On Stein's Identity and Near-Optimal Estimation in High-dimensional Index Models

Zhuoran Yang*[1], Krishnakumar Balasubramanian†[1], and Han Liu‡[1]

[1]Department of Operations Research and Financial Engineering, Princeton University

September 27, 2017

Abstract

We consider estimating the parametric components of semi-parametric multiple index models in a high-dimensional non-Gaussian setting. Our estimators leverage the score function based second-order Stein's lemma and do not require Gaussian or elliptical symmetry assumptions made in the literature. Moreover, to handle score functions and response variables that are heavy-tailed, our estimators are constructed via carefully thresholding their empirical counterparts. We show that our estimator achieves near-optimal statistical rate of convergence in several settings. We supplement our theoretical results via simulation experiments that confirm the theory.

*zy6@princeton.edu

†kb18@princeton.edu

‡hanliu@princeton.edu

1 Introduction

Consider the semi-parametric index model relating the response (Y) and the covariate (X) by

$$Y = f\left(\langle \beta_1^*, X\rangle, \ldots, \langle \beta_k^*, X\rangle\right) + \epsilon, \tag{1.1}$$

where $X, \{\beta_\ell^*\}_{\ell\in[k]} \in \mathbb{R}^d$ and $\{\beta_\ell^*\}_{\ell\in[k]}$ are assumed to be structured and ϵ is a zero-mean noise that is independent of X. Here the vectors $\{\beta_\ell^*\}_{\ell\in[k]}$ are the parametric components and the function f is the nonparametric component or the link function. Such a model is called as sparse multiple index model (MIM) in the literature (Chen et al., 2010). In this work, given n i.i.d samples $\{X_i, Y_i\}_{i=1}^n$ from the above model, where $n < d$, we are concerned with estimating the parametric components $\{\beta_\ell^*\}_{\ell\in[k]}$ when f is unknown. More importantly, we do not impose the assumption that X is Gaussian, which is commonly made in the literature. Two special cases of our model include phase retrieval, for which $k = 1$, and dimensionality reduction, for which $k \geq 1$. Motivated by these applications, we make a distinction between the case of $k = 1$ (which is also called as single index model (SIM)) and $k > 1$ in the rest of the paper.

Furthermore, note that when $k = 1$ and f is the identity function, we obtain the standard linear regression model. Estimators for high-dimensional linear models have been developed and analyzed extensively in the last decade (see for example (Bühlmann and van de Geer, 2011; Vershynin, 2015) for comprehensive overviews). While being a useful testbed for illustrating conceptual phenomenon, they often suffer from a lack of flexibility in modeling real-world situations. On the other hand, completely nonparametric models, although flexible, suffer from the curse of dimensionality unless restrictive additive sparsity or smoothness assumptions are imposed (Yuan et al., 2016). An interesting compromise between the parametric and nonparametric models is provided by the semi-parametric index models (Horowitz, 2009) described in (1.1).

Estimating the parametric components $\{\beta_\ell^*\}_{\ell\in[k]}$ without the dependence on the exact form of the link function appears naturally in several situations. For example, in one-bit compressed sensing (Boufounos and Baraniuk, 2008) and sparse generalized linear models (Loh and Wainwright, 2015), we are interested in recovering the true signal vector based on nonlinear measurements. In sufficient dimensionality reduction, where k is typically a fixed number greater than one but much less than d, we would like to estimate the projection

onto the subspace spanned by the parametric components $\{\beta_\ell^*\}_{\ell\in[k]}$ without depending on the specific form of the function f. Furthermore, in deep neural networks (DNN), which are cascades of the MIM, the nonparametric component corresponds to the activation function which is pre-specified and the task is to estimate the linear components, which are used for prediction in the test stage. Hence, it is crucial to develop estimators for the linear component with both statistical accuracy and computational efficiency for a wide class of link functions.

1.1 Intriguing Aspects of Index Models.

Several subtle issues arise when we consider optimal estimation in SIM and MIM. Specifically, most existing results depend crucially on the assumption made on X or f and fail to hold when those assumptions are relaxed. Such issues arise even in low-dimensional settings, where $n > d$. Consider, for example, the case of $k = 1$ and a known link function $f(u) = u^2$. This corresponds to phase retrieval, which is a challenging inverse problem that has regained interest in the last few years along with the success of compressed sensing. A straightforward way to estimate β^* is to do nonlinear least squares regression (Lecué and Mendelson, 2015), which is a nonconvex optimization problem. Candès et al. (2013) propose an estimator based on convex relaxations. Although their estimator is optimal when X is sub-Gaussian, they are not agnostic to the link function., i.e., the same result does not hold if the link function is changed. Direct optimization of the nonconvex phase retrieval problem was considered by Candès et al. (2015) and Sun et al. (2016), which propose estimators based on iterative algorithms that are statistical optimal. However, they rely on the assumption that X is Gaussian. A careful look at their proofs reveal that extending them to a wider class of distributions is significantly challenging – for example, they require sharp concentration inequalities for polynomials of degree four of X, which would lead to suboptimal rate even when X is sub-Gaussian. Furthermore, their results are not agnostic to the link function as well. Similar observations could be made for both convex (Li and Voroninski, 2013) and nonconvex estimators (Cai et al., 2015) for sparse phase retrieval in high dimensions.

In addition, a surprising result for SIM was established in Plan and Vershynin (2016). They show that when X is Gaussian, even when the link function is unknown, one could estimate β^* at the optimal statistical rate with the convex Lasso estimator. Unfortunately,

their assumptions on the link function is rather restrictive and rule out several interesting models including phase retrieval. Furthermore, none of the above procedures are applicable to the case of MIMs. A line of work pioneered by Ker-Chau Li, focused on estimation in MIMs in the low-dimensional setting. We provide a discuss about them in the related work section, but they again require restrictive assumption on either the link functions or on the distribution of X. For example, X is required to be elliptically symmetric, which prevents it from being widely applicable.

1.2 Motivation

Our work is primarily motivated by an interesting phenomenon illustrated in (Plan and Vershynin, 2016) for a class of high-dimensional SIM. Below, we first briefly summarize the result from (Plan and Vershynin, 2016) and then provide our *alternative justification* for the same result via Stein's identity. We mainly leverage this alternative justification and propose our estimators for the more general setting we consider. Assuming, for simplicity, we work in the one-dimensional setting and are given n i.i.d. samples from the SIM. Consider the least-squares estimator

$$\widehat{\beta}_{LS} = \underset{\beta \in \mathbb{R}}{\operatorname{argmin}} \ \frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i \beta)^2 .$$

Note that the above estimator is the standard least-squares estimator assuming a linear model (i.e., identity link function). The surprising observation from (Plan and Vershynin, 2016) is that, under the *crucial* assumption that X is standard Gaussian, $\widehat{\beta}_{LS}$ is a good estimator of β^* (up to a scaling) even when the data is generated from a nonlinear SIM. The same holds true for the high-dimensional setting when the minimization is performed in an appropriately constrained norm-ball (for example, the ℓ_1-ball). Hence the theory developed for the linear setting could be leveraged to understand the performance in the SIM setting. Below, we give an alternative justification for the above estimator as an implication of Stein's identity in the Gaussian case, which is summarized as follows.

Proposition 1.1 (Gaussian Stein's Identity (Stein, 1972))**.** *Let $X \sim N(0,1)$ and $g : \mathbb{R} \to \mathbb{R}$ be a continuos function such that $\mathbb{E}|g'(X)| \leq \infty$. Then we have $\mathbb{E}[g(X)X] = \mathbb{E}[g'(X)]$.*

Note that in our context, we have $\mathbb{E}[f'(X)] \propto \beta^*$ and $\mathbb{E}[f(X)X] = \mathbb{E}[Y \cdot X]$. Now consider the following estimator, which is based on performing least-squares on the sample version of

the above proposition:

$$\widehat{\beta}_{SL} = \underset{\beta \in \mathbb{R}}{\operatorname{argmin}} \ \frac{1}{n} \sum_{i=1}^{n} (Y_i X_i - \beta)^2$$

Note that $\widehat{\beta}_{LS}$ and $\widehat{\beta}_{SL}$ are the same estimators assuming $X \sim N(0,1)$, as $n \to \infty$. This observation leads to an alternative interpretation of the estimator proposed by (Plan and Vershynin, 2016) via Stein's identity for Gaussian random variables. Thus it provides an alternative justification for why the linear least-squares estimator should work in the SIM setting. Interestingly, a similar procedure based on second-order Stein's identity (see §2 for precise definitions) was used in Candès et al. (2015) to provide a favorable initializer for their gradient descent algorithm for phase retrieval. Our observation also provides an alternative interpretation of the initialization method used in Candès et al. (2015) by appealing to Stein's identity. These observations also naturally leads to leveraging non-Gaussian versions of Stein's identity for dealing with non-Gaussian covariates. Our estimators based on this motivation is described in detail in §3 and §4.

1.3 Related Work

There is a significant body of work on SIMs in the low-dimensional setting. They are based on assumptions on either the distribution of the covariate or the link functions. Assuming a monotonic link function, Han (1987); Sherman (1993) propose the maximum rank correlation estimator exploiting the relationship between monotonic functions and rank-correlations. Furthermore, Li and Duan (1989) propose an estimator for a wide class of unknown link functions under the assumption that the covariate follows a symmetric elliptical distribution. This assumption is restrictive as often times the covariates are not from a symmetric distribution. For example, in several economic applications where the covariates are usually highly skewed and heavy-tailed (Horowitz, 2009).

The success of Lasso and related linear estimators in high-dimensions (Bühlmann and van de Geer, 2011), also enabled the exploration of high-dimensional SIMs. Although, this is very much work in progress. As mentioned previously, Plan and Vershynin (2016) show that the Lasso estimator works for the SIMs in high dimensions when the data is Gaussian. A more tighter albeit an asymptotic results under the same setting was proved in Thrampoulidis et al. (2015). Very recently Goldstein et al. (2016) extend the results of Li and Duan

(1989) to the high dimensional setting but it suffers from similar problems as mentioned in the low-dimensional setting. For the case of monotone nonparametric component, Yang et al. (2015) analyze a non-convex least squares approach under the assumption that the data is sub-Gaussian. However, the success of their method hinges on the knowledge of the link function. Furthermore, Jiang and Liu (2014); Lin et al. (2015); Zhu et al. (2006) analyze the sliced inverse regression estimator in the high-dimensional setting concentrating mainly on support recovery and consistency properties. Similar to the low-dimensional case, the assumptions made on the covariate distribution restrict them from several real-world applications involving non-Gaussian or non-symmetric covariate, for example high-dimensional problems in economics (Fan et al., 2011). Furthermore, several results are established on a case-by-case basis for fixed link function. Specifically Boufounos and Baraniuk (2008); Ai et al. (2014) and Davenport et al. (2014) consider 1-bit compressed sensing and matrix completion respectively, where the link is assumed to be the sign function. Also, Waldspurger et al. (2015) and Cai et al. (2015) propose and analyze convex and non-convex estimators for phase retrieval respectively, in which the link is the square function. All the above works, except Ai et al. (2014) make Gaussian assumptions on the data and are specialized for the specific link functions. The non-asymptotic result obtained in Ai et al. (2014) is under sub-Gaussian assumptions, but the estimator is not consistent. Finally, there is a line of work focussing on estimating both the parametric and the nonparametric component Kalai and Sastry (2009); Kakade et al. (2011); Alquier and Biau (2013); Radchenko (2015). We do not focus on this situation in this paper as mentioned before.

For multiple index models, relatively less work exist in the high-dimensional setting. In the low-dimensional setting, a line of work for estimation in MIMs is proposed by Ker-Chau Li, which include inverse regression (Li, 1991), principal Hessian directions (Li, 1992) and regression under link violation (Li and Duan, 1989). The proposed estimators are applicable for a class of unknown link functions under the assumption that the covariate follows a Gaussian or symmetric elliptical distribution. Such an assumption is restrictive as often times the covariates are heavy-tailed or skewed (Horowitz, 2009; Fan et al., 2011). Furthermore, they concentrate only on the low-dimensional setting establishing asymptotic statements. Estimation in high-dimensional MIM under the subspace sparsity assumption was considered in Chen et al. (2010), where the results are asymptotic and the proposed

estimators are not computable in polynomial time.

To summarize, all the above works require restrictive assumption on either the data distribution or on the link function. We propose and analyze an estimator for a class of (unknown) link functions for the case when the covariates are drawn from a non-Gaussian distribution – under the assumption that we know the distribution *a priori*. Note that in several situations, one could fit specialized distributions, to real-world data that is often times skewed and heavy-tailed, so that it provides a good generative model of the data. Also, mixture of Gaussian distribution, with the number of components selected appropriately, approximates the set of all square integrable distributions to arbitrary accuracy (see for example McLachlan and Peel (2004)). Furthermore, since this is a density estimation problem it is unlabeled and there is no issue of label scarcity. Hence it is possible to get accurate estimate of the distribution in most situations of interest. Thus our work is complementary to the existing literature and provides an estimator for a class of models that is not addressed in the previous works.

1.4 Contributions

As we saw before, there are several subtleties based on the interplay between the assumptions made on X and f when dealing with estimation in SIM and MIM. Thus an interesting question is, whether it is possible to estimate the linear components in SIMs and MIMs with milder assumptions on both X and f in the high-dimensional setting. In this work, we provide a partial answer to this question. We construct estimators that work for a wide class of link functions, including the phase retrieval link function, and for a large family of distributions of X, which is assumed to be known *a priori*. We particularly focus on the case when X follows a non-Gaussian distribution that need not be elliptically symmetric or sub-Gaussian, thus making our method applicable to several situations not possible before. Our estimators are based on Stein's identity for non-Gaussian distributions, which utilizes the score function. Estimating with the score function is challenging due to their heavy tails. In order to illustrate that, consider the univariate histograms provided in Figure-1. The dark shaded, more concentrated one corresponds to the histogram of 10000 samples from Gamma distribution with shape and scale parameters set to 5 and 0.2 respectively. The transparent histogram corresponds to the distribution of the score function of the same Gamma dis-

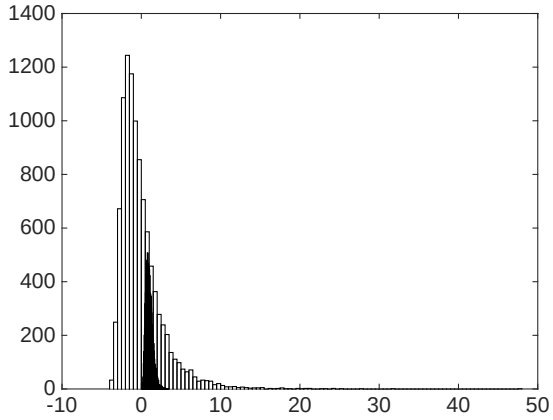


Figure 1: Histogram of Score Function based on 10000 independent samples from the Gamma distribution with shape 5 and scale 0.2. The dark histogram (we recommend the reader to zoom in to notice it) concentrated around zero corresponds to the Gamma distribution and the transparent histogram corresponds to the distribution of the score of the same Gamma distribution.

tribution. Note that even when the actual Gamma distribution is well concentrated, the distribution is the corresponding score function is well-spread and heavy-tailed. In the high dimensional setting, in order to estimate with the score functions, we require certain vectors or matrices based on the score functions to be well-concentrated in appropriate norms. In order to achieve that, we construct robust estimators via careful truncation arguments to balance the bias (due to thresholding)-variance (of the estimator) tradeoff and achieve the required concentration.

- We construct estimators for the parametric component of a sparse SIM and MIM for a class of unknown link function under the assumption that the covariate distribution is non-Gaussian but known a priori.

- We establish near-optimal statistical rates for our estimators. Our results complement the existing ones in the literature and hold in several case not possible before.

- We provide alternative justifications based on the Stein's identity for the estimator used in Plan and Vershynin (2016) for sparse SIM and the initializer used in Candès et al. (2015) for phase retrieval.

- As a consequence of our results for SIM and MIM, we also obtain a near-optimal estimator for sparse PCA with heavy-tailed data in the moderate sample size regime.
- We provide numerical simulations that confirm our theoretical results.

1.5 Notations

In this section, we introduce the notation and define the single index models. Throughout this work, we use $[n]$ to denote the set $\{1, \dots, n\}$. In addition, for a vector $v \in \mathbb{R}^d$, we denote by $\|v\|_p$ the ℓ_p-norm of v for any $p \geq 1$. We use $\mathcal{S}^{d-1}$ to denote the unit sphere in $\mathbb{R}^d$, which is defined as $\mathcal{S}^{d-1} = \{v \in \mathbb{R}^d \colon \|v\|_2 = 1\}$. In addition, we define the support of $v \in \mathbb{R}^d$ as $\mathrm{supp}(v) = \{j \in [d], v_j \neq 0\}$. Moreover, we denote the nuclear norm, operator norm, element-wise max norm and Frobenius norm of a matrix $A \in \mathbb{R}^{d_1 \times d_2}$ by $\|\cdot\|_\star$, $\|\cdot\|_{\mathrm{op}}$, $\|\cdot\|_\infty$ and $\|\cdot\|_{\mathrm{fro}}$, respectively. We denote by $\mathrm{vec}(A)$ the vectorization of matrix A, which is a vector in $\mathbb{R}^{d_1 \cdot d_2}$. For two matrices $A, B \in \mathbb{R}^{d_1 \times d_2}$ we define the trace inner product as $\langle A, B \rangle = \textsc{trace}(A^\top B)$. Note that it can be viewed as the standard inner product between $\mathrm{vec}(A)$ and $\mathrm{vec}(B)$. In addition, for an univariate function $g \colon \mathbb{R} \to \mathbb{R}$, we denote by $g \circ (v)$ and $g \circ (A)$ the output of applying g to each element of a vector v and a matrix A, respectively. Finally, for a random variable $X \in \mathbb{R}$ with density p, we use $p^{\otimes d} \colon \mathbb{R}^d \to \mathbb{R}$ to denote the joint density of $\{X_1, \cdots, X_d\}$, which are d identical copies of X. We also require some notations about tensors. We concentrate on fourth-order tensors for simplicity. For any fourth-order tensor $Z \in \mathbb{R}^{d_1 \times d_2 \times d_3 \times d_4}$, we denote its (j_1, j_2, j_3, j_4)-th entry by $Z(j_1, j_2, j_3, j_4)$. If $d_\ell = d$ for all $\ell \in [4]$, we denote the tensor as $Z \in \mathbb{R}^{d^{\otimes 4}}$. Similar to the matrix case, we define $\mathrm{vec}(Z) \in \mathbb{R}^{d^4}$ as the vectorization of the tensor Z. For two tensors $W, Z \in \mathbb{R}^{d^{\otimes 4}}$, we define their inner inner product as

$$\langle Z, W \rangle = \mathrm{vec}(Z)^\top \mathrm{vec}(W) = \sum_{j_1, j_2, j_3, j_4 \in [d]} Z(j_1, j_2, j_3, j_4) \cdot W(j_1, j_2, j_3, j_4) \tag{1.2}$$

The tensor Frobenius norm of is also denoted by $\|\cdot\|_{\mathrm{fro}}$.

2 Index Models

Now we are ready to define the precise statistical models that we consider in this work. As mentioned above, we consider the case of $k = 1$ (SIM) and $k > 1$ (MIM) separately. We

primarily distinguish our models based on the assumption made on the link functions. We also require the following definition of score function of random variable. Let $p\colon \mathbb{R}^d \to \mathbb{R}$ be a probability density function defined on $\mathbb{R}^d$. The score function $S_p\colon \mathbb{R}^d \to \mathbb{R}$ associated to p is defined as

$$S_p(x) = -\nabla_x[\log p(x)] = -\nabla_x p(x)/p(x).$$

Note that in the above definition, the derivative is taken with respect to x. This is different from the more traditional definition of the score function where the density belongs to a parametrized family and the derivative is taken with respect to the parameters. In the rest of the paper to simplify the notation, we omit the subscript x from ∇_x. We also omit the subscript p from S_p when the underlying density p is clear from the context.

2.1 First-order Link Functions

We first discuss a class of SIM that are based on a certain first-order link functions. We discuss the motivation for our estimator, which automatically highlights the first-order assumption on the link function as well. Recall that our estimators are based on Stein's identity. To begin with, we present the first-order non-Gaussian Stein's identity.

Proposition 2.1 (First-order Stein's Identity (Stein et al., 2004)). *Let $X \in \mathbb{R}^d$ be a real-valued random vector with density p. Assume that $p\colon \mathbb{R}^d \to \mathbb{R}$ is differentiable. In addition, let $g : \mathbb{R}^d \to \mathbb{R}$ be a continous function such that $\mathbb{E}[\nabla g(X)]$ exists. Then it holds that*

$$\mathbb{E}[g(X) \cdot S(X)] = \mathbb{E}[\nabla g(X)],$$

where $S(x) = -\nabla p(x)/p(x)$ is the score function of p.

One could apply the above Stein's Identity to SIMs to obtain an estimate of β^*. To see this, note that when $X \sim N(0, I_d)$ we have $S(x) = x$, $\forall x \in \mathbb{R}^d$. In this case, as $\mathbb{E}(\epsilon) = 0$, we have

$$\mathbb{E}(Y \cdot X) = \mathbb{E}[f(\langle X, \beta^* \rangle) \cdot X] = \mathbb{E}[f'(\langle X, \beta^* \rangle)] \cdot \beta^*.$$

Hence one could estimate β^* based on estimating the moment $\mathbb{E}(Y \cdot X)$. This observation leads to the estimator proposed in Plan and Vershynin (2016). This motivates the following definition of SIM with first order link functions.

Definition 2.2 (Vector SIM with First-order Links)**.** *Under this model, we assume that the response variable $Y \in \mathbb{R}$ and the covariate $X \in \mathbb{R}^d$ are linked via*

$$Y = f(\langle X, \beta^* \rangle) + \epsilon, \tag{2.1}$$

where $f \colon \mathbb{R} \to \mathbb{R}$ is an unknown univariate function, $\beta^* \in \mathbb{R}^d$ is the parameter of interest, and $\epsilon \in \mathbb{R}$ is the exogenous random noise such that $\mathbb{E}(\epsilon) = 0$. In addition, we assume that the entries of X are i.i.d. random variables with density p_0 and that β^* is s^*-sparse, that is, β^* contains only s^* nonzero entries such that $s^* \ll n \ll d$. Moreover, since the norm of β^* can be absorbed in f, we further let $\|\beta^*\|_2 = 1$ for identifiability. Finally, we assume f and X are such that $\mathbb{E}[f'(\langle X, \beta^* \rangle)] \neq 0$.

Note that the SIM depends only on covariate only via inner products. Hence it is natural to generalize it to the case of matrix and tensor valued covariates. To enable estimation in a high-dimensional setting, we enforce low-rank constraints that we describe below.

Definition 2.3 (Matrix SIM with First-order Links)**.** *For the low-rank case SIM, we assume that $\beta^* \in \mathbb{R}^{d_1 \times d_2}$ has rank $r^* \ll \min\{d_1, d_2\}$. In this scenario, $X \in \mathbb{R}^{d_1 \times d_2}$ and the inner product in (2.1) is $\langle X, \beta^* \rangle = \textsc{trace}(X^\top \beta^*)$. For model identifiability, we further assume that $\|\beta^*\|_{fro} = 1$, similar to the sparse case. Finally, we assume f and X are such that $\mathbb{E}[f'(\langle X, \beta^* \rangle)] \neq 0$.*

Before we lay out the first-order low-rank tensor single index model, we first introduce additional notation for tensors. Denote by $u \otimes v \otimes s \otimes t \in \mathbb{R}^{d^{\otimes 4}}$ a rank-one tensor. The minimum value of r such that the tensor Z could be written as a summation of r rank-one tensors, i.e., $Z = \sum_{j=1}^{r} u_j \otimes v_j \otimes s_j \otimes t_j$, is called as the CP-rank of the tensor, denoted by $\text{rank}_{CP}(Z) = r$. We now describe the low-rank tensor model that we consider in this work.

Definition 2.4 (Tensor SIM with First-order Links)**.** *For the low-rank Tensor SIM model, we assume that $\beta^* \in \mathbb{R}^{d^{\otimes 4}}$ and has CP-rank, $\text{rank}_{CP}(\beta^*) = r^*$. In this scenario, $X \in \mathbb{R}^{d^{\otimes 4}}$ and the inner product in (2.1) is understood as defined in (1.2). For model identifiability, we further assume that $\|\beta^*\|_{fro} = 1$, similar to the matrix case. Finally, we assume f and X are such that $\mathbb{E}[f'(\langle X, \beta^* \rangle)] \neq 0$.*

2.2 Second-order Link Functions

In the above models, it is crucial that $\mathbb{E}[f'(\langle X, \beta^* \rangle)] \neq 0$, for it to work. Such a restriction prevents it from being applicable to some widely used cases of SIM, for example, phase retrieval where f is the quadratic function. This limitation of the first order Stein's identity, motivates us to examine the second order Stein's identity which is summarized below.

Proposition 2.5 (Second-order Stein's Identity (Janzamin et al., 2014))**.** *Assume that the density of X is twice differentiable. In addition, define the second-order score function $T\colon \mathbb{R}^d \to \mathbb{R}^{d\times d}$ as $T(x) = \nabla^2 p(x)/p(x)$. Then for any twice differentiable function $g : \mathbb{R}^d \to \mathbb{R}$ such that $\mathbb{E}[\nabla^2 g(X)]$ exists, we have*

$$\mathbb{E}\big[g(X) \cdot T(X)\big] = \mathbb{E}\big[\nabla^2 g(X)\big]. \tag{2.2}$$

Back to the phase retrieval example, when $X \sim N(0, I_d)$, the second order score function now becomes $T(x) = xx^\top - I_d$, $\forall x \in \mathbb{R}^d$. Setting $g(x) = \langle x, \beta^* \rangle^2$ in (2.2), we have

$$\mathbb{E}[g(X) \cdot T(X)] = \mathbb{E}[g(X) \cdot (XX^\top - I)] = \mathbb{E}[\langle X, \beta^* \rangle^2 \cdot (XX^\top - I)] = 2\beta^* {\beta^*}^\top. \tag{2.3}$$

Thus for phase retrieval, one could extract $\pm\beta^*$ based on second order Stein's identity even in the situation where the first order Stein's identity fails. Indeed, (2.3) used in Candès et al. (2015) *implicitly* to provide a spectral initialization for the Wirtinger flow algorithm in the case of Gaussian phase retrieval. Here, we provided an alternative justification based on Stein's identity, for why such an initializer works. Motivated by the this observation, we propose to use the second order Stein's identity to estimate the parametric component of SIMs and MIMs with a class of unknown link functions with non-Gaussian covariates. The precise statistical models that we consider are defined as follows.

Definition 2.6 (Vector SIM with Second-order Links)**.** *Under this model, we assume that the response variable $Y \in \mathbb{R}$ and the covariate $X \in \mathbb{R}^d$ are linked via*

$$Y = f(\langle X, \beta^* \rangle) + \epsilon, \tag{2.4}$$

where $f\colon \mathbb{R} \to \mathbb{R}$ is an unknown univariate function, $\beta^* \in \mathbb{R}^d$ is the parameter of interest, and $\epsilon \in \mathbb{R}$ is the exogenous random noise such that $\mathbb{E}(\epsilon) = 0$. In addition, we assume that the entries of X are i.i.d. random variables with density p_0 and that β^* is s^*-sparse, that

is, β^* contains only s^* nonzero entries. Moreover, since the norm of β^* can be absorbed in f, we further let $\|\beta^*\|_2 = 1$ for identifiability. Finally, we assume f and X are such that $\mathbb{E}[f''(\langle X, \beta^* \rangle)] > 0$.

Note that in the definition of the SIMs, we require that $\mathbb{E}[f''(\langle X, \beta^* \rangle)]$ positive. Since if $\mathbb{E}[f''(\langle X, \beta^* \rangle)]$ is negative , we could always replace f by $-f$ by flipping the sign of Y, we essentially assume that $\mathbb{E}[f''(\langle X, \beta^* \rangle)]$ is nonzero. Intuitively, such restriction on f implies that the second order moments contains the information of β^*, thus we call such a function the second order link. Similar to the first-order case, one could define matrix and tensor versions of the second-order SIMs but we do not concentrate on such models in this work. Thus far, we considered SIMs. We now define a class of MIMs with second order links. For MIMs the notion of first order link functions is naturally not sufficient to estimate the projector onto the subspace.

Definition 2.7 (MIM with Second-order Links)**.** *Under this model, we assume that the response variable $Y \in \mathbb{R}$ and the covariate $X \in \mathbb{R}^d$ are linked via*

$$Y = f\left(\langle X, \beta_1^* \rangle, \ldots, \langle X, \beta_k^* \rangle\right) + \epsilon, \tag{2.5}$$

where $f \colon \mathbb{R}^k \to \mathbb{R}$ is an unknown function, $\{\beta_\ell^*\}_{\ell \in [k]} \subseteq \mathbb{R}^d$ are the parameters of interest, and $\epsilon \in \mathbb{R}$ is the exogenous random noise such that $\mathbb{E}(\epsilon) = 0$. In addition, we assume that the entries of X are i.i.d. random variables with density p_0 and that $\{\beta_\ell^*\}_{\ell \in [k]}$ span a k-dimensional subspace of $\mathbb{R}^d$. Moreover, we denote $B^* = (\beta_1^* \ldots \beta_k^*) \in \mathbb{R}^{d \times k}$. Then the model in (2.5) can be written as $Y = f(XB^*) + \epsilon$. By the QR-factorization, we can write B^* as Q^*R^*, where $Q^* \in \mathbb{R}^{d \times k}$ is an orthonormal matrix and $R^* \in \mathbb{R}^{k \times k}$ is invertible. Since f is unknown, R^* can be absorbed into the link function. Thus, we assume that B^* is orthonormal for identifiability. Furthermore, we further assume that B^* is s^*-row sparse, that is, B^* contains only s^* nonzero rows. We note that such a definition of sparsity for B^* does not depends on the choice of coordinate system. Finally, we assume f and X are such that $\lambda_{\min}\left(\mathbb{E}[\nabla^2 f(XB^*)]\right) > 0$.

The assumption $\mathbb{E}[\nabla^2 f(XB^*)]$ is positive definite, in MIM, is a multivariate generalization of the condition that $\mathbb{E}[f''(\langle X, \beta^* \rangle)] > 0$ in SIM. It essentially guarantees that estimation of the projector onto the subspace spanned by the k components is well-defined. We now

introduce our estimators and provide theoretical results that are near-optimal in several settings.

3 Theoretical Results for Index Models with First-order Links

Recall that in the single index models introduced in §2.1, X in (2.1) has i.i.d. entries with density p_0. To unify the vector, matrix and tensor settings, we identify X with $\mathrm{vec}(X) \in \mathbb{R}^d$ where $d = d_1 \cdot d_2 \cdot d_3 \cdot d_4$. In this case, X has density $p = p_0^{\otimes d}$ and the corresponding score function $S \colon \mathbb{R}^d \to \mathbb{R}^d$ is given by

$$S(x) = -\nabla \log p(x) = -\nabla p(x)/p(x) = s_0 \circ (x), \tag{3.1}$$

where the univariate function $s_0 = p_0'/p_0$ is applied to each entry of x. Thus $S(X)$ has i.i.d. entries. In addition, by Lemma 2.1, we have $\mathbb{E}[S(X)] = 0$ by setting g to be a constant function. Moreover, in the context of SIMs specified in (2.1), we have

$$\mathbb{E}[Y \cdot S(X)] = \mathbb{E}\big[f(\langle X, \beta^* \rangle) \cdot S(X)\big] = \mathbb{E}[f'(\langle X, \beta^* \rangle)] \cdot \beta^*,$$

as long as the density and the link function satisfy the conditions stated in Lemma 2.1. This implies that optimization problem

$$\underset{\beta \in \mathbb{R}^d}{\text{minimize}} \big\{ \langle \beta, \beta \rangle - 2\mathbb{E}[Y \cdot \langle S(X), \beta \rangle] \big\} \tag{3.2}$$

has solution $\beta = \mu \cdot \beta^*$, where $\mu = \mathbb{E}[f'(\langle X, \beta^* \rangle)]$. Hence the above program could be used to obtain the unknown β^* as long as $\mu \neq 0$. Before we proceed to describe the sample version of the above program, we make the following brief remark. The requirement $\mu \neq 0$ rules out in particular the use of our approach for non-Gaussian phase retrieval (where $f(u) = u^2$) as in that case we have $\mu = 0$ when X is centered. But we emphasize that the same holds true in the Gaussian and elliptical setting as well, as noted in Plan and Vershynin (2016) and Goldstein et al. (2016). Their methods also fail to recover the true β^* when the SIM model corresponds to phase retrieval. We refer the reader to §4 for overcoming this limitation using second-order Stein's identity.

We use a sample version of the above program as an estimator for the unknown β^*. In order to deal with the high-dimensional setting, we consider a regularized version of the

above formulation. More specifically, we use the ℓ_1-norm and nuclear norm regularization in the vector and matrix/tensor settings respectively. However, a major difficulty in the sample setting for this procedure is that $\mathbb{E}[Y \cdot S(X)]$ and its empirical counterpart may not be close enough due to a lack of concentration. Recall our discussion from §1 that even if the random variable X is light-tailed, its score-function $S(x)$ might be arbitrarily heavy-tailed. Furthermore, bounded-fourth moment assumption on the noise, Y too can be heavy-tailed. Thus the naive method of using the sample version of (3.2) to estimate β^* leads to sub-optimal statistical rates of convergence.

To improve concentration and obtain optimal rates of convergence, we replace $Y \cdot S(X)$ with a transformed random variable $\mathcal{T}(Y, X)$, which will be defined precisely later for the sparse and low-rank cases. In particular, $\mathcal{T}(Y, X)$ is a carefully truncated version of $Y \cdot S(X)$, introduced and analyzed in Catoni et al. (2012); Fan et al. (2016) for related problems, that enables us to obtain well-concentrated estimators. Thus our final estimator $\widehat{\beta}$ is defined as the solution to the following regularized optimization problem

$$\underset{\beta \in \mathbb{R}^d}{\text{minimize}} \;\; L(\beta) + \lambda \cdot R(\beta), \quad L(\beta) = \langle \beta, \beta \rangle - \frac{2}{n} \sum_{i=1}^{n} \langle \mathcal{T}(Y_i, X_i), \beta \rangle, \tag{3.3}$$

where $\lambda > 0$ is the regularization parameter which will be specified later and $R(\cdot)$ is the ℓ_1-norm in the vector case and the nuclear norm in the matrix/tensor case. We now introduce our main moment assumption for first-order SIM. This assumption is made apart from the assumptions made on the noise and the link function. Recall that each entry of the score function defined in (3.1) is equal to $s_0(u) = -p_0'(u)/p_0(u)$. We first state the assumption and make a few remarks about it.

Assumption 3.1 (Moment Assumptions)**.** *There exists an absolute constant $M > 0$ such that $\mathbb{E}(Y^4) \leq M$ and $\mathbb{E}_{p_0}[s_0^4(U)] \leq M$, where random variable $U \in \mathbb{R}$ has density p_0.*

Consider the assumption $\mathbb{E}(Y^4) \leq M$. By Cauchy-Schwarz inequality we have $\mathbb{E}(Y^4) \leq 4\mathbb{E}(\epsilon^4) + 4\mathbb{E}[f^4(\langle X, \beta^* \rangle)]$. Note that we assum ϵ to be centered, independent of X and has bounded fourth moment (see §2). If the covariate X has bounded fourth moment along the direction of true parameter, since $f(\cdot)$ is continuously differentiable, $f(\langle X, \beta^* \rangle)$ has bounded fourth moment as well if $f(\cdot)$ is defined on a compact subset of $\mathbb{R}$. . Hence the condition $\mathbb{E}(Y^4) \leq M$ is relatively easy to satisfy and significantly milder than assuming that Y is bounded or has lighter tails. Furthermore, $\mathbb{E}_{p_0}[s_0^4(U)] \leq M$ is relatively mild and satisfied by

a wide class of random variables. Specifically random variables that are non-symmetric and non-Gaussian satisfy this property thereby allowing our approach to work with covariates not previously possible. We believe it is highly non-trivial to weaken this condition without losing significantly in the rates of convergence that we discuss below.

3.1 Sparse Vector SIM

Under the above assumptions, we first state our theorem on the sparse SIM. As discussed above, $Y \cdot S(X)$ can by heavy-tailed and hence we apply truncation to achieve concentration. Denote the j-th entry of the score function S in (3.1) as $S_j \colon \mathbb{R}^d \to \mathbb{R}$, $j \in [d]$. We define the truncated response and score function as

$$\widetilde{Y} = \operatorname{sign}(Y) \cdot (|Y| \wedge \tau), \quad \widetilde{S}_j(x) = \operatorname{sign}[S_j(x)] \cdot \big[|S_j(x)| \wedge \tau\big], \tag{3.4}$$

where $\tau > 0$ is a predetermined threshold value. We define $\widetilde{Y}_i$ similarly for all Y_i, $i \in [n]$. Then we define the estimator $\widehat{\beta}$ as the solution to the optimization problem in (3.3) with $\mathcal{T}(Y_i, X_i) = \widetilde{Y}_i \cdot \widetilde{S}(X_i)$ and $R(\beta) = \|\beta\|_1$. Here we apply elementwise truncation in $\mathcal{T}$ to ensure the sample average of $\mathcal{T}$ converges to $\mathbb{E}[Y \cdot S(X)]$ in the ℓ_∞-norm for an appropriately chosen τ. Note that the ℓ_∞-norm is the dual norm of the ℓ_1-norm. Such a convergence requirement in the dual norm is standard in the analysis of regularized M-estimators (Negahban et al., 2012) to achieve optimal rates. The following theorem characterizes the convergence rates of $\widehat{\beta}$.

Theorem 3.2 (Signal Recovery for Sparse Vector SIM)**.** *For the sparse SIM defined in §2, we assume that $\beta^* \in \mathbb{R}^d$ has s^* nonzero entries. Under Assumption 4.1, we let*

$$\tau = 2(M \cdot \log d/n)^{1/4}$$

in (3.4) and set the regularization parameter λ in (3.3) as

$$\lambda = C\sqrt{M \cdot \log d/n},$$

where $C > 0$ is an absolute constant. Then with probability at least $1 - d^{-2}$, the ℓ_1-regularized estimator $\widehat{\beta}$ defined in (3.3) satisfies

$$\|\widehat{\beta} - \mu\beta^*\|_2 \le \sqrt{s^*} \cdot \lambda, \quad \|\widehat{\beta} - \mu\beta^*\|_1 \le 4s^* \cdot \lambda.$$

From this theorem, the ℓ_1- and ℓ_2-convergence rates of $\widehat{\beta}$ are $\|\widehat{\beta}-\mu\beta^*\|_1 = \mathcal{O}(s^*\sqrt{\log d/n})$ and $\|\widehat{\beta}-\mu\beta^*\|_2 = \mathcal{O}(\sqrt{s^*\log d/n})$, respectively. These rates match the convergence rates of sparse generalized linear models (Loh and Wainwright, 2015) and sparse single index models with Gaussian and symmetric elliptical covariates (Plan and Vershynin, 2016; Goldstein et al., 2016) which are known to be minimax-optimal for this problem via matching lower bounds.

3.2 Low-rank Matrix SIM

We next state our theorem for the low-rank Matrix SIM. In this case, we apply the nuclear norm regularization to promote low-rankness. Note that by definition, $\mathcal{T}$ is matrix-valued. Since the dual norm of the nuclear norm is the operator norm, we need the sample average of $\mathcal{T}$ to converge to $\mathbb{E}[Y \cdot S(X)]$ in the operator norm rapidly to achieve optimal rates of convergence. To achieve such a goal, we leverage the truncation argument from Catoni et al. (2012); Minsker (2016) to construct $\mathcal{T}(Y, X)$.

Let $\phi\colon \mathbb{R} \to \mathbb{R}$ be a non-decreasing function such that

$$-\log(1 - x + x^2/2) \le \phi(x) \le \log(1 + x + x^2/2), \quad \forall x \in \mathbb{R}.$$

Based on ϕ, we define a linear mapping $\psi\colon \mathbb{R}^{d_1\times d_2} \to \mathbb{R}^{d_1\times d_2}$ as follows. For any $A \in \mathbb{R}^{d_1\times d_2}$, let

$$\widetilde{A} = \begin{bmatrix} 0 & A \\ A^\top & 0 \end{bmatrix}$$

and let $\Upsilon\Lambda\Upsilon^\top$ be the eigenvalue composition of $\widetilde{A}$. In addition, let $B = \Upsilon\psi\circ(\Lambda)\Upsilon^\top$, where ψ is applied elementwisely on Λ. Then we write B in block from as

$$B = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}$$

and define $\psi(A) = B_{12}$. Finally, we define $\mathcal{T}(Y, X) = 1/\kappa \cdot \psi\big[\kappa \cdot Y \cdot S(X)\big]$, where $\kappa > 0$ will be specified later. Therefore, our final estimator $\widehat{\beta} \in \mathbb{R}^{d_1\times d_2}$ is defined as the solution to the optimization problem in (3.3) with $R(\beta) = \|\beta\|_\star$. We note here the minimization in (3.3) is taken over $\mathbb{R}^{d_1\times d_2}$. The following theorem quantifies the convergence rates of the proposed estimator.

Theorem 3.3 (Signal Recovery for Low-rank Matrix SIM). *For the low-rank single index model defined in §2, we assume that* $\mathrm{rank}(\beta^*) = r^*$. *Under Assumption 4.1, we let*

$$\kappa = 2\sqrt{n \cdot \log(d_1 + d_2)}/\sqrt{(d_1 + d_2)M}$$

in $\mathcal{T}(Y, X)$ *and set* λ *in* (3.3) *as*

$$\lambda = C\sqrt{M \cdot (d_1 + d_2) \cdot \log(d_1 + d_2)/n},$$

where $C > 0$ *is an absolute constant. Then with probability at least* $1 - (d_1 + d_2)^{-2}$, *the nuclear norm regularized estimator* $\widehat{\beta}$ *satisfies*

$$\|\widehat{\beta} - \mu\beta^*\|_{fro} \leq 3\sqrt{r^*} \cdot \lambda, \quad \|\widehat{\beta} - \mu\beta^*\|_\star \leq 12r^* \cdot \lambda.$$

By this theorem, we have $\|\widehat{\beta} - \mu\beta^*\|_{\mathrm{fro}} = \mathcal{O}(\sqrt{r^*(d_1 + d_2) \cdot \log(d_1 + d_2)/n})$ and $\|\widehat{\beta} - \mu\beta^*\|_\star = \mathcal{O}(r^* \cdot \sqrt{(d_1 + d_2) \cdot \log(d_1 + d_2)/n})$. Note that the rate obtained is minimax-optimal up to a logarithmic factor. Furthermore, it matches the rates for low-rank single index models with Gaussian and symmetric elliptical distributions up to a logarithmic factor Plan and Vershynin (2016); Goldstein et al. (2016).

3.3 Low-rank Tensor SIM

We now state our result for low-rank tensor SIM. The notion of rank of a tensor is more delicate compared to that of a matrix. Several generalizations of the matrix rank exist for the case of tensors. Recall from Definition 2.4, that we assumed that the structure on β^* is that it has low CP-rank. Unfortunately, enforcing such a constraint via a direct tensor nuclear norm relaxation (similar to that of the matrix nuclear norm) is NP-hard (Friedland and Lim, 2014).

One way to overcome such a computational hurdle is to deal with tensors via appropriately matricized forms. In order to enable computable estimators, we specifically leverage the results of Mu et al. (2014) and define the following square-unfolding of a tensor. Denote by $\mathrm{Mat}\colon \mathbb{R}^{d^{\otimes 4}} \to \mathbb{R}^{d^2 \times d^2}$ the operation of tensor square-unfolding, which maps a fourth-order tensor to a square matrix. More specifically, the entries of $\mathrm{Mat}(Z)$ are specified by $[\mathrm{Mat}(Z)]_{k_1,k_2} = Z(j_1, j_2, j_3, j_4)$, where the indices satisfy the relationship $k_1 = 1 + (j_1 - 1) + (j_2 - 1) \cdot d$ and $k_2 = 1 + (j_3 - 1) + (j_4 - 1) \cdot d$. Intuitively, the matrix

obtained by the square-unfolding operation is as square as possible, i.e., it is $d^2 \times d^2$ rather than the rectangular $d \times d^3$ or $d^3 \times d$ matrices. It is shown in Mu et al. (2014) such a square matricization preserves the low CP-rank of the original tensor. Hence one could use the matrix nuclear norm relaxation on the square-unfolded tensor. Furthermore, for the case of tensor SIM as in Definition 2.4, note that we have $\langle X, \beta^* \rangle = \langle \mathrm{Mat}(X), \mathrm{Mat}(\beta^*) \rangle$. Combining the above observations, the low CP-rank tensor SIM problem could be reduced to that of low-rank matrix SIM problem, where matrix low-rank constraint, via nuclear norm, is enforced on $\mathrm{Mat}(\beta^*)$. Thus, we use the estimator in (3.3) with $R(\beta) = \|\mathrm{Mat}(\beta)\|_*$ and $\mathrm{Mat}(X_i)$ for all $i = 1, \ldots, n$ along with the truncation operation $\mathcal{T}$ described in §3.2. We now have the following theorem for the low-rank tensor SIM.

Theorem 3.4 (Signal Recovery for Low-rank Tensor SIM)**.** *For the low-rank tensor single index model defined in §2, Definition 2.4, we assume that* $\mathrm{rank}_{CP}(\beta^*) = r^*$. *Under Assumption 4.1, we let*

$$\kappa = 2\sqrt{2n \cdot \log d}/\sqrt{(2d^2)M}$$

in $\mathcal{T}(Y, X)$ *and set* λ *in* (3.3) *as*

$$\lambda = C\sqrt{2M \cdot (2d^2) \cdot \log d/n},$$

where $C > 0$ *is an absolute constant. Then with probability at least* $1 - 2d$*, the nuclear-norm regularized estimator* $\widehat{\beta}$ *satisfies*

$$\|\widehat{\beta} - \mu\, Mat(\beta^*)\|_{fro} \leq 3\sqrt{r^*} \cdot \lambda$$

We omit the proof of the above theorem as it is follows the exact steps of Theorem 3.3 proved in Appendix A.2. From the above theorem, we see that as long as $n = \Omega(r^* d^2)$, we achieve consistent estimation of β^* up to scaling. This improves upon recent results established in Chen et al. (2016), that established similar results under restrictive Gaussian covariate assumption and required knowledge of the link functions (i.e., generalized linear models). Furthermore our results significantly generalizes the results of Mu et al. (2014) that considered only linear link functions. Finally, although our structure on β^* was a low CP-rank structure, the square matricization technique also applies for the case of low Tucker-rank, which is yet another notion of rank for tensors with several applications. It is straightforward to extend our results to this case of low Tucker-rank.

4 Theoretical Results for Index Models with Second-Order Links

We now introduce our estimators and establish their statistical rates of convergence for the case of index models with second-order link functions. Discussions on optimality of the established rates and connection to sparse PCA problem is deferred to §4.3. Similar to the first-order case, we focus on the case where X has i.i.d. entries with density $p_0 \colon \mathbb{R} \to \mathbb{R}$. Thus the joint density of X is $p(x) = p_0^{\otimes d}(x) = \prod_{j=1}^{d} p_0(x_j)$. We define a univariate function $s_0 \colon \mathbb{R} \to \mathbb{R}$ by $s_0(u) = p_0'(u)/p_0(u)$. Then the first-order score function associated with p is given by $S(x) = s_0 \circ (x)$. Equivalently, the j-th entry of the first-order score function associated with p is given by $[S(x)]_j = s_0(x_j)$. Moreover, the second order score function is

$$T(x) = S(x)S^{\top}(x) - \nabla S(x) = S(x)S^{\top}(x) - \operatorname{diag}[s_0' \circ (x)]. \tag{4.1}$$

Before we present our estimator, we introduce the assumption on Y and $s_0(\cdot)$.

Assumption 4.1 (Moment Assumptions)**.** *We assume that there exists a constant M such that $\mathbb{E}_{p_0}[s_0^6(U)] \le M$ and $\mathbb{E}(Y^6) \le M$. We denote $\sigma_0^2 = \mathbb{E}_{p_0}[s_0^2(U)] = \operatorname{Var}_{p_0}[s_0(U)]$.*

The assumption that $\mathbb{E}_{p_0}[s_0^6(U)] \le M$ allows wide family of distributions of including Gaussian and more heavy-tailed random variables. Furthermore, we do not require the covariate to be elliptically symmetric as is commonly seen in prior work, which enables our method to be applicable for skewed covariates. As for the assumption that $\mathbb{E}(Y^6) \le M$, in the case of SIMs, we have $\mathbb{E}(Y^6) \le C\left(\mathbb{E}(\epsilon^6) + \mathbb{E}[f^6(\langle X, \beta^* \rangle)]\right)$. Thus this assumption is satisfied as long as both ϵ and $f(\langle X, \beta^* \rangle)$ have bounded sixth moments. This is a significantly milder assumption which allows for heavy-tailed response as opposed to bounded or light-tailed response.

4.1 Sparse Vector SIM

Now we are ready to describe our estimator for the sparse SIMs in Definition 2.6. Note that by Proposition 2.5 we have

$$\mathbb{E}\big[Y \cdot T(X)\big] = C_0 \cdot \beta^* {\beta^*}^{\top}, \tag{4.2}$$

where $C_0 = 2\mathbb{E}[f''(\langle X, \beta^* \rangle)] > 0$ as per Definition 2.6. Therefore, one way to estimator β^* is to obtain the leading eigenvector of $\mathbb{E}[Y \cdot T(X)]$ from the samples. Since β^* is sparse, we formulate our estimator as a sparsity constrained semi-definite program:

$$\begin{aligned} &\text{maximize} \quad \langle W, \Sigma^* \rangle + \lambda \|W\|_1 \\ &\text{subject to} \quad 0 \preceq W \preceq I_d, \quad \text{TRACE}(W) = 1. \end{aligned} \tag{4.3}$$

where $\Sigma^* = \mathbb{E}[Y \cdot T(X)]$. Note that both the score $T(X)$ and the response variable Y can be heavy-tailed. In order to obtain near-optimal estimates in the sample setting, we apply truncation to handle the heavy-tails. Specifically, for a positive parameter $\tau \in \mathbb{R}$, we define the truncated random variables by

$$\widetilde{Y}_i = \text{sign}(Y_i) \cdot \min\{|Y_i|, \tau\} \quad \text{and} \quad \widetilde{T}_{jk}(X_i) = \text{sign}\{T_{jk}(X_i)\} \cdot \min\big\{|T_{jk}(X_i)|, \tau^2\big\}. \tag{4.4}$$

Then we define an robust estimator of Σ^* as

$$\widetilde{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} \widetilde{Y}_i \cdot \widetilde{T}(X_i). \tag{4.5}$$

Given $\widetilde{\Sigma}$, let $\widehat{W}$ be the solution of the following convex optimization problem

$$\begin{aligned} &\text{maximize} \quad \langle W, \widetilde{\Sigma} \rangle + \lambda \|W\|_1 \\ &\text{subject to} \quad 0 \preceq W \preceq I_d, \quad \text{TRACE}(W) = 1. \end{aligned} \tag{4.6}$$

Here λ is a regularization parameter to be specified later. The final estimator is defined as the leading eigenvector of $\widehat{W}$. The following theorem quantifies the statistical rates of convergence of the proposed estimator.

Theorem 4.2 (Signal Recovery for Spare SIM)**.** *Let $\widehat{W}$ be the solution of the optimization in (4.6) and let $\widehat{\beta}$ be the leading eigenvector of $\widehat{W}$. We set the regularization parameter λ in (4.6) as $\lambda = 10\sqrt{M \log d / n}$ and set $\tau = (1.5Mn/\log d)^{1/6}$ in (4.4). Under Assumption 4.1, we have $\|\widehat{\beta} - \beta^*\|_2 \leq 4\sqrt{2}s^*\lambda$ with probability at least $1 - d^{-2}$.*

By this theorem, the ℓ_2-error of the proposed estimator is $\mathcal{O}(s^*\sqrt{\log d / n})$, which implies that consistent estimation requires $n = \Omega(s^{*2} \log d)$ samples.

4.2 Subspace-Sparse MIM

Now we introduce the estimator for B^* of the sparse MIM in Definition 2.7. Proposition 2.5 implies that $\mathbb{E}[Y \cdot T(X)] = B^* D_0 B^*$, where $D_0 = \mathbb{E}[\nabla^2 f(XB^*)]$ is positive definite. Similar to (4.6), we recover the column space of B^* by solving

$$\begin{aligned} &\text{maximize} \quad \langle W, \widetilde{\Sigma} \rangle + \lambda \|W\|_1, \\ &\text{subject to} \quad 0 \preceq W \preceq I_d, \quad \text{TRACE}(W) = k. \end{aligned} \tag{4.7}$$

where $\widetilde{\Sigma}$ is defined in (4.5), $\lambda > 0$ is a regularization parameter and k is the number of indices which is assumed to be known. Let $\widehat{W}$ be the solution of (4.7), the final estimator is the top k eigenvectors of $\widehat{W}$. For the above estimator, we have the following theorem quantifying the statistical rate of convergence. Let $\rho_0 = \lambda_{\text{MIN}}(\mathbb{E}[\nabla^2 f(XB^*)])$.

Theorem 4.3 (Signal Recovery for Sparse MIM)**.** *Let $\widehat{W}$ be the solution of the optimization problem in (4.7) and let $\widehat{B}$ be the leading eigenvector of $\widehat{W}$. We set the regularization parameter in (4.6) as $\lambda = 10\sqrt{M \log d / n}$ and let the truncation parameter in (4.4) be $\tau = (1.5Mn/\log d)^{1/6}$. Under Assumption 4.1, with probability at least $1 - d^{-2}$, we have*

$$\inf_{O \in \mathbb{O}_k} \|\widehat{B} - B^* O\|_2 \leq 4\sqrt{2}/\rho_0 \cdot s^* \lambda.$$

Minimax lower bounds for subspace estimation for MIM was established recently in Lin et al. (2017). For a fixed k, the above theorem is near-optimal from a minimax point of view. That is, the difference between the optimal rate and the above theorem is a factor of $\sqrt{s}$. We discuss more about this gap in Section 4.3. The proofs of Theorem 4.2 and Theorem 4.3 are in the supplementary material.

Remark 1. *Recall that our discussion in §3 and §4 was under the assumption that the entries in X are i.i.d. This could be relaxed to the case of weak dependence between the covariates without any significant loss in the statistical rates we present in the theorems above. We do not focus on such an extension in this paper as we wanted to clearly convey the main message of the paper in a simpler setting.*

4.3 Optimality and Relation to Sparse PCA

Now we discuss the optimality of the results presented in §4. Throughout the discussion we assume that k is fixed and does not increase with n. Note that the estimator for SIM in

(4.6) and MIM in (4.7) are closely related to the semidefinite program based estimator for Sparse PCA problem (Vu et al., 2013). Let $X \in \mathbb{R}^d$ be a random vector such that $\mathbb{E}(X) = 0$ and covariance matrix $\Sigma = \mathbb{E}(XX^\top)$ which is symmetric and positive definite. The problem of sparse PCA is to estimate projector onto the subspace spanned by top k eigenvectors, $\{v_\ell^*\}_{\ell \in [k]}$ of Σ under the subspace sparsity assumption as discussed in Definition 2.7. An estimator based on semidefinite programing with sparsity constraints was analyzed in Vu et al. (2013); Wang et al. (2016), which is based on solving the following program

$$\begin{aligned} &\text{maximize} \quad \langle W, \widehat{\Sigma} \rangle + \lambda \|W\|_1 \\ &\text{subject to} \quad 0 \preceq W \preceq I_d, \quad \textsc{trace}(W) = k. \end{aligned} \tag{4.8}$$

Here $\widehat{\Sigma} = n^{-1} \sum_{i=1}^n X_i X_i^\top$ is the sample covariance matrix given n i.i.d copies $\{X_i\}_{i=1}^n$ of X. Note that the main difference between the SIM estimator and the sparse PCA estimator is the use of $\widetilde{\Sigma}$ in place of $\widehat{\Sigma}$. It is known that sparse PCA problem exhibits interesting statistical-computational tradeoff (Krauthgamer et al., 2015; Wang et al., 2016) which naturally appears in the context of SIM as well. Indeed while the minimax optimal statistical rate for sparse PCA is $\mathcal{O}(\sqrt{s^* \log d/n})$, the SDP estimator achieves $\mathcal{O}(s^* \sqrt{\log d/n})$ under the assumption that X is light-tailed. It is also known that when $n = \Omega(s^{*2} \log d)$, one can obtain the optimal statistical rate of $\mathcal{O}(\sqrt{s^* \log d/n})$ either by nonconvex methods (Wang et al., 2014), or refinements to the output of the SDP estimator (Wang et al., 2016). However their results rely on the sharp concentration of $\widehat{\Sigma}$ to Σ in the restricted operator norm:

$$\|\widehat{\Sigma} - \Sigma^*\|_{op,s} = \sup\big\{w^\top (\widehat{\Sigma} - \Sigma) w \colon \|w\|_2 = 1, \|w\|_0 \le s\big\} = \mathcal{O}(\sqrt{s \log d/n}). \tag{4.9}$$

When X has heavy-tailed entries, for example bounded fourth moment assumptions, its highly unlikely that, (4.9) holds. Indeed the results in Wang et al. (2016) and Wang et al. (2014) are applicable only to the case of Gaussian or light-tailed X.

4.3.1 Heavy-tailed Sparse PCA

Recall that our estimators utilize a data-driven truncation argument to handle heavy-tailed distributions. Owing to the close relationship between our SIM/MIM estimators and the sparse PCA estimator, it is natural to ask whether such a truncation argument could lead to sparse PCA estimators for heavy tailed X. Below we show that it is indeed possible

to obtain a near-optimal estimator for sparse PCA with heavy-tailed data based on the truncation argument. For a vector $v \in \mathbb{R}^d$, let $\vartheta(v)$ be a truncation operation that operators entry-wise as $\vartheta_j(v) = \text{sign}[v_j] \cdot \min\{|v_j|, \tau\}$ for $j = 1, \dots d$. Then, our estimator is defined as follows.

$$\begin{aligned} &\text{maximize} \quad \langle W, \overline{\Sigma} \rangle + \lambda \|W\|_1 \\ &\text{subject to} \quad 0 \preceq W \preceq I_d, \quad \text{TRACE}(W) = k. \end{aligned} \tag{4.10}$$

where $\overline{\Sigma} = n^{-1} \sum_{i=1}^n \overline{X}_i \overline{X}_i^\top$ and $\overline{X}_i = \vartheta(X_i)$, for $i = 1, \dots n$. For the above estimator, we have the following theorem under the assumption that X has heavy-tailed marginals. Let $V^* = (v_1^* \dots v_k^*) \in \mathbb{R}^{d \times k}$ and assume that $\rho_0 = \lambda_k(\Sigma) - \lambda_{k+1}(\Sigma) > 0$.

Theorem 4.4. *Let $\widehat{W}$ be the solution of the optimization in (4.10) and let $\widehat{V}$ be the leading eigenvector of $\widehat{W}$. We set the regularization parameter in (4.10) as $\lambda = C_1 \sqrt{M \log d / n}$ and set the truncation parameter by $\tau = (C_2 M n / \log d)^{1/4}$, where C_1 and C_2 are some positive constants. Furthermore, assume that V^* contains only s^* nonzero rows and that X satisfies $\mathbb{E}|X_j|^4 \le M$ and $\mathbb{E}|X_i \cdot X_j|^2 \le M$. Then, with probability at least $1 - d^{-2}$, we have*

$$\inf_{O \in \mathbb{O}_k} \|\widehat{V} - V^* O\|_2 \le 4\sqrt{2}/\rho_0 \cdot s^* \lambda.$$

The proof of the above theorem is similar to that of Theorem 4.3 and hence we omit it. The above theorem shows that with elementwise truncation, as long as X satisfies a bounded fourth moment condition, the SDP estimator for sparse PCA achieves the near-optimal statistical rate of $\mathcal{O}(s^* \sqrt{\log d / n})$. We end this section with the following questions based on the above discussions:

1. Can we obtain optimal statistical rate for sparse PCA problem ($\mathcal{O}(\sqrt{s \log d / n})$) when X has only bounded fourth moment in the high sample size regime $n = \Omega(s^{*2} \log d)$?

2. Can we obtain optimal statistical rate ($\mathcal{O}(\sqrt{s^* \log d / n})$) when $n = \Omega(s^{*2} \log d)$ and when f, X and Y satisfies the heavy-tail condition in Assumption 4.1 for the MIM problem ?

The answer to both questions lie in constructing truncation based estimators that concentrate sharply in restricted operator norm as in (4.9) or more realistically exhibit one-sided

concentration bounds (see Mendelson (2014) and Oliveira (2013) for related results and discussion). Obtaining such an estimator seems to be challenging for heavy-tailed sparse PCA and it it not immediately clear if it is possible. We plan to report our findings for the above problem in the near future.

5 Numerical Experiments

We now provide simulation experiments for the case of first-order and second-order SIMs. For the first-order SIM, we concentrate on the sparse vector and low-rank matrix model. Note that our tensor estimator is similar to the low-rank matrix estimator. Furthermore, for the second-order case, we concentrate on the problem of robust sparse phase retrieval.

First-order SIM: We let $\epsilon \sim N(0,1)$ and set the link function in (2.1) as one of $f_1(u) = 3u + 10\sin(u)$ and $f_2(u) = \sqrt{2}u + 4\exp(-2u^2)$, which are plotted in Figure 2. We set p_0 to be one of (i) Gamma distribution with shape parameter 5 and scale parameter 1, (ii) Student's t-distribution with 5 degrees of freedom, and (iii) Rayleigh distribution with scale parameter 2. To measure the estimation accuracy, we use the cosine distance $\cos\theta(\widehat{\beta}, \beta^*) = 1 - \|\widehat{\beta}\|_\bullet^{-1}|\langle\widehat{\beta}, \beta^*\rangle|$, where $\bullet$ stands for the Euclidean norm in the vector case and the Frobenius norm when β^* is a matrix. Here we report the cosine distance rather than $\|\widehat{\beta} - \mu\beta^*\|_\bullet$ to compare the performances for X having different distributions, where μ may have different values.

For the vector case, we fix $d = 2000$, $s^* = 5$ and vary n. The support of β^* is chosen uniformly random among all subsets of $\{1, \ldots, d\}$. For each $j \in \mathrm{supp}(\beta^*)$, we set $\beta_j^* = 1/\sqrt{s^*} \cdot \gamma_j$, where each γ_j is an i.i.d. Rademacher random variable. In addition, the regularization parameter λ is set to $4\sqrt{\log d/n}$. We plot the cosine distance against the signal strength $\sqrt{s^* \log d/n}$ in Figure 5-(a) and (b) for f_1 and f_2 respectively, based on 200 independent trials for each n. As shown in this figure, the estimation error grows sub-linearly as a function of the signal strength.

As for the matrix case, we fix $d_1 = d_2 = 20$, $r^* = 3$ and let n vary. The signal parameter β^* is equal to $USV^\top$, where $U, V \in \mathbb{R}^{d\times d}$ are random orthogonal matrices and S is a diagonal matrix with r^* nonzero entries. Moreover, we set the nonzero diagonal entries of S as $1/\sqrt{r^*}$, which implies $\|\beta^*\|_{\mathrm{fro}} = 1$. We set the regularization parameter as $\lambda = 2\sqrt{(d_1 + d_2)\log(d_1 + d_2)/n}$. Furthermore, we use the proximal gradient descent algo-

rithm (with the learning rate fixed to 0.05) to solve the nuclear norm regularization problem in (3.3). To present the result, we plot the cosine distant against the signal strength $\sqrt{r^*(d_1+d_2)\log(d_1+d_2)/n}$ in Figure 5-(b) based on 200 independent trials. As shown in this figure, the error is bounded by a linear function of the signal strength, which corroborates Theorem 3.3.

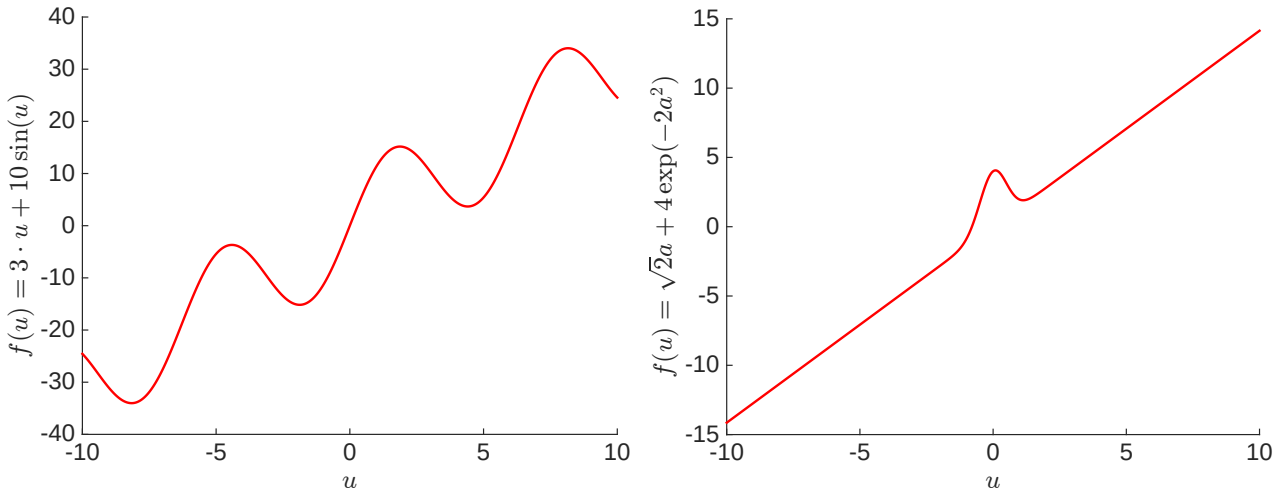


Figure 2: Plot of the link functions $f_1(u) = 3u + 10 \cdot \sin(u)$ and $f_2(u) = \sqrt{2}u + 4\exp(-2u^2)$.

Second-order SIM: We now concentrate on the problem of sparse phase retrieval using the SDP based estimators proposed based on second-order Stein's identity. Recall that in this case, the link function is known and existing convex and non-convex based estimators are applicable predominantly for the case of Gaussian or light-tailed data. The question of de-randomization or what are the necessary assumptions on the measurement vectors for (sparse) phase retrieval to work is an intriguing one (Gross et al., 2015). Here we demonstrate that using the proposed score-based estimators, one could deal with heavy-tailed and skewed measurement as well, which significantly extend the class of measurement vectors applicable for sparse phase retrieval.

Recall that the covariate X has i.i.d. entries with distribution p_0. We set p_0 to be one of Gamma distribution with shape parameter 5 and scale parameter 1 and Rayleigh distribution with scale parameter 2. The random noise ϵ is set to be standard Gaussian. Moreover, we solve the optimization problems in (4.6) and (4.7) via the alternating direction method of multipliers (ADMM) algorithm proposed in Vu et al. (2013), which introduces a dual variable to encode the constrains and updates the primal and dual variables iteratively.

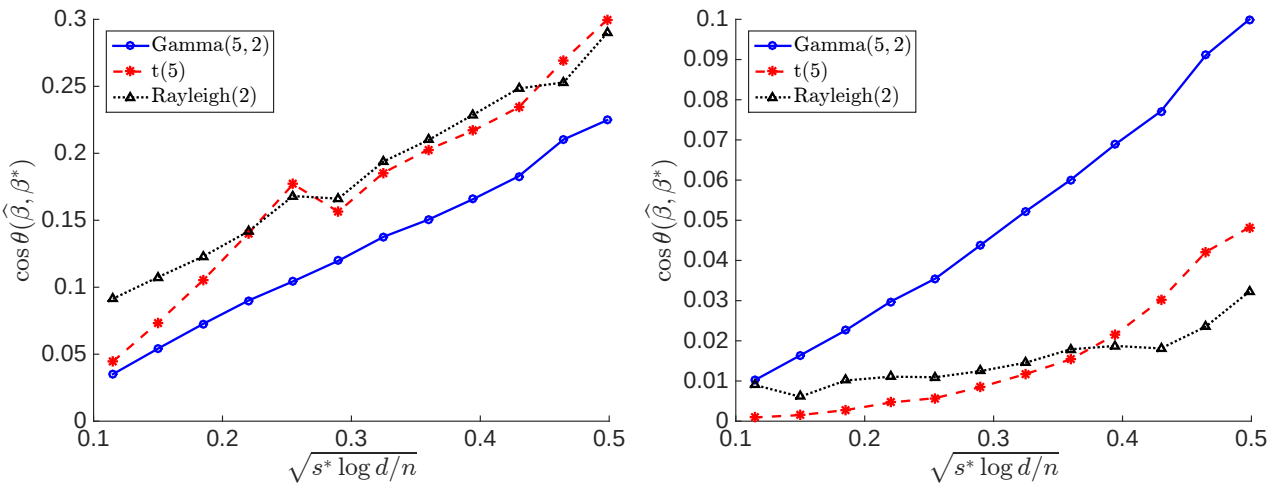


Figure 3: Cosine distances between the true parameter and the estimated parameter in the sparse SIM with the link function in 2.1 set to one of f_1 and f_2. Here we set $d = 2000$. $s^* = 5$ and vary n.

We set the link function to be one of $f_3(u) = u^2$, $f_4 = |u|$, and $f_5(u) = 4u^2 + 3\cos(u)$. Here f_3 corresponds to the phase retrieval model and f_4 and f_5 can be viewed as its robust extension. Throughout the experiment we fix $d = 500$, $s^* = 5$ and vary n. The support of β^* is chosen uniformly random among all subsets of $\{1, \ldots, d\}$with cardinality s^*. For each $j \in \mathrm{supp}(\beta^*)$, we set $\beta_j^* = 1/\sqrt{s^*} \cdot \gamma_j$, where γ_j's are i.i.d. Rademacher random variables. Furthermore, we fix the regularization parameter $\lambda = 4\sqrt{\log d/n}$ and threshold parameter $\tau = 20$. In addition, we adopt the cosine distance $\cos\theta(\widehat{\beta}, \beta^*) = 1 - |\langle\widehat{\beta}, \beta^*\rangle|$, to measure the estimation error. We plot the cosine distance against the statistical rate of convergence $s^*\sqrt{\log d/n}$ in Figure 5-(a)-(c) for each link function, respectively. The plot is based on 100 independent trials for each n, which shows that the estimation error is bounded by a linear function of $s^*\sqrt{\log d/n}$, which corroborate the theory.

6 Conclusion

In this work, we consider estimating the parametric components of single and multiple index models in the high-dimensional setting, under fairly general assumptions on the link function f and response Y. Furthermore, our estimators are applicable in the non-Gaussian setting where X is not required to satisfy restrictive Gaussian or elliptical symmetry assumptions.

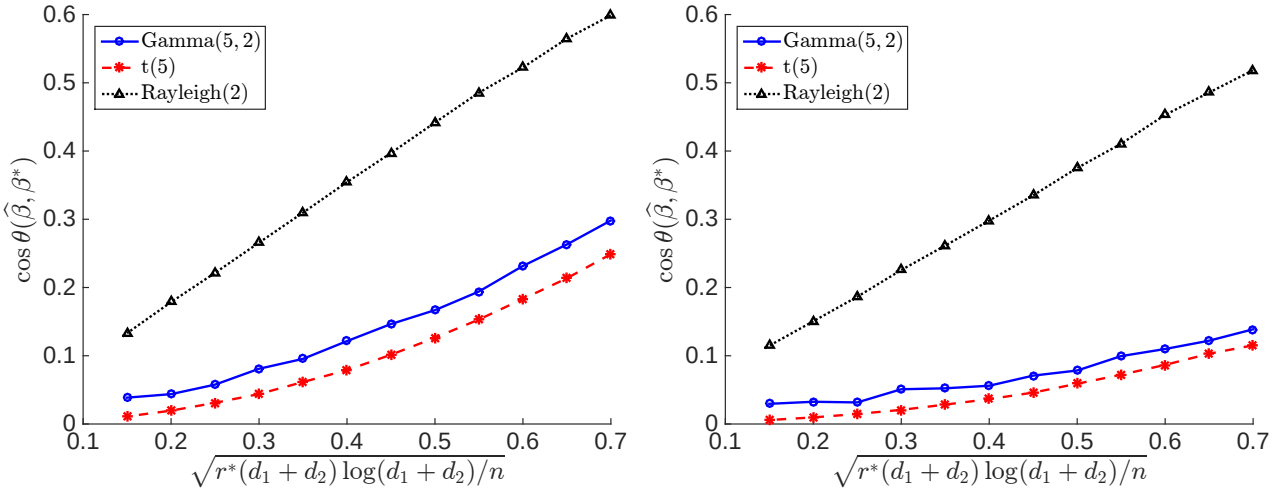


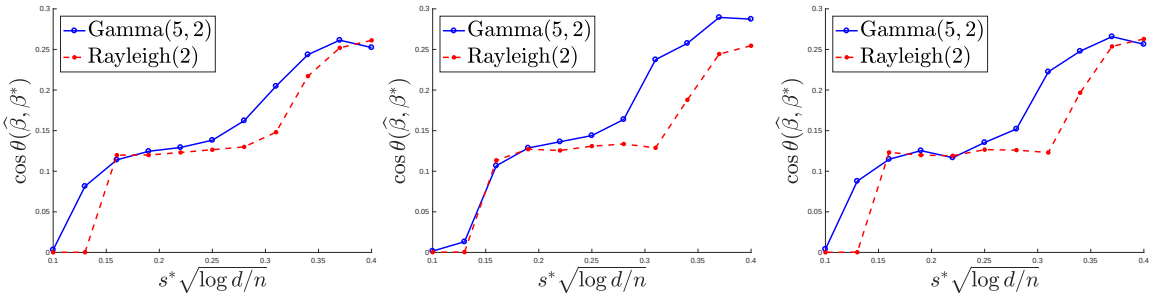
Figure 4: Cosine distances between the true parameter and the estimated parameter in the low-rank SIM for with link function in 2.1 set to one of f_1 and f_2. Here we set $d_1 = d_2 = 20$. $r^* = 3$ and vary n.

Our estimators are based on a data-driven truncation argument in combination with first and second-order Stein's identity. Furthermore, we show that proposed estimators are near-optimal for several different settings.

Recently in the low-dimensional setting, for 2-layer neural networks Janzamin et al. (2015) proposed a tensor-based method for estimating the parametric components. Their estimators are sub-optimal even when we consider the low-dimensional Gaussian setting. An immediate application of our truncation based estimators enables us to obtain optimal results for a fairly general class of covariates in the low-dimensional setting. Obtaining similar optimal or near-optimal results in the high-dimensional setting is of great interest for 2-layer neural networks, albeit challenging. We plan to extend the result of this paper for 2-layer neural networks in the high-dimensional setting and report our results in the near future.

A Proofs of the Main Results

In this section, we lay out the proofs of the theorems in §3 and §4, which establish the statistical rates of convergence of our estimators.



$f_3(u) = u^2$ $f_4(u) = |u|$, $f_5(u) = 4u^2 + 3\cos(u)$

Figure 5: Cosine distances between the true parameter β^* and the estimated parameter $\widehat{\beta}$ in the sparse SIM with the link function in one of f_1, f_2, and f_3. Here we set $d = 500$. $s^* = 5$ and vary n.

A.1 Proof of Theorem 3.2

Proof. Since $\widehat{\beta}$ is the solution of the optimization problem in (3.3), the first-order optimality condition states that

$$\nabla L(\widehat{\beta}) + \lambda\xi = 0, \quad \text{where} \quad \xi \in \partial\|\widehat{\beta}\|_1. \tag{A.1}$$

Then the entries of $\xi \in \mathbb{R}^d$ are given by

$$\xi_j = \mathrm{sign}(\widehat{\beta}_j), \quad \forall j \in \mathrm{supp}(\widehat{\beta}); \quad \xi_j \in [-1, 1], \quad \forall j \notin \mathrm{supp}(\widehat{\beta}).$$

For any index set $\mathcal{A} \subseteq [d]$ and $z \in \mathbb{R}^d$, we define the restriction of z to $\mathcal{A}$, $z_{\mathcal{A}} \in \mathbb{R}^d$, by letting

$$[z_{\mathcal{A}}]_j = z_j \quad \text{if} \quad j \in \mathcal{A}, \quad [z_{\mathcal{A}}]_j = 0 \quad \text{otherwise.}$$

Here $[z_{\mathcal{A}}]_j$ is the j-th entry of $z_{\mathcal{A}}$. Let $\mathcal{S} = \mathrm{supp}(\beta^*)$, then we can write $\xi = \xi_{\mathcal{S}} + \xi_{\mathcal{S}^c}$. For notational simplicity, in the sequel, we define $\theta = \widehat{\beta} - \mu \cdot \beta^*$. Thus by (A.1) it holds that

$$\begin{aligned} \langle \nabla L(\widehat{\beta}) - \nabla L(\mu\beta^*), \theta \rangle &= \langle -\lambda \cdot \xi - \nabla L(\mu\beta^*), \theta \rangle \\ &\le \langle -\lambda \cdot \xi_{\mathcal{S}} - \lambda \cdot \xi_{\mathcal{S}^c}, \theta \rangle + \|\nabla L(\mu\beta^*)\|_\infty \cdot \|\theta\|_1. \end{aligned} \tag{A.2}$$

By the definition of ξ, we have

$$\langle -\lambda \cdot \xi_{\mathcal{S}^c}, \widehat{\beta} - \mu\beta^* \rangle = -\lambda \cdot \|\widehat{\beta}_{\mathcal{S}}\|_1. \tag{A.3}$$

Moreover, since $\|\xi\|_\infty \leq 1$, Hölder's inequality implies that

$$\langle -\lambda \cdot \xi_{\mathcal{S}}, \theta \rangle \leq \|\theta_{\mathcal{S}}\|_1. \tag{A.4}$$

Note that $\nabla^2 L(\beta) = 2I_d$. Combining (A.9), (A.3), and (A.4), we obtain

$$2\|\theta\|_2^2 = \langle \nabla L(\widehat{\beta}) - \nabla L(\mu\beta^*), \theta \rangle \leq -\lambda \|\theta_{\mathcal{S}^c}\|_1 + \lambda \|\theta_{\mathcal{S}}\|_1 + \|\nabla L(\mu\beta^*)\|_\infty \cdot \|\theta\|_1. \tag{A.5}$$

For an upper bound of the right-hand side of (A.5), we apply the following lemma to obtain an upper bound on $\|\nabla L(\mu\beta^*)\|_\infty$.

Lemma 1 (Bound on $\|\nabla L(\mu\beta^*)\|_\infty$)**.** *We set the truncation level in* (3.4) *as* $\tau = 2(M \cdot n/\log d)^{1/4}$*. Then we have*

$$\mathbb{P}\Big[\|\nabla L(\mu\beta^*)\|_\infty > 7\sqrt{M \cdot \log d/n}\Big] \leq d^{-2}.$$

Proof. See §B.1 for a detailed proof. □

Thus by Lemma 1 and the choice of λ, we have $\lambda > 2\|\nabla L(\mu\beta^*)\|_\infty$ with probability at least $1 - d^{-2}$. This implies that

$$2\|\theta\|_2^2 \leq -\lambda/2 \cdot \|\theta_{\mathcal{S}^c}\|_1 + 3\lambda/2 \cdot \|\theta_{\mathcal{S}}\|_1 \leq 2\lambda \cdot \|\theta_{\mathcal{S}}\|_1. \tag{A.6}$$

Since the leftmost term in (A.6) is nonnegative, we obtain $\|\theta_{\mathcal{S}^c}\|_1 \leq 3 \cdot \|\theta_{\mathcal{S}}\|_1$. In addition, since $\mathcal{S}| = s^*$, $\|\theta_{\mathcal{S}}\|_1 \leq \sqrt{s^*} \cdot \|\theta_{\mathcal{S}}\|_2$. Thus by (A.6) we have $\|\theta\|_2 \leq \sqrt{s^*} \cdot \lambda$. Moreover, we also have $\|\theta_{\mathcal{S}}\|_1 \leq s^*\lambda$, which further implies that

$$\|\theta\|_1 = \|\theta_{\mathcal{S}}\|_1 + \|\theta_{\mathcal{S}^c}\|_1 \leq 4 \cdot \|\theta_{\mathcal{S}}\|_1 \leq 4s^*\lambda.$$

Therefore, we conclude the proof. □

A.2 Proof of Theorem 3.3

Proof. The proof of Theorem 3.3 is parallel to that of Theorem 3.2. Here the difference is to handle the nuclear norm regularization, instead of the ℓ_1-penalty. Since $\widehat{\beta}$ is the solution of the optimization problem in (3.3), the first order optimality condition states that

$$L(\widehat{\beta}) + \lambda\|\widehat{\beta}\|_\star \leq L(\mu\beta^*) + \lambda\|\mu\beta^*\|_\star. \tag{A.7}$$

To simplify the notation, we define $\Theta = \widehat{\beta} - \mu \cdot \beta^*$. Since L is quadratic,

$$L(\widehat{\beta}) - L(\mu\beta^*) = \langle \nabla L(\mu\beta^*), \Theta \rangle + 2\|\Theta\|_{\text{fro}}^2, \tag{A.8}$$

where ∇L takes values in $\mathbb{R}^{d_1 \times d_2}$. Then combining (A.7), (A.8), and Hölder's inequality, we have

$$\|\Theta\|_{\text{fro}}^2 \leq -\langle \nabla L(\mu\beta^*), \Theta \rangle + \lambda\|\mu\beta^*\|_\star - \lambda\|\widehat{\beta}\|_\star \leq \big\|\nabla L(\mu\beta^*)\big\|_{\text{op}} \cdot \|\Theta\|_\star + \lambda\|\mu\beta^*\|_\star - \lambda\|\widehat{\beta}\|_\star. \tag{A.9}$$

In the following, we focus on the term $\|\mu\beta^*\|_\star - \|\widehat{\beta}\|_\star$ in (A.9). Let $U\Lambda^* V^\top$ be the singular value decomposition of $\mu\beta^*$, where $U \in \mathbb{R}^{d_1 \times d_1}$ and $V \in \mathbb{R}^{d_2 \times d_2}$ are orthogonal matrices, and $\Lambda^* \in \mathbb{R}^{d_1 \times d_2}$ be formed by the singular values of $\mu\beta^*$. Moreover, since $\text{rank}(\beta^*) = r^*$, Λ^* can be written in block form as

$$\Lambda^* = \begin{bmatrix} \Lambda_{11}^* & 0 \\ 0 & 0 \end{bmatrix}, \tag{A.10}$$

where $\Lambda_{11}^* \in \mathbb{R}^{r^* \times r^*}$ is a diagonal matrix whose diagonal elements are the nonzero singular values of $\mu\beta^*$. We define $\Gamma = U^\top \Theta V$, which can be written in block form as

$$\Gamma = \begin{bmatrix} \Gamma_{11} & \Gamma_{12} \\ \Gamma_{21} & \Gamma_{22} \end{bmatrix},$$

where $\Gamma_{11} \in \mathbb{R}^{r^* \times r^*}$. In addition, we define matrices

$$\Gamma^{(1)} = \begin{bmatrix} 0 & 0 \\ 0 & \Gamma_{22} \end{bmatrix} \quad \text{and} \quad \Gamma^{(2)} = \begin{bmatrix} \Gamma_{11} & \Gamma_{12} \\ \Gamma_{21} & 0 \end{bmatrix}.$$

Then by (A.10) and triangle inequality of the nuclear norm, we have

$$\begin{aligned} \|\widehat{\beta}\|_\star &= \|\mu\beta^* + \Theta\|_\star = \|U(\Lambda^* + \Gamma)V^\top\|_\star = \|\Lambda^* + \Gamma\|_\star \\ &\geq \|\Lambda^* + \Gamma^{(1)}\|_\star - \|\Gamma^{(2)}\|_\star = \|\Lambda^*\|_\star + \|\Gamma^{(1)}\|_\star - \|\Gamma^{(2)}\|_\star, \end{aligned} \tag{A.11}$$

where the last equality follows from the fact that $\Lambda^* + \Gamma^{(1)}$ is block diagonal. Since $\|\mu\beta^*\|_\star = \|\Lambda^*\|_\star$, by (A.11) we obtain

$$\|\mu\beta^*\|_\star - \|\widehat{\beta}\|_\star \leq \|\Gamma^{(2)}\|_\star - \|\Gamma^{(1)}\|_\star. \tag{A.12}$$

In addition, triangle inequality implies that

$$\|\Theta\|_\star = \|U\Gamma V^\top\|_\star \le \|\Gamma^{(1)}\|_\star + \|\Gamma^{(2)}\|_\star. \tag{A.13}$$

Thus combining (A.11), (A.12), (A.13), we have

$$\|\Theta\|_{\text{fro}}^2 \le \big(\big\|\nabla L(\mu\beta^*)\big\|_{\text{op}} + \lambda\big) \cdot \|\Gamma^{(2)}\|_\star + \big(\big\|\nabla L(\mu\beta^*)\big\|_{\text{op}} - \lambda\big) \cdot \|\Gamma^{(1)}\|_\star, \tag{A.14}$$

We utilize the following lemma to obtain an upper bound of $\|\nabla L(\mu\beta^*)\|_{\text{op}}$.

Lemma 2 (Upper bound of $\|\nabla L(\mu\beta^*)\|_{\text{op}}$)**.** *Let loss function $L\colon \mathbb{R}^{d_1\times d_2} \to \mathbb{R}$ be defined in (3.3) for the matrix setting. Setting $\kappa = 2\sqrt{n\cdot\log(d_1+d_2)}/\sqrt{(d_1+d_2)M}$, then it holds that*

$$\mathbb{P}\Big[\|\nabla L(\mu\beta^*)\|_{op} > 6\sqrt{(d_1+d_2)/n}\Big] \le (d_1+d_2)^{-2}.$$

Proof. See B.2 for a detailed proof. □

By Lemma 2 and the choice of λ, we conclude that $\lambda > 2\cdot\|\nabla L(\mu\beta^*)\|_{\text{op}}$ with probability at least $1-(d_1+d_2)^{-3}$. Thus by (A.14) we have

$$\|\Theta\|_{\text{fro}}^2 \le 3\lambda/2\cdot\|\Gamma^{(2)}\|_\star - \lambda/2\cdot\|\Gamma^{(1)}\|_\star \tag{A.15}$$

which implies that $\|\Gamma^{(1)}\|_\star \le 3\cdot\|\Gamma^{(2)}\|_\star$. Moreover, by the subadditivity of rank, we obtain

$$\text{rank}(\Gamma^{(2)}) \le \text{rank}\left(\begin{bmatrix}\Gamma_{11}/2 & \Gamma_{12}\\ 0 & 0\end{bmatrix}\right) + \text{rank}\left(\begin{bmatrix}\Gamma_{11}/2 & 0\\ \Gamma_{21} & 0\end{bmatrix}\right) = 2r^*,$$

which implies that $\|\Gamma^{(2)}\|_\star \le \sqrt{2r^*}\cdot\|\Gamma^{(2)}\|_{\text{fro}}$ Then by (A.15) we obtain that $\|\Theta\|_{\text{fro}} \le 3/\sqrt{2}\cdot\sqrt{r^*}\cdot\lambda$. Finally, by triangle inequality for the nuclear norm,

$$\|\Theta\|_\star = \|\Gamma\|_\star \le \|\Gamma^{(1)}\|_\star + \|\Gamma^{(2)}\|_\star \le 4\cdot\|\Gamma^{(2)}\|_\star \le 4\sqrt{2r^*}\|\Gamma^{(2)}\|_{\text{fro}} = 12r^*\lambda.$$

Thus we conclude the proof of Theorem 3.3.

□

A.3 Proof of Theorem 4.2

Proof. We denote by $\widehat{W}$ the solution of the optimization problem in (4.6). In addition, we let $W^* = \beta^*\beta^{*\top}$. In the following, we establish an upper bound for $\|\widehat{W} - W^*\|_{\text{op}}$.

Since W^* is feasible for the optimization problem in (4.6), we have

$$\langle \widehat{W}, \widetilde{\Sigma} \rangle + \lambda \|\widehat{W}\|_1 \geq \langle W^*, \widetilde{\Sigma} \rangle + \lambda \|W^*\|_1. \tag{A.16}$$

We denote $\Sigma^* = \mathbb{E}[Y \cdot T(X)]$. Note that β^* is the leading eigenvector of Σ^*. Then (A.16) is equivalent to

$$\langle \widehat{W} - W^*, \widetilde{\Sigma} - \Sigma^* \rangle + \lambda \|\widehat{W}\|_1 - \lambda \|W^*\|_1 \geq \langle \Sigma^*, W^* - \widehat{W} \rangle. \tag{A.17}$$

The following Lemma in Vu et al. (2013) establishes an upper bound for the first term on the left-hand side of (A.17).

Lemma 3. *Let $\Omega \in \mathbb{R}^{d \times d}$ be a symmetric matrix and let $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_d$ the eigenvalues of Ω in descending order. For any $\ell \in [d-1]$ such that $\lambda_\ell - \lambda_{\ell+1} > 0$, let $\Pi_\ell \in \mathbb{R}^{d \times d}$ be the projection matrix for the subspace spanned by the eigenvectors of Ω corresponding to $\lambda_1, \ldots, \lambda_\ell$. Then for any $\Lambda \in \mathbb{R}^{d \times d}$ satisfying $0 \preceq \lambda \preceq I_d$ and $\textsc{trace}(\Lambda) = k$, we have*

$$(\lambda_\ell - \lambda_{\ell+1}) \cdot \|\Pi_k - \Lambda\|_{fro}^2 \leq 2\langle \Omega, \Pi_k - \Lambda \rangle.$$

Note that W^* is the projection matrix for the subspace spanned by β^*. Applying Lemma 3 to Σ^* with $\ell = 1$, we have

$$\langle \Sigma^*, W^* - \widehat{W} \rangle \geq C_0/2 \cdot \|\widehat{W} - W^*\|_{\text{fro}}^2, \tag{A.18}$$

where $C_0 > 0$ is defined in (4.2). In addition, by Hölder's inequality, we have

$$\langle \widehat{W} - W^*, \widetilde{\Sigma} - \Sigma^* \rangle \leq \|\widetilde{\Sigma} - \Sigma^*\|_\infty \cdot \|\widehat{W} - W^*\|_1. \tag{A.19}$$

In what follows, we bound $\|\widetilde{\Sigma} - \Sigma^*\|_\infty$.

Lemma 4. *Let $\widetilde{\Sigma}$ be defined in (4.5) and we define $\Sigma^* = \mathbb{E}[Y \cdot T(X)]$. Under Assumption 4.1, for any truncation level $\tau > 0$ in (4.4), with probablity at least $1 - d^{-2}$, we have*

$$\|\widetilde{\Sigma} - \Sigma^*\|_\infty \leq 9M \cdot \tau^{-3} + 2\tau^3 \cdot \log d/n + 2\sqrt{5M \cdot \log d/n}. \tag{A.20}$$

Proof. See §B.3 for a detailed proof. □

By this lemma, if we set $\tau = (1.5Mn/\log d)^{1/6}$, then with probability at least $1 - d^{-1}$,

$$\|\widetilde{\Sigma} - \Sigma^*\|_\infty \leq (2\sqrt{5} + 2\sqrt{6}) \cdot \sqrt{M \log d/n} \leq 10\sqrt{M \log d/n}. \tag{A.21}$$

Thus by setting $\lambda = 10\sqrt{M \log d/n}$ we have $\|\widetilde{\Sigma} - \Sigma^*\|_\infty \le \lambda$ with probability at least $1 - d^{-2}$.

Then combining (A.17), (A.18), and (A.19) we have

$$\lambda \Big(\|\widehat{W} - W^*\|_1 + \|\widehat{W}\|_1 - \|W^*\|_1 \Big) \ge C_0/2 \cdot \|\widehat{W} - W^*\|_{\text{fro}}^2. \tag{A.22}$$

Note that $W^* = \beta^* \beta^{*\top}$ and that β^* is s^*-sparse. We denote the support of W^* by $\mathcal{J}$, which is given by

$$\mathcal{J} = \big\{ (j,k) \in [d] \times [d] \colon \beta_j^* \cdot \beta_k^* \neq 0 \big\}.$$

Then by seperation of the ℓ_1-norm, we have

$$\|\widehat{W}\|_1 = \|\widehat{W}_{\mathcal{J}}\|_1 + \|\widehat{W}_{\mathcal{J}^c}\|_1 \quad \text{and} \quad \|\widehat{W} - W^*\|_1 = \|\widehat{W}_{\mathcal{J}} - W^*_{\mathcal{J}}\|_1 + \|\widehat{W}_{\mathcal{J}^c}\|_1,$$

which implies that

$$\begin{aligned} &\|\widehat{W} - W^*\|_1 + \|\widehat{W}\|_1 - \|W^*\|_1 = \|\widehat{W}_{\mathcal{J}} - W^*_{\mathcal{J}}\|_1 + \|\widehat{W}_{\mathcal{J}}\|_1 - \|W^*_{\mathcal{J}}\|_1 \\ &\le 2\|\widehat{W}_{\mathcal{J}} - W^*_{\mathcal{J}}\|_1 \le 2s^* \|\widehat{W} - W^*\|_{\text{fro}}. \end{aligned} \tag{A.23}$$

Here the last inequality in (A.23) follows from the fact that $|\mathcal{J}| = s^{*2}$. Combining (A.22) and (A.23), we obtain

$$\|\widehat{W} - W^*\|_{\text{fro}} \le 4/C_0 \cdot s^* \lambda. \tag{A.24}$$

Since $\widehat{\beta}$ is the leading eigenvector of $\widehat{W}$, we have $\|\widehat{\beta} - \beta^*\|_2 \le \sqrt{2}\|\widehat{W} - W^*\|_{\text{fro}} \le 4\sqrt{2}/C_0 \cdot s^* \lambda$, which concludes the proof. □

A.4 Proof of Theorem 4.3

Proof. The proof is similar to that of Theorem 4.2. In the case of sparse MIM, we denote $W^* = B^* B^{*\top}$. Note that $\widehat{W}$ is the solution to the optimization problem in (4.7) and that $\widehat{B}$ consists of the top-k eigenvectors of $\widehat{W}$. Then by Corollary 3.2 in Vu et al. (2013), we have

$$\inf_{O \in \mathbb{O}_k} \|\widehat{B} - B^* O\|_{\text{fro}} \le \sqrt{2}\|\widehat{W} - W^*\|_{\text{fro}}. \tag{A.25}$$

In what follows, we derive an upper bound for $\widehat{W} - W^*$. Note that since B^* is orthonormal, $\textsc{trace}(W^*) = k$. Thus W^* is feasible for (4.7), which implies

$$\langle \widehat{W} - W^*, \widetilde{\Sigma} - \Sigma^* \rangle + \lambda \|\widehat{W}\|_1 - \lambda \|W^*\|_1 \ge \langle \Sigma^*, W^* - \widehat{W} \rangle. \tag{A.26}$$

Here we define $\Sigma^* = \mathbb{E}[Y \cdot T(X)]$. Note that W^* is the projection matrix for the subspace spanned by the top-k leading eigenvectors of Σ^*. By Lemma 3 with $\ell = k$, we have

$$\langle \Sigma^*, W^* - \widehat{W} \rangle \geq \rho_0/2 \cdot \|\widehat{W} - W^*\|_{\mathrm{fro}}^2,$$

where ρ_0 is the smallest eigenvalue of $\mathbb{E}[\nabla^2 f(XB^*)]$. Similar to the proof of Theorem 4.2, by Hölder's inequality and (A.26), we have

$$\|\widetilde{\Sigma} - \Sigma^*\|_\infty \cdot \|\widehat{W} - W^*\|_1 + \lambda \|\widehat{W}\|_1 - \lambda \|W^*\|_1 \geq \rho_0/2 \cdot \|\widehat{W} - W^*\|_{\mathrm{fro}}^2. \tag{A.27}$$

By Lemma **??**, if we set $\lambda = 10\sqrt{M \log d/n}$, with probability at least $1 - d^{-2}$, we have

$$\|\widehat{\Sigma} - \Sigma^*\|_\infty \leq \lambda. \tag{A.28}$$

Note that the support of W^* is

$$\mathcal{J} \subseteq \big\{(j,k) \in [d] \times [d] \colon \|B^*_{j\cdot}\|_2 \cdot \|B^*_{k\cdot}\|_2 \neq 0\big\}.$$

Since B^* is s^*-row sparse, $|\mathcal{J}| \leq s^{*2}$. Thus (A.23) also hold for the MIM. Combining (A.27), (A.28), and (A.23), we obtain

$$\|\widehat{W} - W^*\|_{\mathrm{fro}} \leq 4/\rho_0 \cdot s^* \lambda. \tag{A.29}$$

Finally, combining (A.25) and (A.29), we conclude the proof. □

B Proof of Auxiliary Results

B.1 Proof of Lemma 1

Proof. By definition of the loss function L in (3.3), we have

$$\nabla L(\mu\beta^*) = 2\mu\beta^* - \frac{2}{n}\sum_{i=1}^n \widetilde{Y}_i \cdot \widetilde{S}(X_i) = \mathbb{E}\big[2Y_i \cdot S(X_i)\big] - \frac{2}{n}\sum_{i=1}^n \widetilde{Y}_i \cdot \widetilde{S}(X_i).$$

By triangle inequality,

$$\|\nabla L(\mu\beta^*)\|_\infty \leq \Big\|\mathbb{E}\big[2Y \cdot S(X)\big] - \mathbb{E}\big[2\widetilde{Y} \cdot \widetilde{S}(X)\big]\Big\|_\infty + \Big\|\mathbb{E}\big[2\widetilde{Y} \cdot \widetilde{S}(X)\big] - \frac{2}{n}\sum_{i=1}^n \widetilde{Y}_i \cdot \widetilde{S}(X_i)\Big\|_\infty. \tag{B.1}$$

For any $j \in [d]$, by the definition of the truncated response $\widetilde{Y}$ and truncated score $\widetilde{S}$, we obtain

$$\begin{aligned}\left|\mathbb{E}\big[\widetilde{Y}\cdot\widetilde{S}_j(X)\big]-\mathbb{E}\big[Y\cdot S_j(X)\big]\right| &\le \left|\mathbb{E}\Big\{\widetilde{Y}\cdot\big[\widetilde{S}_j(X)-S_j(X)\big]\Big\}\right|+\left|\mathbb{E}\big[(\widetilde{Y}-Y)\cdot S_j(X)\big]\right| \\ &= \underbrace{\left|\mathbb{E}\big[\widetilde{Y}\cdot S_j(X)\cdot\mathbb{1}\{|S_j(X)|>\tau\}\big]\right|}_{a_1}+\underbrace{\left|\mathbb{E}\big[Y\cdot S_j(X)\cdot\mathbb{1}\{|Y|>\tau\}\big]\right|}_{a_2}. \end{aligned} \tag{B.2}$$

By Cauchy-Schwarz inequality, we have

$$\begin{aligned} a_1^2 &\le \mathbb{E}\big[\widetilde{Y}^2 S_j^2(X)\big]\cdot\mathbb{P}\big[|S_j(X)|\ge\tau\big] \\ &\le \sqrt{\mathbb{E}(\widetilde{Y}^4)\cdot\mathbb{E}\big[S_j^4(X)\big]}\cdot\mathbb{E}\big[S_j^4(X)\big]\cdot\tau^{-4} \\ &= M^2\cdot\tau^{-4}, \end{aligned} \tag{B.3}$$

where the second inequality follows from Chebyshev's inequality. Similarly, for a_2 we have

$$\begin{aligned} a_2^2 &\le \mathbb{E}\big[Y^2 S_j^2(X)\big]\cdot\mathbb{P}\big(|Y|\ge\tau\big) \\ &\le \sqrt{\mathbb{E}(\widetilde{Y}^4)\cdot\mathbb{E}\big[S_j^4(X)\big]}\cdot\mathbb{E}(Y^4)\cdot\tau^{-4} \\ &\le M^2\cdot\tau^{-4}. \end{aligned} \tag{B.4}$$

Thus combining (B.2), (B.3), and (B.4), we conclude that

$$\left|\mathbb{E}\big[\widetilde{Y}\cdot\widetilde{S}_j(X)\big]-\mathbb{E}\big[Y\cdot S_j(X)\big]\right| \le a_1+a_2 \le 2M\cdot\tau^{-2}$$

for all $j\in[d]$. Thus choosing $\tau = 2(M\cdot n/\log d)^{1/4}$, we have

$$\left\|\mathbb{E}\big[\widetilde{Y}\cdot\widetilde{S}_j(X)\big]-\mathbb{E}\big[Y\cdot S_j(X)\big]\right\|_\infty \le 1/2\cdot\sqrt{M\cdot\log d/n}. \tag{B.5}$$

Furthermore, under Assumption 4.1, the variance of $\widetilde{Y}\cdot\cdot\widetilde{S}_j(X)$ is bounded by

$$\mathrm{Var}[\widetilde{Y}\cdot\widetilde{S}_j(X)] \le \mathbb{E}[\widetilde{Y}^2\cdot\widetilde{S}_j^2(X)] \le \mathbb{E}[Y^2\cdot S_j^2(X)] \le \sqrt{\mathbb{E}(Y^4)\cdot\mathbb{E}[S_j^4(X)]} \le M.$$

Thus for the second term in (B.1), since $|\widetilde{Y}\cdot\widetilde{S}_j(X)|\le\tau^2$, by the Bernstein inequality in Boucheron et al. (2013) (Theorem 2.10), for any $j\in[d]$ and any $t>0$, we have

$$\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^n \widetilde{Y}_i\cdot\widetilde{S}_j(X_i)-\mathbb{E}\big[\widetilde{Y}\cdot\widetilde{S}_j(X)\big]\right| \ge \sqrt{\frac{2M\cdot t}{n}}+\frac{\tau^2\cdot t}{3n}\right\} \le \exp(-t). \tag{B.6}$$

Taking union bound over $j \in [t]$ in (B.6) yields

$$\mathbb{P}\Big\{\Big\|\frac{1}{n}\sum_{i=1}^{n}\widetilde{Y}_i\cdot\widetilde{S}_j(X_i)-\mathbb{E}\big[\widetilde{Y}\cdot\widetilde{S}_j(X)\big]\Big\|_{\infty}\geq\sqrt{\frac{2M\cdot t}{n}}+\frac{\tau^2\cdot t}{3n}\Big\}\leq\exp(-t+\log d). \tag{B.7}$$

Finally, we plug in $\tau = 2(M\cdot n/\log d)^{1/4}$ and set $t = 3\log d$ in (B.7) to obtain that

$$\Big\|\frac{1}{n}\sum_{i=1}^{n}\widetilde{Y}_i\cdot\widetilde{S}_j(X_i)-\mathbb{E}\big[\widetilde{Y}\cdot\widetilde{S}_j(X)\big]\Big\|_{\infty}\leq(4+\sqrt{6})\sqrt{\frac{M\cdot\log d}{n}} \tag{B.8}$$

with probability at least $1-d^{-2}$. Finally, combining (B.1), (B.5), and (B.8), we conclude the proof. □

B.2 Proof of Lemma 2

Proof. For loss function L defined in (3.3) in the matrix setting, we have

$$\nabla L(\mu\beta^*)=2\mu\beta^*-\frac{2}{\kappa\cdot n}\sum_{i=1}^{n}\psi\big[\kappa\cdot Y_i\cdot S(X_i)\big]=2\mathbb{E}[Y\cdot S(X)]-\frac{2}{\kappa\cdot n}\sum_{i=1}^{n}\psi\big[\kappa\cdot Y_i\cdot S(X_i)\big]. \tag{B.9}$$

Here the last equality follows from the generalized Stein's identity. In the sequel, we apply results in Minsker (2016) to bound $\|\nabla L(\mu\beta^*)\|_{\mathrm{op}}$. To begin with, we first consider the operator norm of $\mathbb{E}[Y^2\cdot S(X)S(X)^\top]\in\mathbb{R}^{d_1\times d_2}$ and $\mathbb{E}[Y^2\cdot S(X)^\top S(X)]\in\mathbb{R}^{d_2\times d_2}$. For notational simplicity, we denote by $S_{j,\cdot}(\cdot)\in\mathbb{R}^{d_2}$ $S_{\cdot,k}(\cdot)\in\mathbb{R}^{d_1}$ the j-th row and k-the column of the score function $S(\cdot)$, respectively. For any $u\in\mathcal{S}^{d_1-1}$, by Cauchy-Schwarz inequality we have

$$\mathbb{E}[Y^2\cdot u^\top S(X)S(X)^\top u]=\sum_{k=1}^{d_2}\mathbb{E}\big\{[Y^2\cdot S_{\cdot,k}(X)^\top u]^2\big\}\leq d_2\cdot\sqrt{\mathbb{E}(Y^4)\cdot\mathbb{E}\big\{[S_{\cdot,1}(X)^\top u]^4\big\}}, \tag{B.10}$$

where we use the fact that the entries of $S(X)$ are i.i.d. Since $\mathbb{E}[S_{ij}(X)]=0$ and $\mathbb{E}[S_{ij}^4(X)]\leq M$, by Cauchy-Schwarz inequality we obtain that

$$\begin{aligned}\mathbb{E}\big\{[S_{\cdot,1}(X)^\top u]^4\big\}&=\sum_{j_1=1}^{d}\sum_{j_2=1}^{d}\mathbb{E}[S_{j_1,1}(X)^2\cdot S_{j_2,1}^2(X)]\cdot u_{j_i}^2u_{j_2}^2\\&\leq\sum_{j_1=1}^{d}\sum_{j_2=1}^{d}\sqrt{\mathbb{E}[S_{j_1,1}^4(X)]\cdot\mathbb{E}[S_{j_2,1}^4(X)]}\cdot u_{j_i}^2u_{j_2}^2\leq M\sum_{j_1=1}^{d}\sum_{j_2=1}^{d}u_{j_i}^2u_{j_2}^2=M.\end{aligned} \tag{B.11}$$

Thus combining (B.10) and (B.11) we obtain that

$$\mathbb{E}[Y^2 \cdot u^\top S(X) S(X)^\top u] \le d_2 \cdot M,$$

which implies that $\|\mathbb{E}[Y^2 \cdot S(X)S(X)^\top]\|_{\mathrm{op}} \le d_2 \cdot M$. Similarly, we obtain $\|\mathbb{E}[Y^2 \cdot S(X)^\top S(X)]\|_{\mathrm{op}} \le d_1 \cdot M$. Thus by Corollary 3.1 in Minsker (2016), we have

$$\begin{aligned}
&\mathbb{P}\biggl\{\biggl\|\frac{1}{\kappa \cdot n}\sum_{i=1}^{n}\psi\bigl[\kappa \cdot Y_i \cdot S(X_i)\bigr] - \mathbb{E}[Y \cdot S(X)]\biggr\|_{\mathrm{op}} \ge \frac{t}{\sqrt{n}}\biggr\} \\
&\quad \le 2(d_1+d_2)\exp\bigl[-\kappa t\sqrt{n} + \kappa^2(d_1+d_2)M/2\bigr] \qquad \text{(B.12)}
\end{aligned}$$

for any $t > 0$ and $\kappa > 0$. We set $\kappa = 2\sqrt{n \cdot \log(d_1+d_2)}/\sqrt{(d_1+d_2)M}$ and $t = \sqrt{(d_1+d_2)M} \cdot s$ in (B.12), which implies that

$$\begin{aligned}
&\mathbb{P}\biggl\{\biggl\|\frac{1}{\kappa \cdot n}\sum_{i=1}^{n}\psi\bigl[\kappa \cdot Y_i \cdot S(X_i)\bigr] - \mathbb{E}[Y \cdot S(X)]\biggr\|_{\mathrm{op}} \ge \sqrt{\frac{(d_1+d_2)M}{n}} \cdot s\biggr\} \\
&\quad \le 2(d_1+d_2)\exp\bigl[-2\sqrt{\log(d_1+d_2)} \cdot s + 2 \cdot \log(d_1+d_2)\bigr]. \qquad \text{(B.13)}
\end{aligned}$$

Now we set $s = 3 \cdot \sqrt{\log(d_1+d_2)}$, which implies that the right-hand side of (B.13) is less than

$$\begin{aligned}
&2(d_1+d_2)\exp\bigl[-6\log(d_1+d_2) + 2 \cdot \log(d_1+d_2)\bigr] \\
&\quad \le (d_1+d_2)^2 \cdot \exp\bigl[-4 \cdot \log(d_1+d_2)\bigr] = (d_1+d_2)^{-2}.
\end{aligned}$$

Therefore, combining (B.9) and (B.13) we conclude that

$$\|\nabla L(\mu\beta^*)\|_{\mathrm{op}} \le 6\sqrt{(d_1+d_2) \cdot M/n} \cdot \sqrt{\log(d_1+d_2)}$$

with probability at least $1 - (d_1+d_2)^{-2}$, which concludes the proof.

□

B.3 Proof of Lemma 4

Proof. By triangle inequailty, we have

$$\|\widetilde{\Sigma} - \Sigma^*\|_\infty \le \|\widetilde{\Sigma} - \mathbb{E}\widetilde{\Sigma}\|_\infty + \|\mathbb{E}\widetilde{\Sigma} - \Sigma^*\|_\infty. \qquad \text{(B.14)}$$

In the sequel, we bound the second term on the right-hand side of (B.14), which controls the bias of truncation. For each $j, k \in [d]$, we have

$$\begin{aligned}\left|\mathbb{E}\widetilde{\Sigma}_{jk} - \Sigma^*_{jk}\right| &\le \left|\mathbb{E}\big[\widetilde{Y} \cdot \widetilde{T}_{jk}(X)\big] - \mathbb{E}\big[Y \cdot T_{jk}(X)\big]\right| \\ &\le \left|\mathbb{E}\big\{\widetilde{Y} \cdot \big[\widetilde{T}_{jk}(X) - T_{jk}(X)\big]\big\}\right| + \left|\mathbb{E}\big[(\widetilde{Y} - Y) \cdot T_{jk}(X)\big]\right|. \end{aligned} \tag{B.15}$$

For the first term in (B.15), note that $\widetilde{T}_{jk}(X) - T_{jk}(X) = T_{jk}(X) \cdot \mathbb{1}\{|T_{jk}(X)| \ge \tau^2\}$. Then by Cauchy-Schwarz inequality we have

$$\begin{aligned}\left|\mathbb{E}\left\{\widetilde{Y} \cdot \left[\widetilde{T}_{jk}(X) - T_{jk}(X)\right]\right\}\right|^2 &= \left|\mathbb{E}\left[\widetilde{Y} \cdot T_{jk}(X) \cdot \mathbb{1}\{|T_{jk}(X)| \ge \tau^2\}\right]\right|^2 \\ &\le \mathbb{E}\left[\widetilde{Y}^2 \cdot T^2_{jk}(X)\right] \cdot \mathbb{P}\left[|T_{jk}(X)| \ge \tau^2\right]. \end{aligned} \tag{B.16}$$

Furthermore, by Hölder's inequality, we have

$$\mathbb{E}\left[\widetilde{Y}^2 \cdot T^2_{jk}(X)\right] \le \left[\mathbb{E}(\widetilde{Y}^6)\right]^{1/3} \cdot \left\{\mathbb{E}\left[|T_{jk}(X)|^3\right]\right\}^{2/3} \le \left[\mathbb{E}(Y^6)\right]^{1/3} \left\{\mathbb{E}\left[|T^3_{jk}(X)|\right]\right\}^{2/3}. \tag{B.17}$$

If $j \ne k$, by the definition of $T(x)$ in (4.1), we have $T_{jk}(x) = S_j(x) \cdot S_k(x)$, $\forall x \in \mathbb{R}^d$. Then by Cauchy-Schwarz inequality, we have

$$\mathbb{E}\left[|T^3_{jk}(X)|\right] = \mathbb{E}\left[|S_j(X)|^3 \cdot |S_k(X)|^3\right] \le \sqrt{\mathbb{E}[S^6_j(X)] \cdot \mathbb{E}[S^6_k(X)]} = \mathbb{E}[S^6_j(X)]. \tag{B.18}$$

In addition, if $j = k$, by (4.1), $T_{jj}(x) = S^2_j(x) - s_1(x_j)$. Since $(a+b)^3 \le 4(a^3 + b^3)$ for any $a, b > 0$, we have

$$\mathbb{E}\left[|T^3_{jj}(X)|\right] \le 4\mathbb{E}[S^6_j(X)] + 4\mathbb{E}\left[|s^3_1(X_j)|\right]. \tag{B.19}$$

Moreover, by (B.16), (B.17), and the Markov's inequality

$$\mathbb{P}\left[|T_{jk}(X)| \ge \tau^2\right] \le \mathbb{E}\left[|T^3_{jk}(X)|\right] \cdot \tau^{-6},$$

we further have

$$\left|\mathbb{E}\big\{\widetilde{Y} \cdot \big[\widetilde{T}_{jk}(X) - T_{jk}(X)\big]\big\}\right|^2 \le \left[\mathbb{E}(Y^6)\right]^{1/3} \cdot \left\{\mathbb{E}\left[|T^3_{jk}(X)|\right]\right\}^{5/3} \cdot \tau^{-6} \le 32M^2 \cdot \tau^{-6}. \tag{B.20}$$

Here the last inequality follows from combining Assumption 4.1, (B.18), and (B.19).

Similarly, for the second term in (B.15), by the Hölder's inequality and the Markov's inequality we obtain

$$\begin{aligned}\left|\mathbb{E}\left[(\widetilde{Y} - Y) \cdot T_{jk}(X)\right]\right|^2 &\le \left[\mathbb{E}(Y^6)\right]^{1/3} \cdot \left\{\mathbb{E}\left[|T^3_{jk}(X)|\right]\right\}^{2/3} \cdot \mathbb{P}(|Y| \ge \tau) \\ &\le \left[\mathbb{E}(Y^6)\right]^{4/3} \cdot \left\{\mathbb{E}\left[|T^3_{jk}(X)|\right]\right\}^{2/3} \cdot \tau^{-6} \le 4M^2 \cdot \tau^{-6}. \end{aligned} \tag{B.21}$$

Therefore, combining (B.15), (B.20), and (B.21) , we obtain

$$\|\mathbb{E}\widetilde{\Sigma} - \Sigma^*\|_\infty \leq 9M \cdot \tau^{-3}. \tag{B.22}$$

In what follows, we give a high-probability bound on $\|\widetilde{\Sigma} - \mathbb{E}\widetilde{\Sigma}\|_\infty$ using concentration inequalities, which combined with B.22, concludes the proof.

For any $j, k \in [d]$, note that $|\widetilde{Y} \cdot \widetilde{T}_{jk}(X)| \leq \tau^3$. In addition, by assumption 4.1, its variance is bounded by

$$\mathrm{Var}\left[\widetilde{Y} \cdot \widetilde{T}_{jk}(X)\right] \leq \mathbb{E}\left[Y^2 \cdot T_{jk}^2(X)\right] \leq \left[\mathbb{E}(Y^6)\right]^{1/3} \cdot \left\{\mathbb{E}\left[|T_{jk}^3(X)|\right]\right\}^{2/3} \leq 2M.$$

Now we apply the Bernstein's inequality (Boucheron et al., 2013) (Theorem 2.10) to $\{\widetilde{Y}_i \cdot \widetilde{T}_{jk}(X_i)\}_{i\in[n]}$ and obtain that

$$\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n} \widetilde{Y}_i \cdot \widetilde{T}_{jk}(X_i) - \mathbb{E}\left[\widetilde{Y} \cdot \widetilde{T}_{jk}(X)\right]\right| \geq \sqrt{\frac{4M \cdot t}{n}} + \frac{\tau^3 \cdot t}{3n}\right\} \leq 2\exp(-t). \tag{B.23}$$

Taking a union bound over $j, k \in [d]$ in (B.23), we obtain that

$$\mathbb{P}\left[\|\widetilde{\Sigma} - \mathbb{E}\widetilde{\Sigma}\|_\infty \geq \sqrt{4M \cdot t/n} + \tau^3 \cdot t/(3n)\right] \leq 2\exp(-t + 2\log d). \tag{B.24}$$

Choosing $t = 5\log d$ in (B.24), we have

$$\|\widetilde{\Sigma} - \mathbb{E}\widetilde{\Sigma}\|_\infty \leq 2\sqrt{5M\log d/n} + 2\tau^3 \cdot \log d/n \tag{B.25}$$

with probablity at least $1 - d^{-2}$. Finally, combining (B.22) and (B.25), we conclude the proof of Lemma 4. □

References

Albert Ai, Alex Lapanowski, Yaniv Plan, and Roman Vershynin. One-bit compressed sensing with non-gaussian measurements. *Linear Algebra and its Applications*, 441:222–239, 2014.

Pierre Alquier and Gérard Biau. Sparse single-index model. *The Journal of Machine Learning Research*, 14(1):243–280, 2013.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

Petros T Boufounos and Richard G Baraniuk. 1-bit compressive sensing. In *Information Sciences and Systems, 2008. CISS 2008. 42nd Annual Conference on*, pages 16–21. IEEE, 2008.

Peter Bühlmann and Sara van de Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

T Tony Cai, Xiaodong Li, and Zongming Ma. Optimal rates of convergence for noisy sparse phase retrieval via thresholded wirtinger flow. *arXiv preprint arXiv:1506.03382*, 2015.

Emmanuel J Candès, Thomas Strohmer, and Vladislav Voroninski. Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming. *Communications on Pure and Applied Mathematics*, 66(8):1241–1274, 2013.

Emmanuel J Candès, Yonina C Eldar, Thomas Strohmer, and Vladislav Voroninski. Phase retrieval via matrix completion. *SIAM Review*, 57(2):225–251, 2015.

Olivier Catoni et al. Challenging the empirical mean and empirical variance: a deviation study. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 48(4):1148–1185, 2012.

Han Chen, Garvesh Raskutti, and Ming Yuan. Non-convex projected gradient descent for generalized low-rank tensor regression. *arXiv preprint arXiv:1611.10349*, 2016.

Xin Chen, Changliang Zou, and R Dennis Cook. Coordinate-independent sparse sufficient dimension reduction and variable selection. *The Annals of Statistics*, 38(6):3696–3723, 2010.

Mark A Davenport, Yaniv Plan, Ewout van den Berg, and Mary Wootters. 1-bit matrix completion. *Information and Inference*, 3(3):189–223, 2014.

J. Fan, J. Lv, and L. Qi. Sparse high-dimensional models in economics. *Annual review of economics*, 3(1):291–317, 2011.

Jianqing Fan, Weichen Wang, and Ziwei Zhu. Robust low-rank matrix recovery. *arXiv preprint arXiv:1603.08315*, 2016.

Shmuel Friedland and Lek-Heng Lim. Nuclear norm of higher-order tensors. *arXiv preprint arXiv:1410.6072*, 2014.

Larry Goldstein, Stanislav Minsker, and Xiaohan Wei. Structured signal recovery from non-linear and heavy-tailed measurements. *arXiv preprint arXiv:1609.01025*, 2016.

David Gross, Felix Krahmer, and Richard Kueng. A partial derandomization of phaselift using spherical designs. *Journal of Fourier Analysis and Applications*, 2015.

Aaron K Han. Non-parametric analysis of a generalized regression model: the maximum rank correlation estimator. *Journal of Econometrics*, 35(2-3):303–316, 1987.

Joel L Horowitz. *Semiparametric and nonparametric methods in econometrics*, volume 12. Springer, 2009.

Majid Janzamin, Hanie Sedghi, and Anima Anandkumar. Score function features for discriminative learning: Matrix and tensor framework. *arXiv preprint arXiv:1412.2863*, 2014.

Majid Janzamin, Hanie Sedghi, and Anima Anandkumar. Beating the perils of non-convexity: Guaranteed training of neural networks using tensor methods. *arXiv preprint arXiv:1506.08473*, 2015.

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

Sham M Kakade, Varun Kanade, Ohad Shamir, and Adam Kalai. Efficient learning of generalized linear and single index models with isotonic regression. In *Advances in Neural Information Processing Systems*, pages 927–935, 2011.

Adam Tauman Kalai and Ravi Sastry. The isotron algorithm: High-dimensional isotonic regression. In *Conference on Learning Theory*, 2009.

Robert Krauthgamer, Boaz Nadler, Dan Vilenchik, et al. Do semidefinite relaxations solve sparse pca up to the information limit? *The Annals of Statistics*, 43(3):1300–1322, 2015.

Guillaume Lecué and Shahar Mendelson. Minimax rate of convergence and the performance of empirical risk minimization in phase retrieval. *Electron. J. Probab*, 20(57):1–29, 2015.

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

Ker-Chau Li. On principal Hessian directions for data visualization and dimension reduction: Another application of Stein's lemma. *Journal of the American Statistical Association*, 87 (420):1025–1039, 1992.

Ker-Chau Li and Naihua Duan. Regression analysis under link violation. *The Annals of Statistics*, 17(3):1009–1052, 1989.

Xiaodong Li and Vladislav Voroninski. Sparse signal recovery from quadratic measurements via convex programming. *SIAM Journal on Mathematical Analysis*, 45(5):3019–3033, 2013.

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

Qian Lin, Xinran Li, Dongming Huang, and Jun S Liu. On the optimality of sliced inverse regression in high dimensions. *arXiv preprint arXiv:1701.06009*, 2017.

Po-Ling Loh and Martin J Wainwright. Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *Journal of Machine Learning Research*, 16:559–616, 2015.

Geoffrey McLachlan and David Peel. *Finite mixture models*. John Wiley & Sons, 2004.

Shahar Mendelson. Learning without concentration. In *Proceedings of The 27th Conference on Learning Theory*, pages 25–39, 2014.

Stanislav Minsker. Sub-gaussian estimators of the mean of a random matrix with heavy-tailed entries. *arXiv preprint arXiv:1605.07129*, 2016.

Cun Mu, Bo Huang, John Wright, and Donald Goldfarb. Square deal: Lower bounds and improved relaxations for tensor recovery. In *Proceedings of The 31st International Conference on Machine Learning*, pages 73–81, 2014.

Sahand N. Negahban, Pradeep Ravikumar, Martin J. Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 11 2012.

Roberto Imbuzeiro Oliveira. The lower tail of random quadratic forms, with applications to ordinary least squares and restricted eigenvalue properties. *arXiv preprint arXiv:1312.2903*, 2013.

Yaniv Plan and Roman Vershynin. The generalized lasso with non-linear observations. *IEEE Transactions on information theory*, 62(3):1528–1537, 2016.

Peter Radchenko. High dimensional single index models. *Journal of Multivariate Analysis*, 139:266–282, 2015.

Robert P Sherman. The limiting distribution of the maximum rank correlation estimator. *Econometrica: Journal of the Econometric Society*, 61(1):123–137, 1993.

C. Stein. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory*. The Regents of the University of California, 1972.

Charles Stein, Persi Diaconis, Susan Holmes, Gesine Reinert, et al. Use of exchangeable pairs in the analysis of simulations. In *Stein's Method*. Institute of Mathematical Statistics, 2004.

Ju Sun, Qing Qu, and John Wright. A geometric analysis of phase retrieval. *arXiv preprint arXiv:1602.06664*, 2016.

Christos Thrampoulidis, Ehsan Abbasi, and Babak Hassibi. Lasso with non-linear measurements is equivalent to one with linear measurements. *Advances in Neural Information Processing Systems*, 2015.

Roman Vershynin. Estimation in high dimensions: a geometric perspective. In *Sampling theory, a renaissance*, pages 3–66. Springer, 2015.

Vincent Q Vu, Juhee Cho, Jing Lei, and Karl Rohe. Fantope projection and selection: A near-optimal convex relaxation of sparse pca. In *Advances in neural information processing systems*, pages 2670–2678, 2013.

Irène Waldspurger, Alexandre dAspremont, and Stéphane Mallat. Phase recovery, max-cut and complex semidefinite programming. *Mathematical Programming*, 149(1-2):47–81, 2015.

Tengyao Wang, Quentin Berthet, Richard J Samworth, et al. Statistical and computational trade-offs in estimation of sparse principal components. *The Annals of Statistics*, 44(5): 1896–1930, 2016.

Zhaoran Wang, Huanran Lu, and Han Liu. Tighten after relax: Minimax-optimal sparse pca in polynomial time. In *Advances in neural information processing systems*, pages 3383–3391, 2014.

Zhuoran Yang, Zhaoran Wang, Han Liu, Yonina C Eldar, and Tong Zhang. Sparse nonlinear regression: Parameter estimation and asymptotic inference. *International Conference on Machine Learning*, 2015.

Ming Yuan, Ding-Xuan Zhou, et al. Minimax optimal rates of estimation in high dimensional additive models. *The Annals of Statistics*, 44(6):2564–2593, 2016.

Lixing Zhu, Baiqi Miao, and Heng Peng. On sliced inverse regression with high-dimensional covariates. *Journal of the American Statistical Association*, 101(474):630–643, 2006.