-2/(2+\xi)} + o(1)\mathbb{P}(Y \leq a_2)^{-2/(2+\xi)} \\
&+ (\widetilde{\mathfrak{m}}(|a_2|) - \widetilde{\mathfrak{m}}(\widetilde{B}_0))^2 + (\widetilde{\mathfrak{m}}(|a_H|) - \widetilde{\mathfrak{m}}(\widetilde{B}_0))^2.
\end{aligned}$$

Since $(\widetilde{\mathfrak{m}}(|a_2|) - \widetilde{\mathfrak{m}}(\widetilde{B}_0))^2 \leq 4(\widetilde{\mathfrak{m}}(|a_2|))^2$, and we know that $\widetilde{\mathfrak{m}}(|a_2|)^{2+\xi}\frac{l}{H} \leq \widetilde{\mathfrak{m}}(|a_2|)^{2+\xi}\mathbb{P}(Y \leq a_2) \to 0$, this means that $\widetilde{\mathfrak{m}}(|a_2|)^2 \frac{1}{H^{2/(2+\xi)}} \to 0$. Furthremore $o(1)\mathbb{P}(Y \leq a_2)^{-2/(2+\xi)}\frac{1}{H^{2/(2+\xi)}} = o(1)$. Finally we recall that by (5) we have:

$$\sup_{b \in \Pi_{2|S_0|+3}(\widetilde{B}_0)} \left(\sum_{i=2}^{2|S_0|+3} |\mathfrak{m}(b_i) - \mathfrak{m}(b_{i-1})|\right)^2 = o\left(|S_0|^{2/(2+\xi)}\right).$$

However $|S_0| \leq \mathbb{P}(-\widetilde{B}_0 \leq Y \leq \widetilde{B}_0)H/l + 1$ and thus:

$$\sup_{b \in \Pi_{2|S_0|+3}(\widetilde{B}_0)} \left(\sum_{i=2}^{2|S_0|+3} |\mathfrak{m}(b_i) - \mathfrak{m}(b_{i-1})|\right)^2 = o\left(H^{2/(2+\xi)}\right),$$

which finishes the proof. □

Proof of Lemma 2. Before we go to the main proof of the lemma we first formulate a simple but useful concentration inequality.

Lemma 7. *Let $\widetilde{X}$ be a sub-Gaussian random variable with $\|\widetilde{X}\|_{\psi_2} \leq \mathcal{K}$. Let $A(\widetilde{X}, \nu) \in \{0, 1\}$ be any (randomized) acceptance rule such that $\mathbb{P}(A = 1) \geq q$, with ν being any random variable. Let $X_1, \ldots, X_r$ be an i.i.d. samples of the distribution $\widetilde{X}|A(\widetilde{X}, \nu) = 1$. Denote with $\mu = \mathbb{E}[X_i]$. Then there exist some absolute constants $C_1, C_2 > 0$ such that:*

$$\mathbb{P}(|\overline{X} - \mu| > \epsilon) \leq 2\exp\left(-\frac{\epsilon^2 r}{C_1\mathcal{K}^2\exp(\sqrt{1 - \log(q)}) + C_2\mathcal{K}\epsilon}\right).$$

As a Corollary to Lemma 7, observe that since $\sup_{q\in[0,1]} \exp(\sqrt{1-\log(q)})q = e$, we then have the following:

$$\mathbb{P}(|\overline{X} - \mu| > \epsilon) \leq 2\exp\left(-\frac{\epsilon^2 r}{C_1\mathcal{K}^2 q^{-1} + C_2\mathcal{K}\epsilon}\right),$$

for some absolute constants $C_1, C_2 > 0$.

By (20) we know that $\mathbb{P}(Y \in S_h) \geq \frac{1}{H} - 2\epsilon$ on E, thus setting $q = \frac{1}{H} - 2\epsilon$, by Lemma 7 conditionally on $\{Y_{(mh)} : h \in [H-1]\}$ we have for all $j \in G$ and all h:

$$\mathbb{P}\left(\left|\overline{\boldsymbol{X}}^j_{h,1:(m-1)} - \boldsymbol{\mu}^j_h\right| > \eta\right) \leq 2\exp\left(-\frac{\eta^2(m-1)}{C_1\mathcal{K}^2 q^{-1} + C_2\mathcal{K}\eta}\right).$$

Note that Lemma 7 is applicable in this case, since the statistics $\boldsymbol{X}^j_{h,\pi_i}$ are i.i.d. conditionally on $Y_{(m(h-1))}$ and $Y_{(mh)}$, where π is a random permutation, as we noticed in the main text. Therefore $\boldsymbol{X}^j_{h,\pi_i} \stackrel{d}{=} \boldsymbol{X}^j | A(\boldsymbol{X}^j, \nu)$ for the acceptance rule $A(\boldsymbol{X}^j, \nu) := A(\boldsymbol{X}, \varepsilon) = \mathbb{1}(f(\boldsymbol{X}^\intercal\boldsymbol{\beta}, \varepsilon) \in S_h)$. Furthermore, notice that the above inequality holds regardless of the values of $\{Y_{(mh)} : h \in [H-1]\}$, on the event E.

Finally, using union bound across the slices and the indexes $j \in G$, we have that this holds for all slices or in other words

$$\mathbb{P}\left(\max_{j\in G, h\in[H]}\left|\overline{\boldsymbol{X}}^j_{h,1:(m-1)} - \boldsymbol{\mu}^j_h\right| > \eta\right) \leq 2|G|H\exp\left(-\frac{\eta^2(m-1)}{C_1\mathcal{K}^2 q^{-1} + C_2\mathcal{K}\eta}\right),$$

on the event E. This is precisely what we wanted to show. □

Proof of Lemma 7. Observe that the random variable $X \stackrel{d}{=} \widetilde{X} | A(\widetilde{X}, \nu)$ satisfies the following inequality:

$$\begin{aligned}\mathbb{P}(|X| \geq t) = \mathbb{P}(|\widetilde{X}| \geq t | A(\widetilde{X}, \nu) = 1) &= \frac{\mathbb{P}(|\widetilde{X}| \geq t, A(\widetilde{X}, \nu) = 1)}{\mathbb{P}(A(\widetilde{X}, \nu) = 1)} \\ &\leq q^{-1}\mathbb{P}(|\widetilde{X}| \geq t) \leq q^{-1} e \exp(-ct^2/\mathcal{K}^2),\end{aligned}$$

where c is an absolute constant, and the last inequality follows by the fact that $\widetilde{X}$ is assumed to be sub-Gaussian. Clearly the above bound can be substituted with the trivial bound 1, for values of $t \leq \sqrt{-\mathcal{K}^2/c \log\left(\frac{q}{e}\right)}$. Let $f(q) := \sqrt{-\mathcal{K}^2/c \log\left(\frac{q}{e}\right)}$. Next, for any positive integer $j \in \mathbb{N}$, we have:

$$\mathbb{E}[|X|^j] = \int_0^\infty \mathbb{P}(|X| \geq t) j t^{j-1} dt \leq \int_0^{f(q)} j t^{j-1} dt + \int_{f(q)}^\infty q^{-1} e \exp(-ct^2/\mathcal{K}^2) j t^{j-1} dt.$$

Multiplying by $\lambda^j/j!$ and summing over $j = 1, 2, \ldots$, we obtain:

$$\begin{aligned}\mathbb{E}[\exp(\lambda|X|)] &\leq 1 + \lambda\int_0^{f(q)} \exp(\lambda t) dt + \lambda \int_{f(q)}^\infty q^{-1} e \exp(-ct^2/\mathcal{K}^2)\exp(\lambda t) dt \\ &= \exp(\lambda f(q)) + \frac{\lambda e\sqrt{\pi}\mathcal{K}}{q\sqrt{c}}\exp(\mathcal{K}^2\lambda^2/4c)\left[1 - \Phi\left(\frac{f(q) - \mathcal{K}^2\lambda/(2c)}{\mathcal{K}/\sqrt{2c}}\right)\right].\end{aligned}$$

Assuming that λ is small enough so that $f(q) - \mathcal{K}^2\lambda/(2c) > 0$, we can use the well known tail bound $1 - \Phi(x) \leq \phi(x)/x$, for $x > 0$ to get:

$$\begin{aligned}\mathbb{E}[\exp(\lambda|X|)] &\leq \exp(\lambda f(q)) + \frac{\lambda e \mathcal{K}^2\sqrt{\pi}\exp(\mathcal{K}^2\lambda^2/4c)}{q(f(q) - \mathcal{K}^2\lambda/(2c))\sqrt{2c}}\phi\left(\frac{f(q) - \mathcal{K}^2\lambda/(2c)}{\mathcal{K}/\sqrt{2c}}\right) \\ &= \exp(\lambda f(q))\left(1 + \frac{\lambda\mathcal{K}^2}{2c(f(q) - \mathcal{K}^2\lambda/(2c))}\right).\end{aligned}$$

Next, by the triangle inequality and the convexity of the exponent and the absolute value we have:

$$\mathbb{E}[\exp(\lambda|X - \mu|)] \leq \mathbb{E}[\exp(\lambda|X|)]\exp[\mathbb{E}[\lambda|X|]] \leq \mathbb{E}[\exp(2\lambda|X|)].$$

Set $Z := X - \mu$. We have showed the following:

$$\mathbb{E}[\exp[\lambda|Z|] - 1 - \lambda|Z|] \leq \mathbb{E}[\exp[\lambda|Z|]] \leq \exp(2\lambda f(q))\left(1 + \frac{\lambda\mathcal{K}^2}{c(f(q) - \mathcal{K}^2\lambda/c)}\right), \tag{51}$$

for a λ such that $f(q) - \mathcal{K}^2\lambda/c > 0$. By selecting $\lambda := \frac{1}{2}\sqrt{\frac{c}{\mathcal{K}^2}}$, one can easily verify that $(f(q) - \mathcal{K}^2\lambda/c) \geq \lambda^{-1}/4$, and hence $\frac{\lambda\mathcal{K}^2}{c(f(q) - \mathcal{K}^2\lambda/c)} \leq 1$, for any $q \leq 1$. With this choice of λ (51) becomes:

$$\mathbb{E}[\exp[\lambda|Z|] - 1 - \lambda|Z|]\lambda^{-2} \leq \frac{8\mathcal{K}^2}{c}\exp\left(\sqrt{\frac{c}{\mathcal{K}^2}}f(q)\right) = \frac{8\mathcal{K}^2}{c}\exp\left(\sqrt{-\log\frac{q}{e}}\right).$$

Recall that a version Bernstein's inequality (see Lemma 2.2.11 Van Der Vaart & Wellner (1996)) states that if the following moment condition $\mathbb{E}|Z|^j \leq j!\lambda^{-(j-2)}v/2$ is met for all $j \geq 2$ and $Z_i \sim Z, i \in [r]$ are mean 0, i.i.d. then:

$$\mathbb{P}\left(\left|\sum_{i=1}^{r} Z_i\right| \geq \epsilon\right) \leq 2\exp\left(-\frac{1}{2}\frac{\epsilon^2}{rv + \lambda^{-1}\epsilon}\right).$$

Observe that by a Taylor expansion the condition $\mathbb{E}[\exp[\lambda|Z|] - 1 - \lambda|Z|]\lambda^{-2} \leq \frac{1}{2}v$ implies the moment condition from above. Hence we conclude:

$$\mathbb{P}\left(\left|\frac{1}{r}\sum_{i=1}^{r} X_i - \mu\right| \geq \epsilon\right) \leq 2\exp\left(-\frac{1}{2}\frac{\epsilon^2 r}{16\mathcal{K}^2\exp(\sqrt{1 - \log(q)})/c + 2\mathcal{K}/\sqrt{c}\epsilon}\right),$$

which completes the proof. □

Proof of Lemma 3. Using the sliced stability condition, for large H we get:

$$\begin{aligned}\left|\mathrm{Var}[\mathfrak{m}_j(Y)] - \sum_{h=1}^{H}(\boldsymbol{\mu}_h^j)^2\mathbb{P}(Y \in S_h)\right| &= \sum_{h=1}^{H}\mathrm{Var}[\mathfrak{m}_j(Y)|Y \in S_h]\mathbb{P}(Y \in S_h) \\ &\leq \frac{CH^\kappa}{s}\left(\frac{1}{H} + 2\epsilon\right).\end{aligned}$$

This shows (22). Consequently we have:

$$\left(\frac{1}{H} - 2\epsilon\right)\sum_{h=1}^{H}(\boldsymbol{\mu}_h^j)^2 \leq \sum_{h=1}^{H}(\boldsymbol{\mu}_h^j)^2\mathbb{P}(Y \in S_h) \leq \frac{C_V}{s} + \frac{CH^\kappa}{s}\left(\frac{1}{H} + 2\epsilon\right).$$

This yields (23). To get (24) we proceed as follows:

$$\left(\frac{1}{H}-2\epsilon\right)\sum_{h=1}^{H}|\boldsymbol{\mu}_h^j| \le \sum_{h=1}^{H}|\boldsymbol{\mu}_h^j|\mathbb{P}(Y\in S_h) \le \sqrt{\sum_{h=1}^{H}\mathbb{P}(Y\in S_h)}\sqrt{\sum_{h=1}^{H}(\boldsymbol{\mu}_h^j)^2\mathbb{P}(Y\in S_h)}$$
$$\le \sqrt{\frac{C_V}{s}+\frac{CH^\kappa}{s}\left(\frac{1}{H}+2\epsilon\right)},$$

and we are done. □

Proof of Lemma 4. Note that on the event $\widetilde{E}$ we have the following chain of inequalities:

$$\frac{1}{H}\sum_{h=1}^{H}\left|\left(\frac{1}{m}\boldsymbol{X}_{h,m}^j+\frac{m-1}{m}\overline{\boldsymbol{X}}_{h,1:(m-1)}^j\right)^2-\frac{(m-1)^2}{m^2}(\boldsymbol{\mu}_h^j)^2\right| \tag{52}$$
$$\le \frac{1}{Hm^2}\sum_{h=1}^{H}(\boldsymbol{X}_{h,m}^j)^2+\frac{2(m-1)}{Hm^2}\sum_{h=1}^{H}|\boldsymbol{X}_{h,m}^j|(\eta+|\boldsymbol{\mu}_h^j|)+\frac{1}{H}\frac{(m-1)^2}{m^2}\sum_{h=1}^{H}\eta(2|\boldsymbol{\mu}_h^j|+\eta)$$
$$\le \frac{1}{m}\frac{1}{n}\sum_{r=1}^{n}(\boldsymbol{X}_r^j)^2+\frac{2}{Hm}\sqrt{\sum_{r=1}^{n}(\boldsymbol{X}_r^j)^2}\sqrt{2\sum_{h=1}^{H}(\eta^2+(\boldsymbol{\mu}_h^j)^2)}+\eta^2+2\frac{\eta}{H}B_3.$$

where we used that we are on the event $\widetilde{E}$ in the first inequality, and (24), Cauchy-Schwartz and the trivial bounds $\frac{m-1}{m}<1$, $\frac{1}{n}\sum_{h=1}^{H}(\boldsymbol{X}_{h,m}^j)^2\le\frac{1}{n}\sum_{r=1}^{n}(\boldsymbol{X}_r^j)^2$ in the second one.

To this end, observe that $(\boldsymbol{X}_r^j)^2, r\in[n]$ are i.i.d. random variables with sub-exponential distributions. This can be seen from the standard inequality:

$$\|(\boldsymbol{X}_r^j)^2\|_{\psi_1}\le 2\|\boldsymbol{X}_r^j\|_{\psi_2}^2\le 2\mathcal{K}^2.$$

Clearly we also have $\mathbb{E}[(\boldsymbol{X}_r^j)^2]\le 2\mathcal{K}^2$. Denote the mean $\mathbb{E}[(\boldsymbol{X}_r^j)^2]=\nu$. Now we are in a position to use a Bernstein type of deviation inequality (see Proposition 5.16 in Vershynin (2010)). We obtain:

$$\mathbb{P}\left(\frac{1}{n}\sum_{r=1}^{n}(\boldsymbol{X}_r^j)^2>\nu+\tau\right)\le\exp\left(-c\min\left(\frac{\tau^2 n}{4\mathcal{K}^4},\frac{\tau n}{2\mathcal{K}^2}\right)\right),$$

for some absolute constant $c>0$. Hence we infer that, when $\tau\le 2\mathcal{K}^2$, there exists a set $\widetilde{\widetilde{E}}\subset\widetilde{E}$ failing with probability at most $p\exp(-c\frac{\tau^2 n}{4\mathcal{K}^4})$, such that $\frac{1}{n}\sum_{r=1}^{n}(\boldsymbol{X}_r^j)^2\le\nu+\tau\le 2\mathcal{K}^2+\tau$ for all $j\in[p]$. Therefore continuing the bound on the event $\widetilde{\widetilde{E}}$, we get:

$$(52)\le\frac{(2\mathcal{K}^2+\tau)}{m}+\frac{2\sqrt{2\mathcal{K}^2+\tau}}{\sqrt{m}}\sqrt{2\eta^2+2\frac{B_2}{H}}+\eta^2+2\eta\frac{B_3}{H},$$

where we used (23). This finishes the proof.

□

Proof of Lemma 6. Fix any $\tau \in (0, 2]$. Define the event:

$$\widetilde{\widetilde{E}}_{\mathcal{S}_{\boldsymbol{\beta}}} = \widetilde{E}_{\mathcal{S}_{\boldsymbol{\beta}}} \cap \{\sup_{j \in \mathcal{S}_{\boldsymbol{\beta}}} n^{-1} \sum_{i=1}^{n} (\boldsymbol{X}_r^j)^2 \leq 2 + \tau\}$$

We first note that the following inequality holds:

$$\left| \mathbf{V}^{jk} - \operatorname{sign}(\boldsymbol{\beta}_j) \operatorname{sign}(\boldsymbol{\beta}_k) \frac{C_V}{s} \right| = \left| \frac{1}{H} \sum_{h=1}^{H} \overline{\boldsymbol{X}}_h^j \overline{\boldsymbol{X}}_h^k - \operatorname{sign}(\boldsymbol{\beta}_j) \operatorname{sign}(\boldsymbol{\beta}_k) \frac{C_V}{s} \right|$$
$$\leq \left| \frac{1}{H} \sum_{h=1}^{H} (\overline{\boldsymbol{X}}_h^j)^2 - \frac{C_V}{s} \right| + \frac{1}{H} \sum_{h=1}^{H} \left| \overline{\boldsymbol{X}}_h^j \right| \left| \operatorname{sign}(\boldsymbol{\beta}_j) \overline{\boldsymbol{X}}_h^j - \operatorname{sign}(\boldsymbol{\beta}_k) \overline{\boldsymbol{X}}_h^k \right|,$$

where the LHS equals, the LHS of (36) after using (34). Fortunately (25) and (27) already give bounds on the first term on the event $\widetilde{\widetilde{E}}_{\mathcal{S}_{\boldsymbol{\beta}}}$, which can be seen with identical arguments to Lemmas 3 and 4. We now show that the second term is small on the same event. Note that the following identity holds:

$$\frac{1}{H} \sum_{h=1}^{H} \left| \overline{\boldsymbol{X}}_h^j \right| \left| \operatorname{sign}(\boldsymbol{\beta}_j) \overline{\boldsymbol{X}}_h^j - \operatorname{sign}(\boldsymbol{\beta}_k) \overline{\boldsymbol{X}}_h^k \right|$$
$$= \frac{1}{H} \sum_{h=1}^{H} \left| \frac{m-1}{m} \overline{\boldsymbol{X}}_{h,1:(m-1)}^j + \frac{1}{m} \boldsymbol{X}_{h,m}^j \right| \left| \operatorname{sign}(\boldsymbol{\beta}_j) \overline{\boldsymbol{X}}_h^j - \frac{m-1}{m} \operatorname{sign}(\boldsymbol{\beta}_k) \boldsymbol{\mu}_h^j \right.$$
$$\left. + \frac{m-1}{m} \operatorname{sign}(\boldsymbol{\beta}_k) \boldsymbol{\mu}_h^k - \operatorname{sign}(\boldsymbol{\beta}_k) \overline{\boldsymbol{X}}_h^k \right|.$$

Thus on the event $\widetilde{E}_{\mathcal{S}_{\boldsymbol{\beta}}}$:

$$\frac{1}{H} \sum_{h=1}^{H} \left| \overline{\boldsymbol{X}}_h^j \right| \left| \operatorname{sign}(\boldsymbol{\beta}_j) \overline{\boldsymbol{X}}_h^j - \operatorname{sign}(\boldsymbol{\beta}_k) \overline{\boldsymbol{X}}_h^k \right|$$
$$\leq \frac{1}{H} \sum_{h=1}^{H} \left(\boldsymbol{\mu}_h + \eta + \frac{1}{m} \left| \boldsymbol{X}_{h,m}^j \right| \right) \left(2\eta + \frac{1}{m} \left| \boldsymbol{X}_{h,m}^j \right| + \frac{1}{m} \left| \boldsymbol{X}_{h,m}^k \right| \right),$$

where $\boldsymbol{\mu}_h = |\boldsymbol{\mu}_h^j| = |\boldsymbol{\mu}_h^k|$, and we used that $\frac{m-1}{m} < 1$, and the fact that on $\widetilde{E}_{\mathcal{S}_{\boldsymbol{\beta}}}$ we have $|\boldsymbol{X}_{h,1:(m-1)}^j - \boldsymbol{\mu}_h^j| \leq \eta$ and similarly $|\boldsymbol{X}_{h,1:(m-1)}^k - \boldsymbol{\mu}_h^k| \leq \eta$. Next we have:

$$\frac{1}{H} \sum_{h=1}^{H} \left(\boldsymbol{\mu}_h + \eta + \frac{1}{m} \left| \boldsymbol{X}_{h,m}^j \right| \right) \left(2\eta + \frac{1}{m} \left| \boldsymbol{X}_{h,m}^j \right| + \frac{1}{m} \left| \boldsymbol{X}_{h,m}^k \right| \right)$$
$$\leq 2\frac{\eta}{H} \sum_{h=1}^{H} \boldsymbol{\mu}_h + \frac{1}{Hm} \sum_{h=1}^{H} (\boldsymbol{\mu}_h + \eta)(|\boldsymbol{X}_{h,m}^j| + |\boldsymbol{X}_{h,m}^k|) + 2\eta^2$$
$$+ \frac{2\eta}{mH} \sum_{h=1}^{H} |\boldsymbol{X}_{h,m}^j| + \frac{1}{m^2 H} \sum_{h=1}^{H} (\boldsymbol{X}_{h,m}^j)^2 + \frac{1}{2m^2 H} \sum_{h=1}^{H} (\boldsymbol{X}_{h,m}^j)^2 + \frac{1}{2m^2 H} \sum_{h=1}^{H} (\boldsymbol{X}_{h,m}^k)^2,$$

where we used the simple inequality $ab \leq (a^2+b^2)/2$. Luckily we have already controlled all of the above quantities. Using Lemma 4 and (25) we get:

$$\frac{1}{H}\sum_{h=1}^{H}\left|\overline{\boldsymbol{X}}_h^j\right|\left|\text{sign}(\boldsymbol{\beta}_j)\overline{\boldsymbol{X}}_h^j - \text{sign}(\boldsymbol{\beta}_k)\overline{\boldsymbol{X}}_h^k\right|$$
$$\leq 2\frac{\eta}{H}B_3 + \frac{2\sqrt{2+\tau}}{\sqrt{m}}\sqrt{2\eta^2 + 2\frac{B_2}{H}} + 2\eta^2 + 2\frac{\eta\sqrt{2+\tau}}{\sqrt{m}} + 2\frac{2+\tau}{m},$$

where we heavily relied on the fact that on $\widetilde{\widetilde{E}}_{\mathcal{S}_{\boldsymbol{\beta}}}$ we have $\frac{1}{mH}\sum_{r=1}^{n}(\boldsymbol{X}_r^j)^2 \leq 2+\tau$, the rest of the bounds can be seen in the proof of Lemma 4. (For the term note $\frac{1}{mH}\sum_{h=1}^{H}|\boldsymbol{X}_{h,m}^j| \leq \frac{1}{m}\sqrt{\sum_{h=1}^{H}(\boldsymbol{X}_{h,m}^j)^2/H} \leq \frac{1}{\sqrt{m}}\sqrt{\sum_{r=1}^{n}(\boldsymbol{X}_r^j)^2/mH}$). Finally noting that $2\frac{\eta\sqrt{2+\tau}}{\sqrt{m}} \leq \eta^2 + \frac{2+\tau}{m}$ gives the desired result. □

Lemma 8. *Let* $\mathbf{N}$ *be a* $s \times s$ *symmetric "noise" matrix satisfying:*

$$\|\mathbf{N}\|_{\max} \leq \frac{C_V}{40s}. \tag{53}$$

Then the following occurs:

(a) Let γ_1 *be the maximum eigenvalue of* $\widetilde{\mathbf{V}}_{\mathcal{S}_{\boldsymbol{\beta}},\mathcal{S}_{\boldsymbol{\beta}}} = \frac{C_V}{2}\boldsymbol{\beta}_{\mathcal{S}_{\boldsymbol{\beta}}}\boldsymbol{\beta}_{\mathcal{S}_{\boldsymbol{\beta}}}^{\intercal} + \mathbf{N}$, *i.e.* $\gamma_1 := \lambda_{\max}(\widetilde{\mathbf{V}}_{\mathcal{S}_{\boldsymbol{\beta}},\mathcal{S}_{\boldsymbol{\beta}}})$. *Then we have* $|\gamma_1 - \frac{C_V}{2}| \leq \frac{C_V}{40}$; *Furthermore if* γ_2 *is the second largest in magnitude eigenvalue of* $\widetilde{\mathbf{V}}_{\mathcal{S}_{\boldsymbol{\beta}},\mathcal{S}_{\boldsymbol{\beta}}}$, *we have* $|\gamma_2| \leq \frac{C_V}{40}$, *and hence* $\gamma_1 > |\gamma_2|$.

(b) The corresponding principal eigenvector of $\widetilde{\mathbf{V}}_{\mathcal{S}_{\boldsymbol{\beta}},\mathcal{S}_{\boldsymbol{\beta}}}$ — $\widetilde{\mathbf{z}}_{\mathcal{S}_{\boldsymbol{\beta}}}$[4] *satisfies the following inequality:*

$$\left\|\widetilde{\mathbf{z}}_{\mathcal{S}_{\boldsymbol{\beta}}} - \boldsymbol{\beta}_{\mathcal{S}_{\boldsymbol{\beta}}}\right\|_\infty \leq \frac{1}{2\sqrt{s}}.$$

Proof. This Lemma is a re-statement of Lemma 6 from Amini & Wainwright (2008), the difference being that we require precise bounds, rather than considering simply the asymptotics. To see part (a), observe that (53) implies:

$$\|\mathbf{N}\|_{2,2} = \sup_{\|\mathbf{v}\|_2=1}|\mathbf{v}^{\intercal}\mathbf{N}\mathbf{v}| \leq \|\mathbf{v}\|_1^2\|\mathbf{N}\|_{\max} \leq \frac{C_V}{40}. \tag{54}$$

Hence by Weyl's inequality we have that:

$$|\gamma_1 - \frac{C_V}{2}| \leq \frac{C_V}{40}, \quad |\gamma_2 - 0| \leq \frac{C_V}{40},$$

which is exactly part (a). For the second part observe that

$$\|\mathbf{N}\|_{\infty,\infty} = \max_i \sum_j |\mathbf{N}_{ij}| \leq s\frac{C_V}{40s} = \frac{C_V}{40}. \tag{55}$$

The proof of part (b) can then be seen as in Lemma 6 of Amini & Wainwright (2008), by carefully exploiting (54) and (55).

□

[4] Here we mean the principal eigenvector oriented so that $\widetilde{\mathbf{z}}_{\mathcal{S}_{\boldsymbol{\beta}}}^{\intercal}\boldsymbol{\beta}_{\mathcal{S}_{\boldsymbol{\beta}}} \geq 0$.

Proof of Proposition 2. The proof is based on an application of Fano's inequality, which in turn is a standard approach for showing minimax lower bounds (e.g. see Cover & Thomas (2012), Wainwright (2009), Yang & Barron (1999), Yu (1997) among others). In particular we base our proof on ideas from Amini & Wainwright (2008), Wainwright (2009).

For two probability measures P and Q, which are absolutely continuous with respect to a third probability measure μ define their KL divergence by $D(P\|Q) = \int p \log \frac{p}{q} d\mu$, where $p = \frac{dP}{d\mu}$, $q = \frac{dQ}{d\mu}$. We proceed with the following lemma:

Lemma 9. *Let us have n observations from a SIM $Y = f(\boldsymbol{X}^\intercal \boldsymbol{\beta}, \varepsilon), \boldsymbol{X} \sim N_p(0, \mathbb{I}_p)$. Assume that for any fixed $u, v \in \mathbb{R}$ the following regularity condition holds for the random variables $f(u, \varepsilon)$ and $f(v, \varepsilon)$:*

$$D(p(f(u,\varepsilon))\|p(f(v,\varepsilon))) \leq \exp(\Xi(u-v)^2) - 1, \tag{56}$$

where Ξ is a positive constant. Then if

$$n < \frac{s \log(p-s+1)}{8\Xi},$$

and $s \geq 8\Xi$, any algorithm recovering the support in our model will have errors with probability at least $\frac{1}{2}$ asymptotically.

To finish the proof simply apply Example 3 with $P(x) = x/(2\sigma^2)$ and $G = h = Id$. □

Proof of Lemma 9. For simplicity of the exposition we will assume that the vector $\boldsymbol{\beta}$ has only non-negative entries (i.e. all non-zero entries are $\frac{1}{\sqrt{s}}$). The proof extends in exactly the same way in the case when the entries of $\boldsymbol{\beta}$ are not restricted to be positive.

Let $\mathbb{S} \subset 2^{[p]}$, the set of all subsets of $[p]$ with s elements. Clearly, $|\mathbb{S}| = \binom{p}{s}$. Let $\widehat{S} : (\mathbb{R}^{p+1})^n \to \mathbb{S}$ be any potentially random function, which is used to recover the support of $\boldsymbol{\beta}$, based on the sample $\{(Y_i, \boldsymbol{X}_i)\}_{i=1}^n$. Under the 0-1 loss the risk equals the probability of error:

$$\frac{1}{\binom{p}{s}} \sum_{\mathcal{S}_{\boldsymbol{\beta}} \in \mathbb{S}} P_{\mathcal{S}_{\boldsymbol{\beta}}}(\widehat{S} \neq \mathcal{S}_{\boldsymbol{\beta}}), \tag{57}$$

where by $P_{\mathcal{S}_{\boldsymbol{\beta}}}$ we are measuring the probability under a dataset generated with $\text{supp}(\boldsymbol{\beta})$ equal to the index of the measure $P_{\mathcal{S}_{\boldsymbol{\beta}}}$.

Instead of directly dealing with the sum above, we first consider the $p - s + 1$ element set $\widetilde{\mathbb{S}} = \{S \in \mathbb{S} : [s-1] \subset S, |S| = s\}$, and we bound the probability of error, on any function $\widehat{S}$ (even if given the knowledge that the true support is drawn from $\widetilde{\mathbb{S}}$). Let J be a uniformly distributed in $\widetilde{\mathbb{S}}$. By Fano's inequality that:

$$\mathbb{P}(\text{error}) \geq 1 - \frac{\mathcal{I}(J; (Y, \boldsymbol{X})^n) + \log(2)}{\log |\widetilde{\mathbb{S}}|}, \tag{58}$$

where $\mathcal{I}(J; (Y, \boldsymbol{X})^n)$ is the mutual information between the sample J and the sample $(Y, \boldsymbol{X})^n$. Note now that for the mutual information we have

$$\begin{aligned} \mathcal{I}(J; (Y, \boldsymbol{X})^n) &= \mathcal{I}(J; (f(\boldsymbol{X}^\intercal \boldsymbol{\beta}, \varepsilon), \boldsymbol{X})^n) \\ &\leq n\mathcal{H}((f(\boldsymbol{X}^\intercal \boldsymbol{\beta}, \varepsilon), \boldsymbol{X})) - n\mathcal{H}((f(\boldsymbol{X}^\intercal \boldsymbol{\beta}, \varepsilon), \boldsymbol{X})|J), \end{aligned}$$ [5]

where the last inequality follows from the chain inequality of entropy. We therefore need an upper bound on the last expression.

Let $i \geq s$ and set $\boldsymbol{\beta}^i := (\beta_1^i, \beta_2^i, \ldots, \beta_p^i)^\intercal$ the vector such that that $\beta_j^i = \frac{\mathbb{1}(j \in [s-1]) + \mathbb{1}(j=i)}{\sqrt{s}}$. When J is unknown the distribution is a mixture, and hence due to the convexity of $-\log$ we have the following inequality:

$$\begin{aligned}
&\mathcal{H}((f(\boldsymbol{X}^\intercal\boldsymbol{\beta}, \varepsilon), \boldsymbol{X})) - \mathcal{H}((f(\boldsymbol{X}^\intercal\boldsymbol{\beta}, \varepsilon), \boldsymbol{X}))|J) \\
&\leq \frac{1}{(p-s+1)^2} \sum_{i,j \geq s} p((f(\boldsymbol{X}^\intercal\boldsymbol{\beta}^i, \varepsilon), \boldsymbol{X})) \log \frac{p(f(\boldsymbol{X}^\intercal\boldsymbol{\beta}^i, \varepsilon), \boldsymbol{X})}{p(f(\boldsymbol{X}^\intercal\boldsymbol{\beta}^j, \varepsilon), \boldsymbol{X})} \\
&= \frac{p-s}{p-s+1} D(p((f(\boldsymbol{X}^\intercal\boldsymbol{\beta}^s, \varepsilon), \boldsymbol{X})) \| p((f(\boldsymbol{X}^\intercal\boldsymbol{\beta}^{s+1}, \varepsilon), \boldsymbol{X}))),
\end{aligned}$$

where the last equality follows by a symmetric argument. Since the KL divergence is invariant under changing variables, setting $U = \boldsymbol{X}^\intercal\boldsymbol{\beta}^s$, $V = \boldsymbol{X}^\intercal\boldsymbol{\beta}^{s+1}$ and $\boldsymbol{W} = \mathbf{P}_{\{\boldsymbol{\beta}^s, \boldsymbol{\beta}^{s+1}\}^\perp}\boldsymbol{X}$, where $\mathbf{P}_{\{\boldsymbol{\beta}^s, \boldsymbol{\beta}^{s+1}\}^\perp}$ denotes the orthogonal projection on the space $\operatorname{span}\{\boldsymbol{\beta}^s, \boldsymbol{\beta}^{s+1}\}^\perp$. We get:

$$\begin{aligned}
&\mathcal{H}((f(\boldsymbol{X}^\intercal\boldsymbol{\beta}, \varepsilon), \boldsymbol{X})) - \mathcal{H}((f(\boldsymbol{X}^\intercal\boldsymbol{\beta}, \varepsilon), \boldsymbol{X}))|J) \\
&\leq \frac{p-s}{p-s+1} D(p((f(U, \varepsilon), U, V)) \| p((f(V, \varepsilon), U, V))),
\end{aligned}$$

where we used the fact that $\boldsymbol{W}$ is independent of U, V, ε. Applying assumption (56) we get:

$$\begin{aligned}
D(p((f(U, \varepsilon), U, V)) \| p((f(V, \varepsilon), U, V))) &\leq \mathbb{E} \exp(\Xi(U-V)^2) - 1 \\
&= \sqrt{\frac{s}{s - 4\Xi}} - 1 \leq \frac{4\Xi}{s}.
\end{aligned}$$

where the first inequality can be obtained by conditioning on U, V, in the equality we used the fact that $U - V \sim N(0, \frac{2}{s})$, and we assume that the value of s is large enough so that $s \geq 8\Xi$. The inequality in the preceding display helps us to conclude that for large values of s:

$$\mathcal{I}(J; (Y, \boldsymbol{X})^n) \leq \frac{4\Xi(p-s)n}{s(p-s+1)} < \frac{4\Xi n}{s}.$$

Consequently by (58) if $n < \frac{s \log(p-s+1)}{8\Xi}$ we will have errors with probability at least $\frac{1}{2}$, asymptotically. To finish the conclusion, note that the sum (57), can be split into $\binom{p}{s-1}$ terms, by the following operation:

1. Repeat each set in $\mathbb{S}$ — s times, and denote this superset by $s \times \mathbb{S}$
2. For each S of the $\binom{p}{s-1}$, subsets of $[p]$ with $s-1$ elements, collect $p-s+1$ distinct elements of $s \times \mathbb{S}$ containing S
3. Apply the $\frac{1}{2}$ error bound obtained from above to this local sum.

In the end we get that the probability of error by selecting $S \subset \mathbb{S}$ uniformly is at least: $\frac{1}{s}\frac{\binom{p}{s-1}}{\binom{p}{s}}(p-s+1)\frac{1}{2} = \frac{1}{2}$. □

[5] Here we use $\mathcal{H}$ to denote the entropy, not to be confused with the number of slices H.

B SIM Examples

In this section we look into models of the type $Y = G(h(\boldsymbol{X}^\intercal\boldsymbol{\beta}) + \epsilon)$ and show that under certain sufficient conditions they belong to a class $\mathcal{F}_A$ for some A. In addition, we provide a sub-class of these models, in which support recovery is impossible with probability at least $\frac{1}{2}$ when Γ is small.

Example 1. For SIM $Y = G(h(\boldsymbol{X}^\intercal\boldsymbol{\beta}) + \varepsilon)$ with strictly monotone h and G there exists $A > 0$ (depending on h, G) such that $\mathrm{Var}(\mathbb{E}[\boldsymbol{X}^\intercal\boldsymbol{\beta}|Y]) \geq A > 0$.

Proof. We will show more generally that if $Y = f(\boldsymbol{X}^\intercal\boldsymbol{\beta}, \varepsilon)$, where f, ε satisfy the condition that the function $g(z) = \mathbb{E}[f(Z, \varepsilon)|Z = z]$ is strictly monotone, we have: $\mathrm{Var}(\mathbb{E}[Z|f(Z, \varepsilon)]) > 0$.

Note that the condition $\exists A > 0 : \mathrm{Var}(\mathbb{E}[Z|f(Z, \varepsilon)]) \geq A$ is equivalent $\mathbb{E}[Z|f(Z, \varepsilon)]$ to not being a constant. We argue that the latter is clearly implied if, for example, $\mathbb{E}[Zf(Z, \varepsilon)] \neq 0$. To see this assume the contrary, i.e. $\mathbb{E}[Z|f(Z, \varepsilon)] = 0$ a.s., but $\mathbb{E}[Zf(Z, \varepsilon)] \neq 0$. Then we have $\mathbb{E}[Zf(Z, \varepsilon)] = \mathbb{E}[\mathbb{E}[Z|f(Z, \varepsilon)]f(Z, \varepsilon)] = 0$, which is a contradiction.

Next we show that our condition implies $\mathbb{E}[Zf(Z, \varepsilon)] \neq 0$. WLOG assume that g is strictly increasing. Observe that since $Z \sim N(0, 1)$ is a continuous random variable, by Chebyshev's association inequality (Boucheron et al. 2013) we have:

$$\mathbb{E}[Zf(Z, \varepsilon)] = \mathbb{E}[Zg(Z)] > \mathbb{E}[Z]\mathbb{E}[g(Z)] = 0,$$

and hence as we argued earlier it follows that $\mathrm{Var}(\mathbb{E}[Z|f(Z, \varepsilon)]) > 0$. Due to the independence of $\boldsymbol{X}$ and ε models with a coordinate-wise monotone f function (strictly monotone in the first coordinate) belong to this class. □

Example 2. Let $Y = G(h(\boldsymbol{X}^\intercal\boldsymbol{\beta}) + \varepsilon)$, where G, h are strictly monotone continuous functions. Furthermore let ε be such that the function $\mathfrak{m}(y)$ is continuous and monotone in y. We argue that such models are sliced stable in the sense of Definition 1.

Proof. Since $\mathfrak{m}$ is a continuous and monotone function in y, it follows that it is of bounded variation on any closed interval, which in turn implies (5). Furthermore, by Example 1, we are guaranteed that $\sqrt{\mathrm{Var}(\mathfrak{m}_j(Y))s} = C_V$ for some $C_V > 0$. Notice also that in the case of a continuous and monotone $\mathfrak{m}$ one can take $\widetilde{\mathfrak{m}}(y) := |\mathfrak{m}(y) - \mathfrak{m}(-y)| < |\mathfrak{m}(y)| + |\mathfrak{m}(-y)|$, for $y > 0$. Finally we need to show that $\mathbb{E}[|\widetilde{\mathfrak{m}}(|Y|)|^{(2+\xi)}]$ is finite. Due to the last inequality it suffices to control:

$$\begin{aligned} C_V^{2+\xi}\mathbb{E}|\widetilde{\mathfrak{m}}(Y)|^{2+\xi} &\leq s^{(2+\xi)/2}\mathbb{E}|\mathbb{E}[\boldsymbol{X}^j|Y]|^{2+\xi} = s^{(2+\xi)/2}\mathbb{E}|\mathbb{E}[s^{-1/2}\boldsymbol{X}^\intercal\boldsymbol{\beta}|Y]|^{2+\xi} \\ &\leq s^{(2+\xi)/2}s^{-(2+\xi)/2}\mathbb{E}\mathbb{E}[|\boldsymbol{X}^\intercal\boldsymbol{\beta}|^{2+\xi}|Y] = \mathbb{E}|Z|^{2+\xi} < \infty, \end{aligned}$$

where $Z \sim N(0, 1)$ and in the first equality we used the fact that $\boldsymbol{X}^j - s^{-1/2}\boldsymbol{X}^\intercal\boldsymbol{\beta} \perp\!\!\!\perp \boldsymbol{X}^\intercal\boldsymbol{\beta}$ and hence $\mathbb{E}[\boldsymbol{X}^j - s^{-1/2}\boldsymbol{X}^\intercal\boldsymbol{\beta}|Y] = \mathbb{E}[\boldsymbol{X}^j - s^{-1/2}\boldsymbol{X}^\intercal\boldsymbol{\beta}] = 0$. This completes the proof. □

Remark 3. To see that the monotonicity condition on $\mathfrak{m}(y)$ is not vacuous, we give concrete examples of SIM satisfying it, which include the simple linear regression model (8) as a special case. Consider the models from Example 2 and let ε have a density function satisfying $p_\varepsilon(x) \propto \exp(-P(x^2))$, where P is any nonzero polynomial with non-negative coefficients such that $P(0) = 0$. To see that $\mathfrak{m}$ is monotonic and continuous, we start by obtaining a precise expression of $\mathfrak{m}$. It is simple to see that:

$$\mathfrak{m}(y) = \frac{\mathbb{E}[Z|h(Z) + \varepsilon = G^{-1}(y)]}{\sqrt{\mathrm{Var}(\mathfrak{m}_j(Y))s}},$$

for almost every $y \in \text{supp}(Y)$. By Example 1, we have that $\sqrt{\text{Var}(\mathfrak{m}_j(Y))s} = C_V$ for some $C_V > 0$. Next we argue that $y \mapsto \mathfrak{m}(y)$ is monotone. This is equivalent to showing that $\mathfrak{m}(G(y))$ is monotone. To this end we apply Lemma A.2 of Duan & Li (1991), which implies that it suffices for the random variable $h(Z) + \varepsilon | Z$ to have a monotone likelihood ratio in order for $\mathfrak{m}(G(y))$ to be monotone. Since the family of random variables $h(z) + \varepsilon$ is a location family, the normalizing constants of their densities do not change with z. This in conjuction with the fact that h is increasing, implies that the monotonicity of the likelihood ratio will be implied if we show that the function $x \mapsto P(x^2) - P((x-c)^2)$ is increasing in x for any fixed $c > 0$. Notice that since $P(x^2)$ is a differentiable convex function by construction, we have that $P(x^2) - P((x-c)^2) \geq c\frac{dP(y^2)}{dy}|_{y=x-c} > 0$. It is worth noting that the same argument applies more generally to the case where ε is a log-concave random variable (i.e. $p_\varepsilon(x) = \exp(-\varphi(x))$ where φ is a convex function). The fact that $\mathfrak{m}$ is continuous follows by the continuity of G and h.

Finally, with the help of Lemma 9 we demonstrate that some models discussed in Remark 3 meet the information theoretic barrier described in Proposition 2, and hence their support cannot be recovered by any algorithm unless Γ is large enough.

Example 3. Let $Y = G(h(\boldsymbol{X}^\intercal \boldsymbol{\beta}) + \varepsilon)$, where G, h are strictly monotone continuous functions and in addition h is an L-Lipschitz function. Furthermore let ε has a density as specified in Remark 3, i.e. $p_\varepsilon(x) \propto \exp(-P(x^2))$, where P is any nonzero polynomial with non-negative coefficients such that $P(0) = 0$. Then if $n < \frac{s\log(p-s+1)}{C}$ for some constant $C > 0$ (depending on P, G, h) and s sufficiently large, any algorithm recovering the support in our model will have errors with probability at least $\frac{1}{2}$ asymptotically.

Proof. Note that all moments of the random variable ε exist. Next we verify that condition (56) holds in this setup. Since G is 1-1 and KL divergence is invariant under changes of variables WLOG we can assume our model is simply $Y = h(\boldsymbol{X}^\intercal \boldsymbol{\beta}) + \varepsilon$ or in other words $f(u, \varepsilon) = h(u) + \varepsilon$. This is a location family for $u \in \mathbb{R}$ and thus the normalizing constant of the densities will stay the same regardless of the value of u. Direct calculation yields:

$$\begin{aligned} D(p(f(u,\varepsilon)) \| p(f(v,\varepsilon))) &= \mathbb{E}[P((\xi + h(u) - h(v))^2) - P(\xi^2)] \\ &= \widetilde{P}((h(u) - h(v))^2), \end{aligned}$$

where ξ has a density $p_\xi(x) \propto \exp(-P(x^2))$, and $\widetilde{P}$ is another nonzero polynomial with nonnegative coefficients, with $\widetilde{P}(0) = 0$ of the same degree as P. The last equality follows from the fact that all odd moments of ξ are 0, since ξ is a symmetric about 0 distribution. Since h is L-Lipschitz we conclude that:

$$D(p(f(u,\varepsilon)) \| p(f(v,\varepsilon))) \leq \widetilde{P}(L^2(u-v)^2).$$

The last can be clearly dominated by $\exp(C(u-v)^2) - 1$ for a large enough constant C. □

References

Amini, A. A. & Wainwright, M. J. (2008), High-dimensional analysis of semidefinite relaxations for sparse principal components, *in* 'Information Theory, 2008. ISIT 2008. IEEE International Symposium on', IEEE, pp. 2454–2458.

Berthet, Q. & Rigollet, P. (2013), Complexity theoretic lower bounds for sparse principal component detection, *in* 'Conference on Learning Theory', pp. 1046–1066.

Boucheron, S., Lugosi, G. & Massart, P. (2013), *Concentration inequalities: A nonasymptotic theory of independence*, OUP Oxford.

Brenner, N., Bialek, W. & Van Steveninck, R. d. R. (2000), 'Adaptive rescaling maximizes information transmission', *Neuron* **26**(3), 695–702.

Cai, T., Liu, W. & Luo, X. (2011), 'A constrained l1 minimization approach to sparse precision matrix estimation 1 minimization approach to sparse precision matrix estimation', *Journal of the American Statistical Association* **106**(494), 594–607.

Cai, T. T., Ma, Z., Wu, Y. et al. (2013), 'Sparse pca: Optimal rates and adaptive estimation', *The Annals of Statistics* **41**(6), 3074–3110.

Candes, E. & Tao, T. (2007), 'The dantzig selector: Statistical estimation when p is much larger than n', *The Annals of Statistics* pp. 2313–2351.

Chen, X., Xu, H., Yuan, P., Fang, F., Huss, M., Vega, V. B., Wong, E., Orlov, Y. L., Zhang, W., Jiang, J. et al. (2008), 'Integration of external signaling pathways with the core transcriptional network in embryonic stem cells', *Cell* **133**(6), 1106–1117.

Cloonan, N., Forrest, A. R., Kolle, G., Gardiner, B. B., Faulkner, G. J., Brown, M. K., Taylor, D. F., Steptoe, A. L., Wani, S., Bethel, G. et al. (2008), 'Stem cell transcriptome profiling via massive-scale mrna sequencing', *Nature methods* **5**(7), 613–619.

Cook, R. D. & Ni, L. (2005), 'Sufficient dimension reduction via inverse regression', *Journal of the American Statistical Association* **100**(470).

Cook, R. D. et al. (2004), 'Testing predictor contributions in sufficient dimension reduction', *The Annals of Statistics* **32**(3), 1062–1092.

Cover, T. M. & Thomas, J. A. (2012), *Elements of information theory*, John Wiley & Sons.

d'Aspremont, A., Bach, F. & Ghaoui, L. E. (2008), 'Optimal solutions for sparse principal component analysis', *The Journal of Machine Learning Research* **9**, 1269–1294.

d'Aspremont, A., El Ghaoui, L., Jordan, M. I. & Lanckriet, G. R. (2007), 'A direct formulation for sparse pca using semidefinite programming', *SIAM review* **49**(3), 434–448.

Duan, N. & Li, K.-C. (1991), 'Slicing regression: a link-free regression method', *The Annals of Statistics* pp. 505–530.

Hsing, T. & Carroll, R. J. (1992), 'An asymptotic theory for sliced inverse regression', *The Annals of Statistics* pp. 1040–1061.

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

Johnstone, I. M. & Lu, A. Y. (2004), 'Sparse principal components analysis', *Unpublished manuscript* .

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

Krauthgamer, R., Nadler, B. & Vilenchik, D. (2013), 'Do semidefinite relaxations really solve sparse pca?', *arXiv preprint arXiv:1306.3690* .

Laurent, B. & Massart, P. (2000), 'Adaptive estimation of a quadratic functional by model selection', *Annals of Statistics* pp. 1302–1338.

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

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

Lin, Q., Zhao, Z. & Liu, S. J. (2015), 'On consistency and sparsity of sliced inverse regression in high dimensions', *arXiv preprint arXiv:1507.03895* .

Massart, P. (1990), 'The tight constant in the dvoretzky-kiefer-wolfowitz inequality', *The Annals of Probability* pp. 1269–1283.

Ouyang, Z., Zhou, Q. & Wong, W. H. (2009), 'Chip-seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells', *Proceedings of the National Academy of Sciences* **106**(51), 21521–21526.

Paul, D. (2007), 'Asymptotics of sample eigenstructure for a large dimensional spiked covariance model', *Statistica Sinica* **17**(4), 1617.

Price, A. L., Zaitlen, N. A., Reich, D. & Patterson, N. (2010), 'New approaches to population stratification in genome-wide association studies', *Nature Reviews Genetics* **11**(7), 459–463.

Tibshirani, R. et al. (1997), 'The lasso method for variable selection in the cox model', *Statistics in medicine* **16**(4), 385–395.

Van Der Vaart, A. W. & Wellner, J. A. (1996), *Weak Convergence*, Springer.

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

Vu, V. Q. & Lei, J. (2012), 'Minimax rates of estimation for sparse pca in high dimensions', *arXiv preprint arXiv:1202.0786* .

Wainwright, M. J. (2009), 'Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting', *Information Theory, IEEE Transactions on* **55**(12), 5728–5741.

Yang, Y. & Barron, A. (1999), 'Information-theoretic determination of minimax rates of convergence', *Annals of Statistics* pp. 1564–1599.

Yu, B. (1997), Assouad, fano, and le cam, *in* 'Festschrift for Lucien Le Cam', Springer, pp. 423–435.

Yu, Z., Zhu, L., Peng, H. & Zhu, L. (2013), 'Dimension reduction and predictor selection in semiparametric models', *Biometrika* p. ast005.

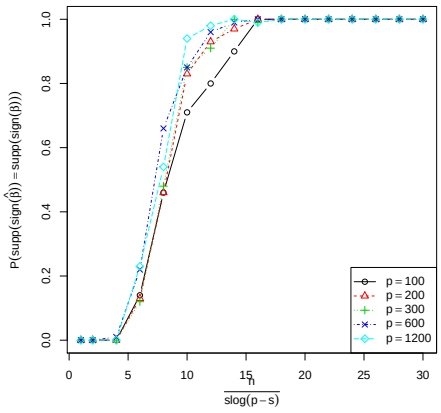
Zhang, Y. & Ghaoui, L. E. (2011), Large-scale sparse principal component analysis with application to text data, *in* 'Advances in Neural Information Processing Systems', pp. 532–539.

Zhong, W., Zhang, T., Zhu, Y. & Liu, J. S. (2012), 'Correlation pursuit: forward stepwise variable selection for index models', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **74**(5), 849–870.

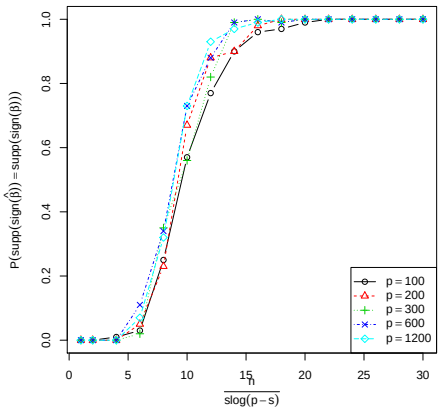
Zhu, L., Miao, B. & Peng, H. (2006), 'On sliced inverse regression with high-dimensional covariates', *Journal of the American Statistical Association* **101**(474).

C Extra Numerical Studies Figures

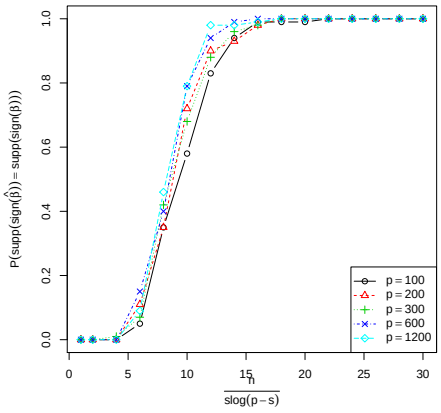
Figure 4: Efficiency Curves for DT-SIR, $s = \log p$, $\Gamma = \frac{n}{s\log(p-s)} \in [0, 30]$



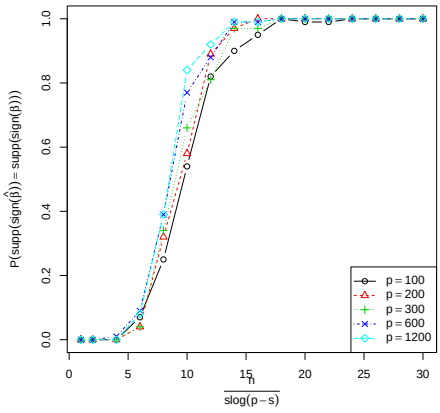
(a) Model (13)



(b) Model (14)

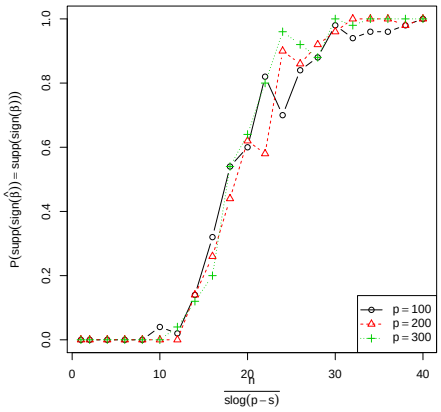


(c) Model (15)

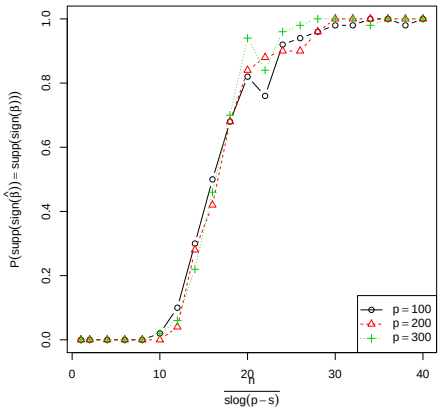


(d) Model (16)

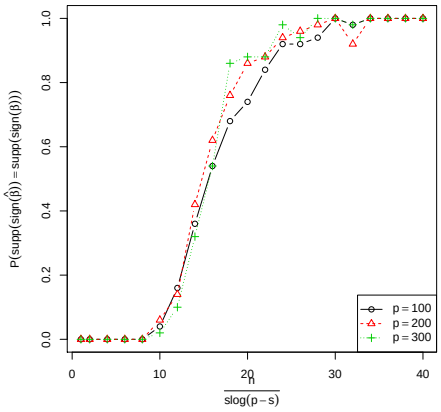
Figure 5: Efficiency Curves for SDP, $s = \log p,\ \Gamma = \frac{n}{s\log(p-s)} \in [0, 40]$



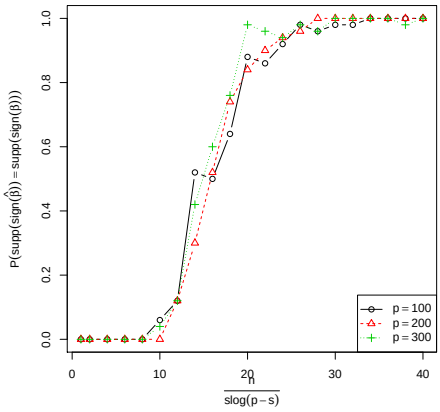
(a) Model (13)



(b) Model (14)

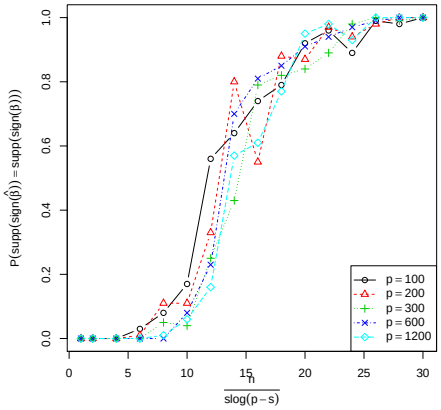


(c) Model (15)

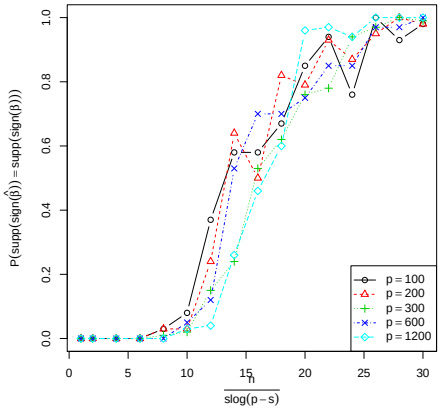


(d) Model (16)

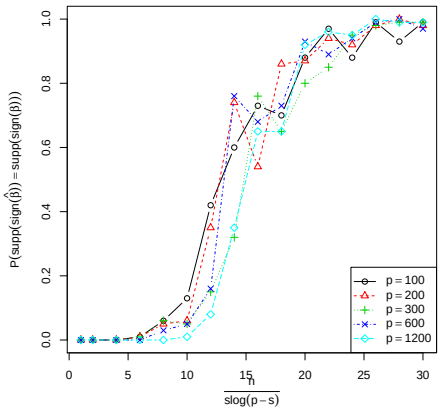
Figure 6: Efficiency Curves for DT-SIR, $s = \sqrt{p}$, $\Gamma = \frac{n}{s\log(p-s)} \in [0, 30]$, $\boldsymbol{\beta} = \widetilde{\boldsymbol{\beta}}/\|\widetilde{\boldsymbol{\beta}}\|_2$ according to (17)



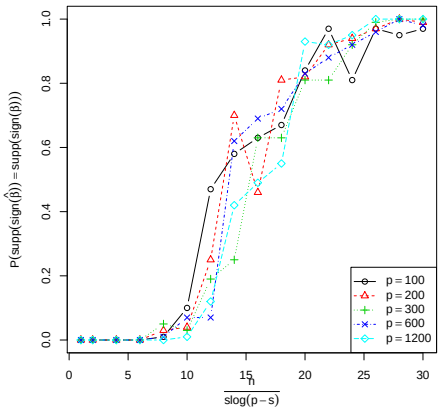
(a) Model (13)



(b) Model (14)

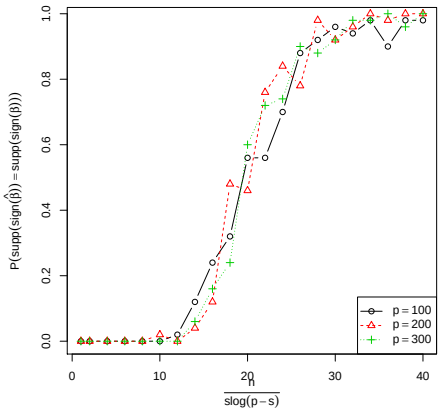


(c) Model (15)

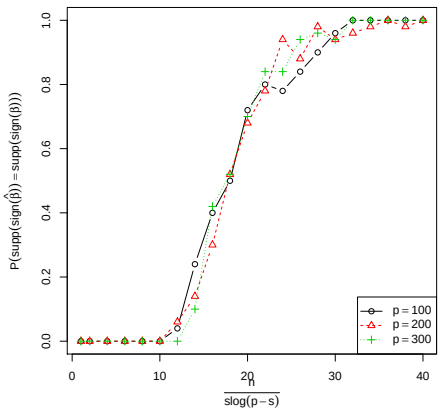


(d) Model (16)

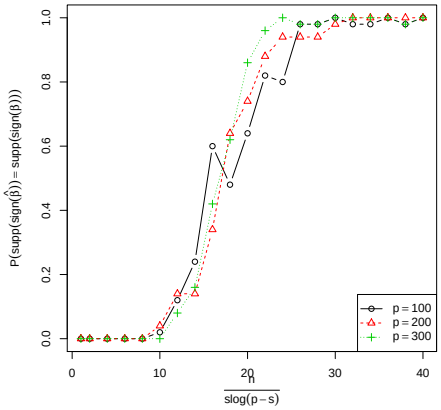
Figure 7: Efficiency Curves for SDP, $s = \sqrt{p}$, $\Gamma = \frac{n}{s\log(p-s)} \in [0, 40]$, $\boldsymbol{\beta} = \widetilde{\boldsymbol{\beta}}/\|\widetilde{\boldsymbol{\beta}}\|_2$ according to (17)



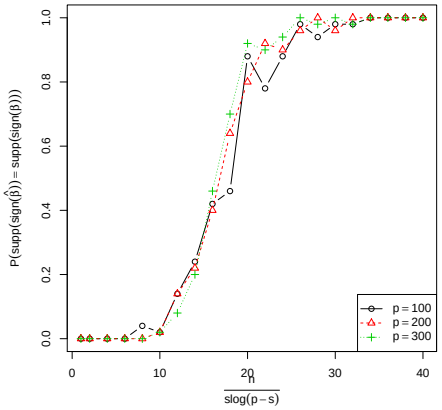
(a) Model (13)



(b) Model (14)



(c) Model (15)



(d) Model (16)