Journal of Machine Learning Research 20 (2019) 1–44 Submitted 10/18; Revised 5/19; Published 10/19

High-dimensional Varying Index Coefficient Models via Stein's Identity

Sen Na SENNA@UCHICAGO.EDU
Department of Statistics
University of Chicago
Chicago, IL 60637, USA

Zhuoran Yang ZY6@PRINCETON.EDU
Department of Operations Research and Financial Engineering
Princeton University
Princeton, NJ 08544, USA

Zhaoran Wang ZHAORAN.WANG@NORTHWESTERN.EDU
Department of Industrial Engineering and Management Sciences
Northwestern University
Evanston, IL 60208, USA

Mladen Kolar MKOLAR@CHICAGOBOOTH.EDU
The University of Chicago Booth School of Business
Chicago, IL 60637, USA

Editor: Kenji Fukumizu

Abstract

We study the parameter estimation problem for a varying index coefficient model in high dimensions. Unlike the most existing works that iteratively estimate the parameters and link functions, based on the generalized Stein's identity, we propose computationally efficient estimators for the high-dimensional parameters without estimating the link functions. We consider two different setups where we either estimate each sparse parameter vector individually or estimate the parameters simultaneously as a sparse or low-rank matrix. For all these cases, our estimators are shown to achieve optimal statistical rates of convergence (up to logarithmic terms in the low-rank setting). Moreover, throughout our analysis, we only require the covariate to satisfy certain moment conditions, which is significantly weaker than the Gaussian or elliptically symmetric assumptions that are commonly made in the existing literature. Finally, we conduct extensive numerical experiments to corroborate the theoretical results.

Keywords: high-dimensional estimation, semiparametric modeling, Stein's identity, varying index coefficient model

1. Introduction

We consider the problem of estimating parameters in a high-dimensional varying index coefficient model with following form

$$y = \sum_{j=1}^{d_2} z_j \cdot f_j(\langle \boldsymbol{x}, \boldsymbol{\beta}_j^\star \rangle) + \epsilon, \tag{1}$$

where y is response variable, $\boldsymbol{x} = (x_1, \ldots, x_{d_1})^\top \in \mathbb{R}^{d_1}$ and $\boldsymbol{z} = (z_1, \ldots, z_{d_2})^\top \in \mathbb{R}^{d_2}$ are given covariates, and ϵ is random noise with $\mathbb{E}[\epsilon \mid \boldsymbol{x}, \boldsymbol{z}] = 0$. For $j \in [d_2]$[1], $\boldsymbol{\beta}_j^\star = (\beta_{j1}^\star, \ldots, \beta_{jd_1}^\star)^\top$ are the coefficient vectors, that is parameters, which vary with different covariates z_j, and $f_j(\cdot)$ are unknown nonparametric link functions. For identification purposes, we can always permute $\boldsymbol{\beta}_j^\star$ and multiply by a scalar such that

$$\boldsymbol{\beta}_j^\star \in \{\boldsymbol{\beta} \in \mathbb{R}^{d_1} : \|\boldsymbol{\beta}\|_2 = 1 \text{ and } \beta_1 > 0\}, \quad j = 1, \ldots, d_2. \tag{2}$$

All further restrictions on parameters will only be considered under (2).

Model (1) has been introduced by Ma and Song (2015) as a flexible generalization of a number of well studied semiparametric statistical models (see also Xue and Wang, 2012). When $z_j = 1$ for all $j \in [d_2]$, the model reduces to the additive single-index model (Chen, 1991; Carroll et al., 1997), which can also be viewed as a two-layer neural network with d_2 hidden nodes. When $d_1 = 1$ and $\boldsymbol{\beta}_j^\star = 1$ for $j = 1, \ldots, d_2$, the model (1) reduces to the varying coefficient model proposed in Cleveland et al. (1991) and Hastie and Tibshirani (1993), with wide applications in scientific areas such as economics and medical science (Fan and Zhang, 2008). Varying coefficient models allow the coefficients of $\boldsymbol{z}$ to be smooth functions of $\boldsymbol{x}$, thus incorporating nonlinear interactions between $\boldsymbol{x}$ and $\boldsymbol{z}$. Model (1) is also easily interpreted in real applications because it inherits features from both single-index model and varying coefficient model, while being able to capture complex multivariate nonlinear structure.

Our focus is on the case when the dimension of $\boldsymbol{x}$ is high, which makes estimation of the coefficients difficult. Existing procedures estimate the unknown functions and coefficients iteratively. First, with the signal parameters $\{\boldsymbol{\beta}_j^\star\}_{j\in[d_2]}$ fixed, one estimates the functions $\{f_j(\cdot)\}_{j\in[d_2]}$ using a nonparametric method, such as local polynomial estimator. Next, using the estimated link functions, one re-estimates the coefficients. While the global minimizer has desirable properties (see Xue and Wang (2012) and Ma and Song (2015) and the references therein), the loss function is usually nonconvex and it is computationally intractable to obtain the global optima. For high-dimensional single-index models ($d_2 = 1$ and $\boldsymbol{z} = 1$), when the distribution of $\boldsymbol{x}$ is known, the signal parameter can be estimated directly by fitting Lasso (Tibshirani, 1996). Such an estimator is shown to achieve minimax-optimal statistical rate of convergence (Plan and Vershynin, 2016; Plan et al., 2017). Thus, the following question naturally arises:

Is it possible to estimate signal parameters $\{\boldsymbol{\beta}_j^\star\}_{j\in[d_2]}$ *in* (1) *with both statistical accuracy and computational efficiency?*

In this work, we provide a positive answer to above question. Specifically, we focus on the problem of estimating the parameter matrix $\boldsymbol{B}^\star = (\boldsymbol{\beta}_1^\star, \ldots, \boldsymbol{\beta}_{d_2}^\star) \in \mathbb{R}^{d_1 \times d_2}$ in the high-dimensional setting where the sample size is much smaller than $d_1 \times d_2$ and $\boldsymbol{B}^\star$ is either sparse or low-rank. We utilize the score functions and the generalized Stein's identity (Stein, 1972; Stein et al., 2004) to estimate the unknown coefficients through a regularized least-square regression problem, without learning the unknown functions $\{f_j(\cdot)\}_{j\in[d_2]}$. We prove that the estimators achieve (near) optimal statistical rates of convergence under weak moment conditions, which make our procedure suitable for heavy-tailed data, using a careful truncation argument. Finally, our estimator can be computed as a solution to a convex optimization problem.

1. For any integer d, we denote $[d] = \{1, 2, ..., d\}$.

Main Contributions. Our contributions are three-fold. First, we propose a computationally efficient estimation procedure for the varying index coefficient model in high dimensions. Different from existing work, our approach does not need to estimate the unknown functions $\{f_j\}_{j\in[d_2]}$. Second, when $\boldsymbol{B}^\star$ is sparse, we prove that the proposed estimator achieves the optimal statistical rate of convergence, while when $\boldsymbol{B}^\star$ is low-rank, our estimator is shown to be near-optimal (up to a logarithmic factor). Finally, we provide thorough numerical experiments with both synthetic and real data to back up our theories.

Related Work. There is a plethora of literature on the varying coefficient model, first proposed in Cleveland et al. (1991) and Hastie and Tibshirani (1993), where the coefficients are modeled as nonparametric functions of $\boldsymbol{x}$. See Fan and Zhang (2008) for a detailed review. Xia and Li (1999), Fan et al. (2003), and Xue and Wang (2012) considered model in (1) with $\boldsymbol{\beta}_j^\star = \boldsymbol{\beta}^\star$ for all $j \in [d_2]$ and estimated it with standard nonparametric techniques. Ma and Song (2015) proposed model (1) and developed a profile least-square approach to estimate the coefficients. Unfortunately, the estimator is defined as a solution to a constrained optimization problem with nonconvex objective function that can be hard to globally optimize in practice. This should be contrasted to estimators that are based on solving convex optimization problems.

Another related line of research is on the high-dimensional single-index model (SIM) with sparse coefficient vector, which is a special case of model (1). Most of the existing results require either knowing the distribution of $\boldsymbol{x}$ or strong assumptions on the link functions. Specifically, Thrampoulidis et al. (2015), Neykov et al. (2016), Plan and Vershynin (2016), and Plan et al. (2017) all showed that when $\boldsymbol{x}$ is a standard Gaussian and the link function satisfies certain conditions, Lasso estimators could also work for SIM with the same theoretical guarantee as if the link function is not present. To relax the Gaussian assumption, Goldstein et al. (2018) proposed a modified Lasso-type estimator when $\boldsymbol{x}$ has an elliptically symmetric distribution. Moreover, using the generalized Stein's identity, Yang et al. (2017) proposed a soft-thresholding estimator for SIM when the distribution of $\boldsymbol{x}$ is known, and Na and Kolar (2018) proposed estimators for single-index volatility model. Our work can be viewed as the extension of this work from two aspects. First, as mentioned before, our model in (1) includes the single-index model considered in Yang et al. (2017) as a special case, which can be recovered by letting $d_2 = 1$ and $\boldsymbol{z} = 1$. In comparison, our model contains d_2 unknown signal vectors $\{\boldsymbol{\beta}_j^\star\}_{j\in[d_2]}$, where d_2 itself can be large compared to the sample size. As we will show later, when the parameter matrix $\boldsymbol{B}^\star$ has a certain low-dimensional structure, such as being sparse or low rank, we construct efficient estimators that enjoy sharp statistical rates of convergence even when d_2 is large and $\{z_j\}_{j\in[d_2]}$ are dependent, heavy-tailed. Second, from the methodology perspective, our estimators are based on combining the first-order Stein's identity and the dependence structure of the covariate $\boldsymbol{z}$, which is a novel observation in index models and generalizes the method in Yang et al. (2017) for SIM. In particular, we extract the information of the parameter matrix $\boldsymbol{B}^\star$ via multiplying the response y by the score function of $\boldsymbol{x}$ and the covariate $\boldsymbol{z}$ adjusted by its precision matrix (see (9)). When $\boldsymbol{z}$ has independent components, our procedure for estimating each single sparse vector $\boldsymbol{\beta}_j^\star$ is equivalent to fitting d_2 SIMs using the procedure proposed in Yang et al. (2017). However, even having independent $\{z_j\}_{j\in[d_2]}$, their method fails when estimating $\boldsymbol{B}^\star$ as a whole with either low-rank or sparse structure, since they only have one coefficient and cannot deal with varying coefficient models without account-

ing for $\boldsymbol{z}$ in Steins identity. Furthermore, to handle dependence structure of $\boldsymbol{z}$ in general, we construct suitable precision matrix estimators for heavy-tailed variable in either low or high dimensions based on the CLIME procedure (Cai et al., 2011) and the soft truncation technique, which is of independent interests. Besides aforementioned related literature, a sequence of work (Zhu et al., 2006; Jiang and Liu, 2014; Zhang et al., 2017; Lin et al., 2017, 2018) applied the sliced inverse regression (SIR) technique on high-dimensional SIM, which is generalized from Li (1991). But all these works require the distribution of $\boldsymbol{x}$ to be Gaussian or elliptical. To resolve this limitation, Babichev and Bach (2018) incorporated SIR with both first-order and second-order score function when fitting a low-dimensional index model, while the high-dimensional analysis is not included.

Furthermore, our work is also related to the study of additive index model, which is more challenging than (1), and there is very much work in this direction. Most existing work focuses on estimating the signal parameters and the link functions together in the low-dimensional setting. See, for example, Yuan (2011), Wang et al. (2015), and Chen and Samworth (2016). When the covariate is Gaussian and the link functions are known, Sedghi et al. (2016) proposed to estimate the signal parameters via tensor decomposition. These works are not comparable with ours as we consider a different model and our goal is to efficiently estimate the high-dimensional parameters.

Lastly, our estimation methodology utilizes the generalized Stein's identity (Stein et al., 2004), which extends the well-known Stein's identity for Gaussian distribution (Stein, 1972) to general distributions whose density satisfies certain regularity conditions. This identity is widely applied in probability, statistics, and machine learning. We point the reader to Chen et al. (2011), Chwialkowski et al. (2016), Liu et al. (2016), Liu and Wang (2016), and Liu et al. (2018) for recent applications.

Notations: Throughout the paper, we use boldface letters, $\boldsymbol{v}, \boldsymbol{V}$, to denote vector or matrix and their elements will be denoted as v_i, V_{ij}. For any vector $\boldsymbol{v}$ and $p \geqslant 1$, $\|\boldsymbol{v}\|_p$ is the vector ℓ_p-norm, with the usual extension $\|\boldsymbol{v}\|_0 = |\mathrm{supp}(\boldsymbol{v})| = |\{i : v_i \neq 0\}|$. Given a matrix $\boldsymbol{V} \in \mathbb{R}^{m\times n}$, we let $\|\boldsymbol{V}\|_p$ be the induced p-norm. We denote the nuclear norm and Frobenius norm as $\|\boldsymbol{V}\|_*$ and $\|\boldsymbol{V}\|_F$, respectively. We also define $\|\boldsymbol{V}\|_{p,q} = \big(\sum_{j=1}^n (\sum_{i=1}^m |V_{ij}|^p)^{q/p}\big)^{1/q}$, which is basically computing the vector ℓ_p-norm for each column and then computing the ℓ_q norm for those n numbers. We also define $\|\boldsymbol{V}\|_{\max} = \|\boldsymbol{V}\|_{\infty,\infty}$ and $\mathrm{supp}(\boldsymbol{V}) = \{(i,j) : V_{ij} \neq 0\}$. For two matrices $\boldsymbol{V}, \boldsymbol{U}$ with the same dimension, we let $\langle \boldsymbol{V}, \boldsymbol{U}\rangle = \mathrm{trace}(\boldsymbol{V}^T\boldsymbol{U}) = \sum_{i,j=1}^n V_{ij}U_{ij}$. When presenting the result, we use $a \lesssim b$ ($\gtrsim$) to denote $a \leqslant c \cdot b$ ($\geqslant$) for some constant c that we are less interested in. Also, we have $a \asymp b \Leftrightarrow a \lesssim b$ and $a \gtrsim b$. Given a threshold λ, we define the soft-thresholding function $\mathcal{T}_\lambda(\cdot)$ as follows: (i) when $\boldsymbol{a} \in \mathbb{R}^d$, we let $\mathcal{T}_\lambda(\boldsymbol{a}) \in \mathbb{R}^d$ with $[\mathcal{T}_\lambda(\boldsymbol{a})]_i = (1 - \lambda/|a_i|)_+ a_i$; (ii) when $\boldsymbol{A} \in \mathbb{R}^{d_1\times d_2}$, suppose its singular value decomposition can be written as $\boldsymbol{A} = \boldsymbol{U}\mathrm{diag}(\boldsymbol{\sigma})\boldsymbol{V}^T$, then we let $\mathcal{T}_\lambda(\boldsymbol{A}) = \boldsymbol{U}\mathrm{diag}(\hat{\boldsymbol{\sigma}})\boldsymbol{V}^T$ where $\hat{\sigma}_i = (\sigma_i - \lambda)_+$.

2. Estimation via the Generalized Stein's Identity

In this section, we present the main idea for estimating coefficients in model (1). Our estimator relies on the generalized Stein's identity (Stein et al., 2004), which we state next.

Theorem 1 (Generalized Stein's identity, Stein et al. (2004)) *Suppose $\boldsymbol{v} \in \mathbb{R}^d$ is a random vector with differentiable positive density $p_{\boldsymbol{v}} : \mathbb{R}^d \to \mathbb{R}$. We define its score function as $S_{\boldsymbol{v}} : \mathbb{R}^d \to \mathbb{R}^d$, $S_{\boldsymbol{v}}(\mathbf{v}) = -\nabla \log p_{\boldsymbol{v}}(\mathbf{v})$. If a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ together with $\boldsymbol{v}$ satisfies regularity condition: $|p_{\boldsymbol{v}}(\mathbf{v})| \to 0$ as $\|\mathbf{v}\| \to \infty$ and $\mathbb{E}[|f(\boldsymbol{v})S_{\boldsymbol{v}}(\boldsymbol{v})|] \vee \mathbb{E}[|\nabla f(\boldsymbol{v})|] < \infty$, then we have*

$$\mathbb{E}[f(\boldsymbol{v})S_{\boldsymbol{v}}(\boldsymbol{v})] = \mathbb{E}[\nabla f(\boldsymbol{v})]. \tag{3}$$

In particular, when $\boldsymbol{v} \sim N(\mathbf{0}, \boldsymbol{I}_d)$, we have

$$\mathbb{E}[g(\boldsymbol{v})\boldsymbol{v}] = \mathbb{E}[\nabla g(\boldsymbol{v})].$$

We drop off the subscript of density and score function to make notation concise. In order to use Theorem 1 for estimation of coefficients in model (1), we require the following regularity condition.

Assumption 2 (Regularity) *We assume that $\boldsymbol{x}$, $\boldsymbol{z}$ in (1) are independent and the density function $p(\cdot)$ of $\boldsymbol{x}$ is positive and differentiable. For any $j \in [d_2]$, we assume function $\tilde{f}_j : \mathbb{R}^{d_1} \to \mathbb{R}$, defined to be $\tilde{f}_j(\mathbf{x}) = f_j(\langle \mathbf{x}, \boldsymbol{\beta}_j^\star \rangle)$, together with variable $\boldsymbol{x}$ satisfies the regularity condition of Theorem 1. Further, let $\mu_j := \mathbb{E}[f_j'(\langle \boldsymbol{x}, \boldsymbol{\beta}_j^\star \rangle)]$ and we assume $\mu_j \neq 0$. In addition, we assume covariate $\boldsymbol{z}$ are standardized with $\mathbb{E}[z_j] = 0$ and $\mathbb{E}[z_j^2] = 1$, $\forall j \in [d_2]$.*

Note that the standardization of $\boldsymbol{z}$ is made only to simplify our presentation. It is easy to extend to a general $\boldsymbol{z}$ using the fact that $\mathbb{E}[\boldsymbol{z}(\boldsymbol{z} - \mathbb{E}[\boldsymbol{z}])^T] = \mathrm{Var}(\boldsymbol{z})$, where diagonal entries of $\mathrm{Var}(\boldsymbol{z})$ can be assumed to be one without loss of generality, as the variance of each z_j can be absorbed into $f_j(\cdot)$. Since $\mathbb{E}[\boldsymbol{z}]$ is easy to estimate at a fast enough rate, we can replace $\boldsymbol{z}$ by $\boldsymbol{z} - \mathbb{E}[\boldsymbol{z}]$ whenever necessary in analysis for the general $\boldsymbol{z}$. See equation (9) for example. Under Assumption 2, Stein's identity will allow us to extract the unknown coefficient parameter, which is proportional to the derivative of the corresponding unknown function in the index model. To clarify, note that

$$\mathbb{E}[f_j(\langle \boldsymbol{x}, \boldsymbol{\beta}_j^\star \rangle)S(\boldsymbol{x})] = \mathbb{E}[\tilde{f}_j(\boldsymbol{x})S(\boldsymbol{x})] \overset{(3)}{=} \mathbb{E}[\nabla \tilde{f}_j(\boldsymbol{x})] = \mu_j \boldsymbol{\beta}_j^\star := \tilde{\boldsymbol{\beta}}_j. \tag{4}$$

The condition $\mu_j \neq 0$ ensures that the above expectation will not vanish and further $\boldsymbol{\beta}_j^\star$ can be fully identifiable from $\tilde{\boldsymbol{\beta}}_j$ due to (2).

With this setup, we start with a warm-up example and illustrate how to estimate coefficients $\{\boldsymbol{\beta}_j^\star\}_{j \in [d_2]}$ when $\boldsymbol{x} \sim N(\mathbf{0}, \boldsymbol{I}_{d_1})$, $\boldsymbol{z} \sim N(\mathbf{0}, \boldsymbol{I}_{d_2})$ and $\boldsymbol{x}$ and $\boldsymbol{z}$ are independent. The extension to heavy-tailed distributions is presented in the next section. Similar to (4), for any $k \in [d_2]$, Stein's identity gives us

$$\mathbb{E}[y \cdot z_k \cdot \boldsymbol{x}] = \sum_{j=1}^{d_2} \mathbb{E}[z_j z_k f_j(\langle \boldsymbol{\beta}_j^\star, \boldsymbol{x} \rangle)\boldsymbol{x}] = \mathbb{E}[f_k(\langle \boldsymbol{\beta}_k^\star, \boldsymbol{x} \rangle)\boldsymbol{x}] \overset{(4)}{=} \mu_k \boldsymbol{\beta}_k^\star = \tilde{\boldsymbol{\beta}}_k. \tag{5}$$

Under Assumption 2 and identifiability condition (2), the above equation allows us to form an estimator for $\boldsymbol{\beta}_k^\star$ by minimizing the following population loss

$$\tilde{\boldsymbol{\beta}}_k = \arg\min_{\boldsymbol{\beta}_k} L_k(\boldsymbol{\beta}_k) = \arg\min_{\boldsymbol{\beta}_k} \left\{ \|\boldsymbol{\beta}_k\|_2^2 - 2\mathbb{E}[y \cdot z_k \cdot \langle \boldsymbol{\beta}_k, \boldsymbol{x} \rangle] \right\}. \tag{6}$$

Given n i.i.d. copies of $(y, \boldsymbol{x}, \boldsymbol{z})$, $\{y_i, \boldsymbol{X}_i, \boldsymbol{Z}_i\}_{i=1}^n$, we obtain an estimator of $\tilde{\boldsymbol{\beta}}_k$ by replacing the expectation in (6) with a sample mean

$$\hat{\boldsymbol{\beta}}_k = \arg\min_{\boldsymbol{\beta}_k} \hat{L}_k(\boldsymbol{\beta}_k) + R_k(\boldsymbol{\beta}_k) = \arg\min_{\boldsymbol{\beta}_k} \left\{ \|\boldsymbol{\beta}_k\|^2 - \frac{2}{n}\sum_{i=1}^n y_i Z_{ik} \langle \boldsymbol{\beta}_k, \boldsymbol{X}_i \rangle + \lambda_k \|\boldsymbol{\beta}_k\|_1 \right\}, \quad (7)$$

where $R_k(\boldsymbol{\beta}_k)$ is a penalty function that imposes desired structural assumptions on the estimate. In a high-dimensional setting, it is common to assume that $\tilde{\boldsymbol{\beta}}_k$ is sparse, so here we use the ℓ_1-norm penalty with $R_k(\boldsymbol{\beta}_k) = \lambda_k \|\boldsymbol{\beta}_k\|_1$. Note that the loss function in (6) can also be written as

$$L(\boldsymbol{\beta}_k) = \mathbb{E}[(y - z_k \langle \boldsymbol{x}, \boldsymbol{\beta}_k \rangle)^2],$$

which leads to an alternative form for the estimator with a design matrix

$$\arg\min_{\boldsymbol{\beta}_k} \left\{ \frac{1}{n}\sum_{i=1}^n (y_i - Z_{ik} \boldsymbol{X}_i^T \boldsymbol{\beta}_k)^2 + \lambda_k \|\boldsymbol{\beta}_k\|_1 \right\}.$$

Finally, we note that the estimator in (7) can be obtained in a closed form

$$\hat{\boldsymbol{\beta}}_k = \mathcal{T}_{\lambda_k/2} \left(\frac{1}{n}\sum_{i=1}^n y_i Z_{ik} \boldsymbol{X}_i \right),$$

where $\mathcal{T}(\cdot)$ is the soft-thresholding operator defined at the end of Section 1.

Our first result establishes the convergence rate for the estimator in (7). We present the result for a slightly more general setting where $\boldsymbol{z}$ has independent sub-Gaussian components with $\|z_j\|_{\psi_2} = \Upsilon_{z_j}$, $\forall j \in [d_2]$.[2]

Theorem 3 (Warm-up) *Consider model* (1) *with* $\|\boldsymbol{\beta}_k^\star\|_0 \leqslant s$ *for* $k \in [d_2]$, $\boldsymbol{x} \sim N(\boldsymbol{0}, \boldsymbol{I}_{d_1})$, *components of* $\boldsymbol{z}$ *are independent with* $\|z_k\|_{\psi_2} = \Upsilon_{z_k} \leqslant \Upsilon_{\boldsymbol{z}}$ *for* $k \in [d_2]$ *and independent of* $\boldsymbol{x}$, *and* y *is sub-exponential with* $\|y\|_{\psi_1} \leqslant \Upsilon_y$. *Furthermore assume that Assumption 2 holds. The estimator in* (7) *with* $\lambda_k = 4\Upsilon\sqrt{\log n/n}$, *for a constant* Υ *that depends on* Υ_y *and* $\Upsilon_{\boldsymbol{z}}$ *only, satisfies*

$$\|\hat{\boldsymbol{\beta}}_k - \tilde{\boldsymbol{\beta}}_k\|_2 \leqslant \frac{3}{2}\sqrt{s}\lambda_k \quad \textit{and} \quad \|\hat{\boldsymbol{\beta}}_k - \tilde{\boldsymbol{\beta}}_k\|_1 \leqslant 6s\lambda_k, \quad \forall k \in [d_2]$$

with probability at least $1 - d_2 d_1/n^2$.

Theorem 3 established rate of convergence for the estimator of $\tilde{\boldsymbol{\beta}}_k$. In particular, it suggests that with high probability we have $\forall k \in [d_2]$,

$$\|\hat{\boldsymbol{\beta}}_k - \tilde{\boldsymbol{\beta}}_k\|_2 \lesssim \sqrt{\frac{s\log n}{n}} \quad \text{and} \quad \|\hat{\boldsymbol{\beta}}_k - \tilde{\boldsymbol{\beta}}_k\|_1 \lesssim s\sqrt{\frac{\log n}{n}},$$

2. For centered random variable x, we define $\|x\|_{\psi_1} = \sup_{p\geqslant 1} p^{-1}(\mathbb{E}|x|^p)^{1/p}$, $\|x\|_{\psi_2} = \sup_{p\geqslant 1} p^{-1/2}(\mathbb{E}|x|^p)^{1/p}$. We call x a sub-exponential random variable if $\|x\|_{\psi_1} < \infty$. We call x a sub-Gaussian random variable with proxy variance $\|x\|_{\psi_2}^2$ if $\|x\|_{\psi_2} < \infty$. See Vershynin (2012) for detailed properties.

which matches the optimal rate of convergence for sparse vectors recovery under the setting when $n \ll d_1 \ll n^2$ (Lin et al., 2017). It is useful to note that sub-exponential assumption on y is mild. For example, it is satisfied when $\{f_j\}_{j\in[d_2]}$ can be dominated by a linear function and ϵ is sub-exponential. The result follows from a bound on $\|\nabla \hat{L}_k(\tilde{\boldsymbol{\beta}}_k)\|_\infty$, which is presented in the following lemma.

Lemma 4 *Under the conditions of Theorem 3, we have* $\forall k \in [d_2]$,

$$P\left(\|\nabla \hat{L}_k(\tilde{\boldsymbol{\beta}}_k)\|_\infty > 2\Upsilon\sqrt{\frac{\log n}{n}}\right) < \frac{d_1}{n^2}.$$

Under the identifiability condition (2), we have the following corollary for the normalized estimator.

Corollary 5 *Suppose the conditions of Theorem 3 are satisfied, then for sufficiently large* n *(threshold depends on* s *and* $\min_{j\in[d_2]}(|\mu_j| \wedge \beta^\star_{j1})$*), we have*

$$\left\|sign(\hat{\beta}_{k1})\frac{\hat{\boldsymbol{\beta}}_k}{\|\hat{\boldsymbol{\beta}}_k\|_2} - \boldsymbol{\beta}^\star_k\right\|_2 \lesssim \sqrt{\frac{s\log n}{n}} \quad \text{and} \quad \left\|sign(\hat{\beta}_{k1})\frac{\hat{\boldsymbol{\beta}}_k}{\|\hat{\boldsymbol{\beta}}_k\|_2} - \boldsymbol{\beta}^\star_k\right\|_1 \lesssim s\sqrt{\frac{\log n}{n}}, \quad \forall k \in [d_2]$$

with probability $1 - d_2d_1/n^2$. *Further we can get* $\|\hat{\boldsymbol{B}} - \boldsymbol{B}^\star\|_F \lesssim \sqrt{\frac{d_2 s\log n}{n}}$ *where* $\hat{\boldsymbol{B}}$ *saves normalized estimators by column.*

Note that the order of $\|\hat{\boldsymbol{B}} - \boldsymbol{B}^\star\|_F$ in Corollary 5 is proven under the column-wise sparsity. Details under the assumption that $\boldsymbol{B}^\star$ is fully sparse will be discussed later.

In the next few sections, we will build on the illustrative example studied in this section and generalize our results to a heavy-tailed setting, which will improve the applicability of the estimator. Furthermore, we will consider estimation of all coefficients $\{\boldsymbol{\beta}^\star_j\}_{j\in[d_2]}$ simultaneously by imposing structural assumptions on the coefficient matrix $\boldsymbol{B}^\star$. Let $\tilde{\boldsymbol{B}} = (\tilde{\boldsymbol{\beta}}_1, ..., \tilde{\boldsymbol{\beta}}_{d_2})$, we usually focus on establishing statistical guarantee for estimating $\tilde{\boldsymbol{B}}$ as it keeps the same structure as $\boldsymbol{B}^\star$, and conversely, $\boldsymbol{B}^\star$ is fully identifiable from $\tilde{\boldsymbol{B}}$ under (2).

3. Overview of Results

In this section, we introduce weak moment assumption and provide an overview of the proposed estimators and their statistical convergence rates. Our theoretical analysis is separated into two cases: (i) estimating a single sparse coefficient vector $\boldsymbol{\beta}^\star_k$; (ii) estimating the coefficient matrix $\boldsymbol{B}^\star$. In the former case, we assume covariate $\boldsymbol{z}$ has independent entries so that we can extract one specific parameter, while in the latter case, we impose either low-rank or sparse structure on $\boldsymbol{B}^\star$ and relax the requirement for independence of $\boldsymbol{z}$ by modifying the estimation procedure to include the estimator of the precision matrix. We build our theoretical results on following weak moment condition.

Assumption 6 (Finite p-th moment) *We say finite* p*-th moment holds if there exists a constant* $M_p > 0$ *such that*

$$\mathbb{E}[y^p] \vee \mathbb{E}[S(\boldsymbol{x})^p_j] \vee \mathbb{E}[z^p_k] \leqslant M_p, \quad \forall j \in [d_1], k \in [d_2].$$

Single sparse vector recovery			
Type	Moment condition	Dimension	Rate
Warm-up	$\boldsymbol{x}, \boldsymbol{z} \sim N(0, \boldsymbol{I})$, indep; $y \sim$ subE	$s \ll n \ll d_1 \ll n^2$	$\sqrt{\frac{s \log n}{n}}$
General	$p = 6$	$s \ll n \ll d_1$	$\sqrt{\frac{s \log d_1 d_2}{n}}$
Low-rank matrix recovery			
Sparse precision	$p = 4$	$(r, w) \ll (d_1, d_2) \ll n$	$\sqrt{\frac{r(d_1+d_2)\log(d_1+d_2)}{n}} \vee w\sqrt{\frac{r \log d_2}{n}}$
General precision	$p = 4$	$r \ll (d_1, d_2) \ll n$	$\sqrt{\frac{r(d_1+d_2)\log(d_1+d_2)}{n}}$
Indep. entries in $\boldsymbol{z}$	$p = 4$	$r \ll (d_1, d_2) \ll n$	$\sqrt{\frac{r(d_1+d_2)\log(d_1+d_2)}{n}}$
Sparse matrix recovery			
Column sparse & sparse precision	$p = 6$	$(s, d_2) \ll n \ll d_1$	$\sqrt{\frac{s d_2 \log d_1 d_2}{n}}$
Column sparse & general precision	$p = 6$	$(s, d_2) \ll n \ll d_1$	$\sqrt{\frac{s d_2 \log d_1 d_2}{n}} \vee \frac{d_2\sqrt{s \log d_2}}{\sqrt{n}}$
Fully sparse & sparse precision	$p = 6$	$(s, w) \ll n \ll (d_1, d_2)$	$\sqrt{\frac{s \log d_1 d_2}{n}}$
Fully sparse & indep. entries in $\boldsymbol{z}$	$p = 6$	$s \ll n \ll (d_1, d_2)$	$\sqrt{\frac{s \log d_1 d_2}{n}}$

Table 1: Convergence Rate for $\|\cdot\|_2$ or $\|\cdot\|_F$.

This condition is used throughout all of our theoretical analysis. In sparse vector recovery, we require finite 6-th moment, while in low-rank matrix recovery, we only require finite 4-th moment. Note that though we cannot assume $S(\boldsymbol{x})$ is sub-Gaussian, it turns out assuming $S(\boldsymbol{x})$ to have finite moment is still reasonable in the sense that even for some heavy-tailed distributions, such as t-distribution and Gamma distribution, their score variables still satisfy the finite moment assumption for some p. On the other hand, assumptions for applying Stein's identity always boil down to finite moment conditions. For example, Yang et al. (2017) required finite 4-th moment when estimating SIM. In order to estimate varying indices, we require additional two moments to be finite.

Our results are summarized in Table 1. In summary, we achieve $\sqrt{s \log d_1/n}$ rate for estimating a single sparse vector, while $\sqrt{s \log d_1 d_2/n}$ for estimating a fully sparse parameter matrix. Both of them attain the minimax rate considering the case where all unknown link functions $f_j(\cdot)$ are identity functions. For low-rank estimation, we achieve $\sqrt{r(d_1 + d_2)\log(d_1 + d_2)/n}$ rate, which is also comparable with results in Plan and Vershynin (2016); Goldstein et al. (2018), though it only attains near-optimal rate up to the logarithmic factor. Note that estimating precision matrix of $\boldsymbol{z}$ can be conducted independently from our main procedure and any advanced, suitable estimators can be plugged into our approach. To make the paper compact but self-contained, we only consider estimating a general low-dimensional precision matrix for heavy-tailed $\boldsymbol{z}$ as an illustration, and discuss the high-dimensional sparse precision matrix estimation in appendix. Basically, if the precision matrix of $\boldsymbol{z}$ is sparse, we can estimate it by CLIME procedure (Cai et al., 2011) with a careful truncation of the sample covariance to attain optimal rate of convergence,

even though $\boldsymbol{z}$ only has finite certain moment. The proposed precision matrix estimator for heavy-tailed variable is used as a plug-in estimator whenever $\boldsymbol{z}$ has non-diagonal covariance matrix. Detailed estimation procedures and corresponding error rates are showed in Section 5 and Appendix A, respectively.

4. Sparse Vector Recovery

In this section, we present an extension of the estimator discussed in Section 2 to heavy-tailed data. Applying Theorem 1 after replacing $\boldsymbol{x}$ by $S(\boldsymbol{x})$ in (5) leads to

$$\mathbb{E}[y \cdot z_k \cdot S(\boldsymbol{x})] = \sum_{j=1}^{d_2} \mathbb{E}[z_j z_k f_j(\langle \boldsymbol{\beta}_j^\star, \boldsymbol{x}\rangle) S(\boldsymbol{x})] = \mathbb{E}[f_k(\langle \boldsymbol{\beta}_k^\star, \boldsymbol{x}\rangle) S(\boldsymbol{x})] \overset{(4)}{=} \mu_k \boldsymbol{\beta}_k^\star = \tilde{\boldsymbol{\beta}}_k,$$

under the independence condition that $\mathbb{E}[z_j z_k] = 0$ for $j \neq k$, which we maintain throughout the section but relax it in Section 5 and 6. The above identity allows us to estimate the direction of $\boldsymbol{\beta}_k^\star$ by estimating the left hand side even in the setting with heavy-tailed data. However, in order to get a fast rate of convergence we will require the covariates and the response to be appropriately truncated.

Given a threshold $\tau > 0$, we define the truncation of a vector $\boldsymbol{v} \in \mathbb{R}^d$ as $\breve{\boldsymbol{v}} \in \mathbb{R}^d$ whose coordinates are defined by $[\breve{\boldsymbol{v}}]_i = v_i$ if $|v_i| \leqslant \tau$ and 0 otherwise. Our estimator for $\tilde{\boldsymbol{\beta}}_k$ is given as

$$\hat{\boldsymbol{\beta}}_k = \arg\min_{\boldsymbol{\beta}_k} \bar{L}_k(\boldsymbol{\beta}_k) + R_k(\boldsymbol{\beta}_k) = \arg\min_{\boldsymbol{\beta}_k} \left\{ \|\boldsymbol{\beta}_k\|^2 - \frac{2}{n} \sum_{i=1}^n \breve{y}_i \widetilde{Z_{ik}} \langle \boldsymbol{\beta}_k, \widetilde{S(\boldsymbol{X}_i)} \rangle + \lambda_k \|\boldsymbol{\beta}_k\|_1 \right\}, \tag{8}$$

which can be obtained in a closed form as

$$\hat{\boldsymbol{\beta}}_k = \mathcal{T}_{\lambda_k/2}\left(n^{-1} \sum_{i=1}^n \breve{y}_i \widetilde{Z_{ik}} \widetilde{S(\boldsymbol{X}_i)} \right).$$

Compared to the estimator in (7), we have replaced $\boldsymbol{X}_i$ by $S(\boldsymbol{X}_i)$ and have carefully truncated the data to obtain the following result.

Theorem 7 (Single sparse vector recovery) *Consider the model* (1) *with* $\|\boldsymbol{\beta}_k^\star\|_0 \leqslant s$, $\forall k \in [d_2]$. *Suppose Assumption 2, Assumption 6* ($p = 6$) *hold and* $\mathbb{E}[z_j z_k] = 0$ *for* $j \neq k$, *then the estimator defined in* (8) *with* $\lambda_k = 76\sqrt{M_6 \log d_1 d_2 / n}$ *and* $\tau = (M_6 n / \log d_1 d_2)^{1/6}/2$ *satisfies*

$$\|\hat{\boldsymbol{\beta}}_k - \tilde{\boldsymbol{\beta}}_k\|_2 \leqslant \frac{3}{2}\sqrt{s}\lambda_k \quad \textit{and} \quad \|\hat{\boldsymbol{\beta}}_k - \tilde{\boldsymbol{\beta}}_k\|_1 \leqslant 6s\lambda_k, \quad \forall k \in [d_2],$$

with probability at least $1 - 2/d_1^2 d_2^2$.

The theorem establishes that

$$\|\hat{\boldsymbol{\beta}}_k - \tilde{\boldsymbol{\beta}}_k\|_2 \lesssim \sqrt{\frac{s \log d_1 d_2}{n}} \quad \text{and} \quad \|\hat{\boldsymbol{\beta}}_k - \tilde{\boldsymbol{\beta}}_k\|_1 \lesssim s\sqrt{\frac{\log d_1 d_2}{n}}$$

with high probability. When $d_2 = 1$ the rate matches the minimax rate established in Lin et al. (2017). Our proof technique requires finite 6-th moment, which ensures that the truncated variables do not lose too much information. This assumption can be compared to boundedness of the 4-th moment in estimation of a single-index model (Yang et al., 2017). We require a stronger assumption due to estimation in a more general model. Theorem 7 follows from a bound on $\|\nabla \bar{L}_k(\tilde{\boldsymbol{\beta}}_k)\|_\infty$ given in the following lemma.

Lemma 8 *Under the conditions of Theorem 7,*

$$P\bigg(\|\nabla \bar{L}_k(\tilde{\boldsymbol{\beta}}_k)\|_\infty \leqslant 38\sqrt{\frac{M_6 \log d_1 d_2}{n}}, \quad \forall k \in [d_2]\bigg) \geqslant 1 - \frac{2}{d_1^2 d_2^2}.$$

From the standard analysis of the ℓ_1-penalized methods in Bühlmann and van de Geer (2011), we know that the penalty parameter λ_k should be set as $c\|\nabla \hat{L}_k(\tilde{\boldsymbol{\beta}}_k)\|_\infty$ for some $c > 0$. Moreover, we see threshold τ has the order $\tau \asymp 1/\lambda_k^{2/p}$, where p is the number of moments variables have. Thus, the more moments the variables have, the smaller the threshold level τ would be, which is also consistent with our intuition.

We note that the estimator in (8) crucially depends on the independence among coordinates of $\boldsymbol{z}$. Without this assumption the estimator is not valid. In what follows, we study estimators of the matrix $\boldsymbol{B}^\star$ as a whole, by imposing either low-rank or sparse structure, instead of estimating the matrix column by column.

5. Low-rank Matrix Recovery

In this section, we propose an estimator for $\boldsymbol{B}^\star$ in model (1), which has near optimal rate of convergence under an assumption that $\boldsymbol{B}^\star$ is low-rank. We relax the condition that $\mathbb{E}[z_j z_k] = 0$ as assumed earlier, by estimating the inverse of the covariance of $\boldsymbol{z}$, also called the precision matrix. Let $\boldsymbol{\Sigma}^\star = \mathbb{E}[\boldsymbol{z}\boldsymbol{z}^\top] \in \mathbb{R}^{d_2 \times d_2}$ and $\boldsymbol{\Omega}^\star = (\boldsymbol{\Sigma}^\star)^{-1}$. We consider two cases: (i) when d_2 is in low dimensions, we have no structural assumptions on the precision matrix $\boldsymbol{\Omega}^\star$; (ii) when d_2 is in high dimensions, we assume that the precision matrix is in the set $\mathcal{F}_w^K$ for some w and K, where

$$\mathcal{F}_w^K = \bigg\{\boldsymbol{\Omega} \succeq \mathbf{0} : \|\boldsymbol{\Omega}\|_{0,\infty} \leqslant w, \|\boldsymbol{\Omega}\|_2 \leqslant K, \|\boldsymbol{\Omega}^{-1}\|_2 \leqslant K\bigg\}.$$

The set above is borrowed from Cai et al. (2011) and it controls upper and lower bounds on eigenvalues of $\boldsymbol{\Omega}^\star$, as well as the maximal sparsity over columns. Since estimating the precision matrix itself is a well studied topic and can be conducted independently from estimating model (1), we only take the former case as an example. For the latter case, the sparsity structure on precision matrix can allow us to study the model with both d_1, d_2 in high dimensions. We will discuss how to make use of the CLIME procedure (Cai et al., 2011) with truncated sample covariance to estimate the precision matrix for heavy-tailed variable in Appendix A.

We start by writing down the identifiability relationship. Under Assumption 2, we have

$$\begin{aligned}\mathbb{E}[y \cdot S(\boldsymbol{x})\boldsymbol{z}^T]\boldsymbol{\Omega}^\star &= \sum_{j=1}^{d_2} \mathbb{E}[f_j(\langle \boldsymbol{\beta}_j^\star, \boldsymbol{x}\rangle)S(\boldsymbol{x})]\mathbb{E}[z_j \cdot \boldsymbol{z}^T]\boldsymbol{\Omega}^\star \\ &= \sum_{j=1}^{d_2} \tilde{\boldsymbol{\beta}}_j \boldsymbol{e}_j^T \boldsymbol{\Sigma}^\star \boldsymbol{\Omega}^\star = (\tilde{\boldsymbol{\beta}}_1, ..., \tilde{\boldsymbol{\beta}}_{d_2}) = \tilde{\boldsymbol{B}}, \end{aligned} \tag{9}$$

where $\boldsymbol{e}_j \in \mathbb{R}^{d_2}$ is the j-th canonical basis vector. This relationship allows us to estimate the $\tilde{\boldsymbol{B}}$ as a minimizer of the population loss,

$$\tilde{\boldsymbol{B}} = \arg\min_{\boldsymbol{B}} \left\{ \|\boldsymbol{B}\|_F^2 - 2\mathbb{E}[y \cdot \langle S(\boldsymbol{x})\boldsymbol{z}^T\boldsymbol{\Omega}^\star, \boldsymbol{B}\rangle] \right\}. \tag{10}$$

In order to use the above relationship, we will separately estimate $\mathbb{E}[y \cdot S(\boldsymbol{x})\boldsymbol{z}^T]$ and $\boldsymbol{\Omega}^\star$.

Let

$$\phi(x) = \begin{cases} -\log(1 - x + x^2/2) & \text{if } x \leqslant 0, \\ \log(1 + x + x^2/2) & \text{if } x > 0 \end{cases} \tag{11}$$

be the soft truncation function. When $|x|$ is small $\phi(x) \approx x$; when $|x|$ is relatively large $\phi(x)$ dramatically shrinks x, while still being monotonically increasing. This function has been widely used in robust mean estimation, especially for variables that only have finite certain moments. For the univariate case, suppose $v_1, \ldots, v_n$ to be a sequence of i.i.d. samples distributed as $v_i \sim v$ that only has finite 2nd moment. Catoni (2012) proposed a simple mean estimator defined as $(n\kappa)^{-1}\sum_{i=1}^n \phi(\kappa v_i)$ with a properly chosen κ and showed sub-Gaussian tail bound. Minsker (2018) generalized to the multivariate case using estimator $(n\kappa)^{-1}\sum_{i=1}^n \phi(\kappa \boldsymbol{V}_i)$, where $\boldsymbol{V}_i$ are independent random Hermitian matrices with finite 2nd moments and $\phi(\cdot)$ is applied to the eigenvalues only. Similarly, when $\boldsymbol{V}_i$ are not Hermitian, we can define a dimension-free matrix soft truncation function $\Phi(\cdot)$ using Hermitian dilation as follows: for a matrix $\boldsymbol{V}$, let $\begin{pmatrix} \boldsymbol{0} & \boldsymbol{V} \\ \boldsymbol{V}^T & \boldsymbol{0} \end{pmatrix} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^T$ be the eigenvalue decomposition of the Hermitian dilation of $\boldsymbol{V}$. Let $\tilde{\boldsymbol{U}} = \boldsymbol{Q}\phi(\boldsymbol{\Lambda})\boldsymbol{Q}^T$, where $\phi(\boldsymbol{\Lambda})$ is applied entrywise. Then $\Phi(\boldsymbol{V})$ is defined to be the upper right corner matrix of $\tilde{\boldsymbol{U}}$ with the same dimension as $\boldsymbol{V}$. Using the truncation function $\Phi(\cdot)$, our estimator for $\mathbb{E}[yS(\boldsymbol{x})\boldsymbol{z}^T]$ is further defined as

$$\frac{1}{n\kappa_1}\sum_{i=1}^n \Phi(\kappa_1 y_i \cdot S(\boldsymbol{X}_i)\boldsymbol{Z}_i^T), \tag{12}$$

where $\kappa_1 > 0$ is a user-specified parameter. In the later section, we will also apply the function $\Phi(\cdot)$ when estimating precision matrix for heavy-tailed variables in low dimensions. With the matrix in (12), our estimator of $\tilde{\boldsymbol{B}}$ is given as

$$\hat{\boldsymbol{B}} = \arg\min_{\boldsymbol{B}} \left\{ \|\boldsymbol{B}\|_F^2 - \frac{2}{n\kappa_1}\sum_{i=1}^n \langle \Phi(\kappa_1 y_i \cdot S(\boldsymbol{X}_i)\boldsymbol{Z}_i^T)\hat{\boldsymbol{\Omega}}, \boldsymbol{B}\rangle + \lambda\|\boldsymbol{B}\|_* \right\}, \tag{13}$$

where $\hat{\boldsymbol{\Omega}}$ is an estimator of $\boldsymbol{\Omega}^\star$. The penalty function $\lambda\|\boldsymbol{B}\|_*$ biases the estimated matrix $\hat{\boldsymbol{B}}$ to be low-rank. Note that the estimator $\hat{\boldsymbol{B}}$ can be obtained in a closed form as

$$\hat{\boldsymbol{B}} = \mathcal{T}_{\lambda/2}\bigg(\frac{1}{n\kappa_1}\sum_{i=1}^{n}\Phi(\kappa_1 y_i \cdot S(\boldsymbol{X}_i)\boldsymbol{Z}_i^T)\hat{\boldsymbol{\Omega}}\bigg).$$

We characterize convergence rate for the estimator (13) in the next theorem and discuss estimation of a general low-dimensional $\boldsymbol{\Omega}^\star$ for heavy-tailed covariate later.

Theorem 9 (Low-rank matrix recovery) *Consider the model (1) with rank$(\boldsymbol{B}^\star) \leqslant r$. Suppose Assumption 2, Assumption 6 ($p = 4$) hold and furthermore suppose a precision matrix estimator $\hat{\boldsymbol{\Omega}}$ satisfies*

$$P\big(\|\hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}^\star\|_2 \leqslant \mathcal{H}(n, d_2)\big) \geqslant 1 - \mathcal{P}(n, d_2).$$

Denote $K = \|\boldsymbol{\Sigma}^\star\|_2 \vee \|\boldsymbol{\Omega}^\star\|_2$. If we set $\kappa_1 = \sqrt{\frac{2\log(d_1+d_2)}{n(d_1+d_2)M_4^{3/2}}}$ and

$$\lambda \geqslant 16KM_4^{3/4}\sqrt{\frac{(d_1+d_2)\log(d_1+d_2)}{n}} + 4K\max_{j\in[d_2]}|\mu_j| \cdot \|\boldsymbol{B}^\star\|_2 \cdot \mathcal{H}(n, d_2),$$

then the estimator (13) satisfies

$$\|\hat{\boldsymbol{B}} - \tilde{\boldsymbol{B}}\|_F \leqslant 3\sqrt{r}\lambda \quad \textit{and} \quad \|\hat{\boldsymbol{B}} - \tilde{\boldsymbol{B}}\|_* \leqslant 24r\lambda.$$

with probability at least $1 - 2/(d_1+d_2)^2 - \mathcal{P}(n, d_2)$.

The theorem follows from the following concentration result.

Lemma 10 *Under the conditions in Theorem 9, we have*

$$\bigg\|\frac{1}{n\kappa_1}\sum_{i=1}^{n}\Phi(\kappa_1 y_i \cdot S(\boldsymbol{X}_i)\boldsymbol{Z}_i^T) - \mathbb{E}[y \cdot S(\boldsymbol{x})\boldsymbol{z}^T]\bigg\|_2 \leqslant 4M_4^{3/4}\sqrt{\frac{(d_1+d_2)\log(d_1+d_2)}{n}},$$

with probability at least $1 - \frac{2}{(d_1+d_2)^2}$.

Different from Theorem 7, the penalty parameter λ depends also on a term that comes from estimating $\boldsymbol{\Omega}^\star$. Specifically, if it holds that $\mathbb{E}[z_j z_k] = 0$, we can get the following corollary immediately.

Corollary 11 *Suppose the conditions of Theorem 9 are satisfied. In addition, suppose that $\mathbb{E}[\boldsymbol{z}\boldsymbol{z}^\top] = \boldsymbol{I}_{d_2}$. If we set $\kappa_1 = \sqrt{\frac{2\log(d_1+d_2)}{n(d_1+d_2)M_4^{3/2}}}$ and $\lambda = 16M_4^{3/4}\sqrt{\frac{(d_1+d_2)\log(d_1+d_2)}{n}}$, then*

$$\|\hat{\boldsymbol{B}} - \tilde{\boldsymbol{B}}\|_F \leqslant 3\sqrt{r}\lambda \quad \textit{and} \quad \|\hat{\boldsymbol{B}} - \tilde{\boldsymbol{B}}\|_* \leqslant 24r\lambda,$$

with probability at least $1 - 2/(d_1+d_2)^2$.

Next, we briefly discuss how to estimate the precision matrix $\boldsymbol{\Omega}^\star$, noting that any suitable estimator for heavy-tailed data can be used. In a general case, when no additional structural assumptions are available, we can invert the soft truncated empirical covariance matrix as

$$\hat{\boldsymbol{\Omega}} = \hat{\boldsymbol{\Sigma}}^{-1} \quad \text{where} \quad \hat{\boldsymbol{\Sigma}} = \frac{1}{n\kappa_2}\sum_{i=1}^{n} \Phi(\kappa_2 \boldsymbol{Z}_i \boldsymbol{Z}_i^T). \tag{14}$$

We will show that $\hat{\boldsymbol{\Sigma}}$ is invertible for sufficiently large n. In particular, we prove $\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^\star\|_2 \lesssim \sqrt{d_2 \log d_2/n}$. Thus, $\hat{\boldsymbol{\Sigma}}$ is invertible when $\sqrt{d_2 \log d_2/n} < \lambda_{\min}(\boldsymbol{\Sigma}^\star)$, where $\lambda_{\min}(\boldsymbol{\Sigma}^\star)$ denotes the minimum eigenvalue of $\boldsymbol{\Sigma}^\star$. The following lemma characterizes the rate of convergence.

Lemma 12 *Set* $\kappa_2 = \sqrt{\frac{2\log d_2}{n d_2 M_4^{1/2}}}$. *If* $n \geqslant 64\sqrt{M_4}K^2 d_2 \log d_2$, *the estimator* (14) *satisfies*

$$P\left(\|\hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}^\star\|_2 \leqslant 8K^2 M_4^{1/4}\sqrt{\frac{d_2 \log d_2}{n}}\right) \geqslant 1 - \frac{2}{d_2^2}.$$

In fact, we only need finite 2nd moment for $\boldsymbol{z}$ to make $\hat{\boldsymbol{\Omega}}$ in (14) consistent. Combining the rate obtained in Lemma 12 with that of Theorem 9, we observe that

$$\|\hat{\boldsymbol{B}} - \tilde{\boldsymbol{B}}\|_F \lesssim \sqrt{\frac{r(d_1+d_2)\log(d_1+d_2)}{n}}.$$

with high probability. In particular, the rate of convergence is governed by the rate obtained in Lemma 10 and the estimation of the precision matrix contributes to the higher order terms. Furthermore, we note that the rate is optimal up to logarithmic terms (Rohde and Tsybakov, 2011). Similar rate is shown in estimating the single-index model (Plan and Vershynin, 2016; Goldstein et al., 2018; Yang et al., 2017). Estimation of a high-dimensional sparse precision matrix is presented in Appendix A.

6. Sparse Matrix Recovery

In this section, we consider the setting as in Section 5, but with the parameter matrix $\boldsymbol{B}^\star$ being sparse rather than low-rank. Different from (12), here we can simply estimate $\mathbb{E}[y \cdot S(\boldsymbol{x})\boldsymbol{z}^T]$ by

$$\frac{1}{n}\sum_{i=1}^{n} \breve{y}_i \cdot \widetilde{S(\boldsymbol{X}_i)}\breve{\boldsymbol{Z}}_i^T \tag{15}$$

for some truncation threshold $\tau > 0$. We apply hard truncation instead of soft truncation (based on $\phi(\cdot)$) in (15), since we are going to bound the max norm of the error, instead of the operator norm as we did in Section 5. The hard truncation can give a fast rate when bounding the max norm, while a slow rate when bounding the operator norm. The soft truncation works in the opposite way. Finally, our estimator is defined as

$$\hat{\boldsymbol{B}} = \arg\min_{\boldsymbol{B}} \left\{ \|\boldsymbol{B}\|_F^2 - \frac{2}{n}\sum_{i=1}^{n} \langle \breve{y}_i \cdot \widetilde{S(\boldsymbol{X}_i)}\breve{\boldsymbol{Z}}_i^T \hat{\boldsymbol{\Omega}}, \boldsymbol{B}\rangle + \lambda\|\boldsymbol{B}\|_{1,1} \right\}. \tag{16}$$

We obtain the following rate of convergence for $\hat{\boldsymbol{B}}$.

Theorem 13 (Sparse matrix recovery (column-wise sparse)) *Consider the model (1) with $\|\boldsymbol{\beta}_k^\star\|_0 \leqslant s$ for all $k \in [d_2]$. Suppose Assumption 2 and Assumption 6 ($p = 6$) hold and furthermore suppose that the precision matrix estimator $\hat{\boldsymbol{\Omega}}$ satisfies*

$$P(\|\hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}^\star\|_{\max} \leqslant \tilde{\mathcal{H}}(n, d_2)) \geqslant 1 - \tilde{\mathcal{P}}(n, d_2).$$

If $\tau = (M_6 n / \log d_1 d_2)^{1/6}/2$ in (15) and

$$\lambda \geqslant 76\|\boldsymbol{\Omega}^\star\|_1 \sqrt{\frac{M_6 \log d_1 d_2}{n}} + 4 \max_{j \in [d_2]} |\mu_j| \cdot \|\boldsymbol{B}^\star \boldsymbol{\Sigma}^\star\|_\infty \tilde{\mathcal{H}}(n, d_2),$$

then

$$\|\hat{\boldsymbol{B}} - \tilde{\boldsymbol{B}}\|_F \leqslant 2\sqrt{s d_2}\lambda \quad \textit{and} \quad \|\hat{\boldsymbol{B}} - \tilde{\boldsymbol{B}}\|_{1,1} \leqslant 8 s d_2 \lambda,$$

with probability at least $1 - 2/d_1^2 d_2^2 - \tilde{\mathcal{P}}(n, d_2)$.

Different from Theorem 9, we require the bound $\|\hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}^\star\|_{\max}$ with high probability here because $\|\cdot\|_{\max}$ is the dual norm of $\|\cdot\|_{1,1}$. Note that $\|\hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}^\star\|_{\max} \leqslant \|\hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}^\star\|_2$, so we can simply have $\tilde{\mathcal{H}}(n, d_2) = \mathcal{H}(n, d_2)$ and $\tilde{\mathcal{P}}(n, d_2) = \mathcal{P}(n, d_2)$ for estimation in low dimensions where $\mathcal{H}(n, d_2)$ and $\mathcal{P}(n, d_2)$ come from Lemma 12. However, this bound is not sharp for CLIME procedure in high dimensions. Above theorem follows from the following lemma.

Lemma 14 *Under the conditions in Theorem 13,*

$$\left\| \mathbb{E}[y \cdot S(\boldsymbol{x})\boldsymbol{z}^T] - \frac{1}{n}\sum_{i=1}^n \breve{y}_i \cdot \widetilde{S(\boldsymbol{X}_i)} \breve{\boldsymbol{Z}}_i^T \right\|_{\max} \leqslant 19\sqrt{\frac{M_6 \log d_1 d_2}{n}}$$

with probability at least $1 - 2/d_1^2 d_2^2$.

Note that the rate obtained in Theorem 13 is the same as the one obtained in Theorem 7, which required the assumption that $\mathbb{E}[z_j z_k] = 0$. Furthermore, we observe that the same proof provided for Theorem 13 can be used under the setting that $n \ll d_1 \wedge d_2$ and $\boldsymbol{B}^\star$ is fully sparse with $\|\boldsymbol{B}^\star\|_{0,1} \leqslant s$. Estimation of a high-dimensional precision matrix is discussed in Appendix A, but we note that the error in estimating the precision matrix in any case only contributes higher order terms and our final rate is

$$\|\hat{\boldsymbol{B}} - \tilde{\boldsymbol{B}}\|_F \lesssim \sqrt{\frac{s \log d_1 d_2}{n}} \quad \text{and} \quad \|\hat{\boldsymbol{B}} - \tilde{\boldsymbol{B}}\|_{1,1} \lesssim s\sqrt{\frac{\log d_1 d_2}{n}}$$

with probability at least $1 - 2/d_1 d_2 - 2/d_2^2$. It follows by combining Theorem 13 with Lemma 15. Finally, similar to Corollary 11, when $\boldsymbol{\Sigma}^\star = \boldsymbol{I}_{d_2}$, we can set $\tilde{\mathcal{H}}(n, d_2) = \tilde{\mathcal{P}}(n, d_2) = 0$ in Theorem 13 to derive the same optimal rate.

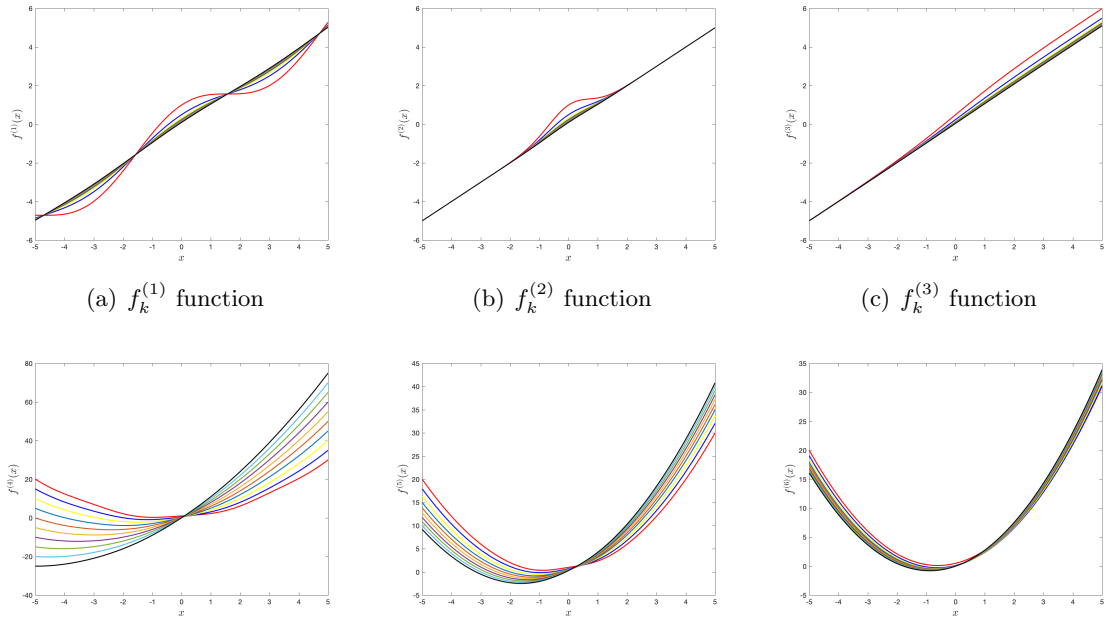
Until now, we have shown a comprehensive theoretical analysis for the model (1). When estimating a single sparse vector, we assume that $\boldsymbol{z}$ has independent entries. When estimating a parameter matrix $\boldsymbol{B}^\star$, we relax this assumption by incorporating an estimate of the precision matrix. Based on our analysis, we see the error occurred in estimating $\mathbb{E}[y \cdot S(\boldsymbol{x})\boldsymbol{z}^T]$ will always be the dominant term, while estimation of the precision matrix only contributes higher order terms.

7. Numerical Experiment

In this section, we illustrate the performance of our proposed estimators in different simulation settings and carry out a real data application in the end. In simulation studies, the link function is set to one of the following six forms:

$$f_k^{(1)}(x) = x + \frac{1}{k}\cos(x); \qquad f_k^{(2)}(x) = x + \frac{1}{k}\exp(-x^2); \qquad f_k^{(3)}(x) = x + \frac{1}{k}\frac{\exp(x)}{1+\exp(x)};$$
$$f_k^{(4)}(x) = x^2 + kx + \frac{1}{k}\cos^2(x); \quad f_k^{(5)}(x) = x^2 + \sqrt{k}x + \frac{1}{\sqrt{k}}\exp(-x^2);$$
$$f_k^{(6)}(x) = x^2 + k^{\frac{1}{4}}x + \frac{1}{k^2}\frac{\exp(x)}{1+\exp(x)}.$$

Their plots are shown in Figure 1. For all simulations we let $\epsilon \sim N(0, 0.01)$. To measure the estimation accuracy we use the cosine distance defined by $\cos(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}^\star) = 1 - |\hat{\boldsymbol{\beta}}^T\boldsymbol{\beta}^\star|/\|\hat{\boldsymbol{\beta}}\|_2$. Note that we do not normalize $\hat{\boldsymbol{\beta}}$ and change its direction according to the sign of its first entry because cosine distance is more suitable for verifying our matrix results and it allows us to generate $\boldsymbol{\beta}^\star$ without restricting the first entry to be positive. We can easily see $\cos(\hat{\boldsymbol{\beta}}_k, \boldsymbol{\beta}_k^\star) \asymp \|\hat{\boldsymbol{\beta}}_k - \tilde{\boldsymbol{\beta}}_k\|_2^2$, $\forall k \in [d_2]$. For a matrix estimator, we will sum up cosine distance



(a) $f_k^{(1)}$ function (b) $f_k^{(2)}$ function (c) $f_k^{(3)}$ function

(d) $f_k^{(4)}$ function (e) $f_k^{(5)}$ function (f) $f_k^{(6)}$ function

Figure 1: The link functions used in simulations. When k varies from 1 to 10, the line moves from red to black. The link functions in the first row are essentially linear functions combined with different fluctuations, while in the second row they are quadratic. As k increases, the fluctuation is more and more moderate.

Distribution	parameter	score function
Gaussian	$\mu = 0$; $\sigma = 1$	$s(x) = x$
Beta	$\alpha = 8$; $\beta = 8$	$s(x) = \frac{14x-7}{(1-x)x}$
Gamma	$k = 8$; $\theta = 0.1$	$s(x) = 10 - \frac{7}{x}$
Student's t	$\nu = 13$	$s(x) = \frac{14x}{13+x^2}$
Rayleigh	$\sigma = 1$	$s(x) = x - \frac{1}{x}$
Weibull	$k = 7$; $\lambda = 1$	$s(x) = 7x^6 - \frac{6}{x}$

Table 2: *Distribution of* $\boldsymbol{x}$

over all columns, which is a rough surrogate of $\|\hat{\boldsymbol{B}} - \tilde{\boldsymbol{B}}\|_F^2$. Our results are averaged of 40 independent runs.

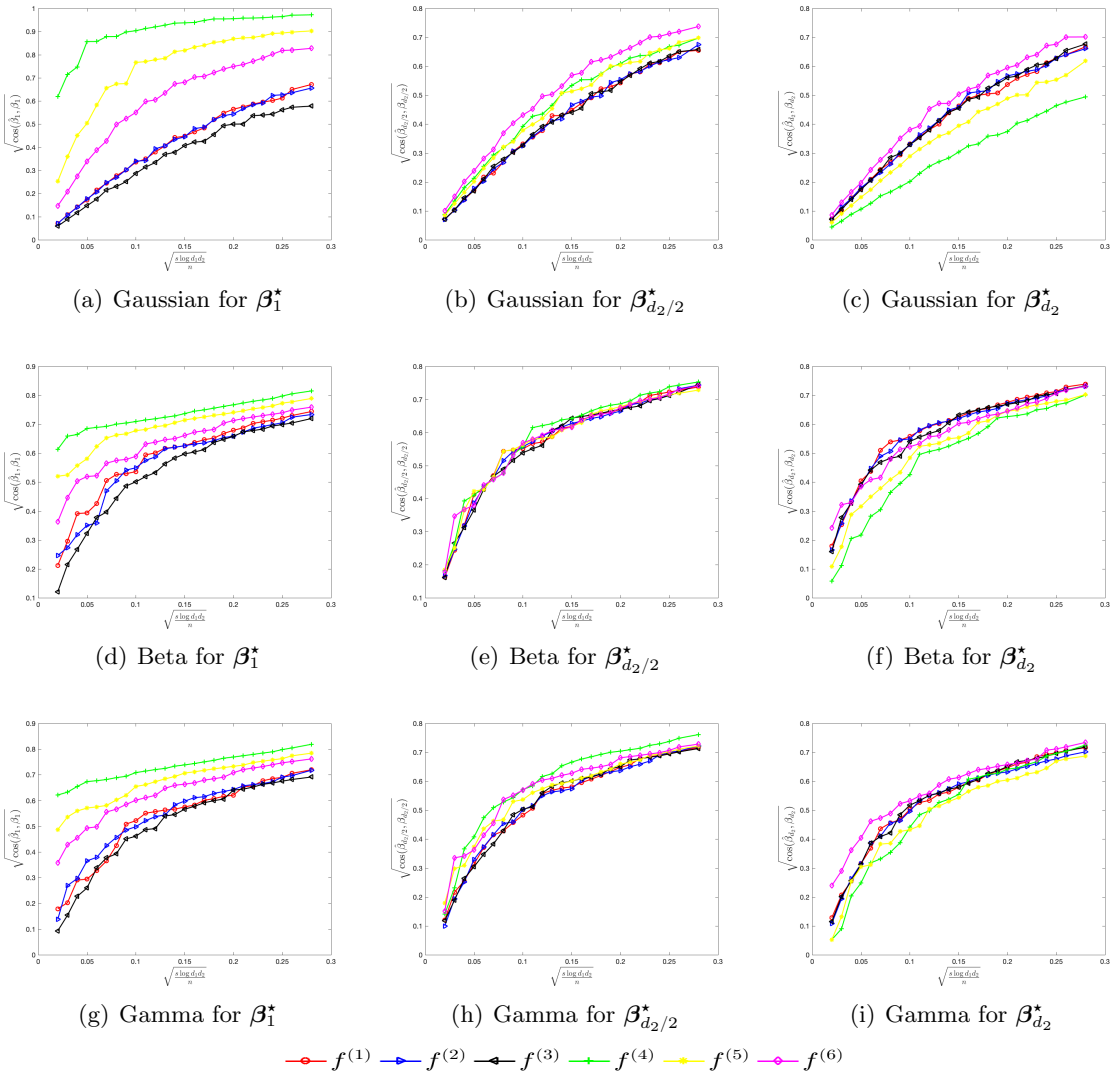
7.1. Single Sparse Vector

We set $d_1 = 100$, $d_2 = 20$, $s = 10$, and vary n only. For $i \in [n]$ and $k \in [d_2]$, we let $\boldsymbol{X}_i$ be independent and identically generated from one of distributions listed in Table 2, and $Z_{ik} \in \{-1, 1\}$ with equal probability and independent from other coordinates. To generate d_1-dimensional parameter $\boldsymbol{\beta}_k^\star$, we first generate the support of nonzero coefficients S_k uniformly at random and then let $\{[\boldsymbol{\beta}_k^\star]_l\}_{l \in S_k} \overset{iid}{\sim} \frac{1}{\sqrt{s}} \cdot \mathrm{Unif}(\{-1, 1\})$. According to Theorem 3 and 7, we set $\lambda_k = 30\sqrt{\log d_1 d_2 / n}$ and $\tau = 2(n/\log d_1 d_2)^{1/6}$.

Figure 2 illustrates the error trend for estimating $\boldsymbol{\beta}_1^\star$, $\boldsymbol{\beta}_{d_2/2}^\star$ and $\boldsymbol{\beta}_{d_2}^\star$ under different link functions and $\boldsymbol{x}$ following one of the following distributions: Gaussian, Beta or Gamma. Additional results are shown in Figure 9 in Appendix D for $\boldsymbol{x}$ following t_{13}, Rayleigh or Weibull distributions. From the plots, we see the error for using linear link functions ($f^{(1)}$ to $f^{(3)}$) is generally smaller than the error for using quadratic link functions ($f^{(4)}$ to $f^{(6)}$), no matter which distribution we use for design. But in all settings the scaled error plots have a linear trend when $n \gg s \log d_1 d_2$, which is consistent with Theorem 7. Another interesting observation is that, as k increases, the estimation will be more and more precise. This can also be seen from our theoretical results. Note that quadratic function $f(x) = x^2$ brings the singularity to the estimation problem even for the Gaussian design since $\mathbb{E}[f'(\langle \boldsymbol{x}, \boldsymbol{\beta}^\star \rangle)] = 0$. Thus, when k is small, the scalar μ_k, appeared in (5), will be close to zero, which implies that the estimation problem is almost singular. Fortunately, when k is large, the linear term in quadratic functions will have more contributions and $\mu_k \approx k$. From Corollary 5 and identity (5), we see that μ_k amplifies the magnitude of the signal $\boldsymbol{\beta}_k^\star$ that characterizes the difficulty of estimation problem. Therefore a larger μ_k implies an easier problem. Figure 10—12 in Appendix D provide additional simulation results where we plot the error trend for all d_2 parameters, which make this observation more clear.

7.2. Low-rank Matrix

We study the estimation of $\boldsymbol{B}^\star$ under the low-rank assumption. We let $d_1 = d_2 = 25$, $r = 5$. The distribution of $\boldsymbol{x}$ is set as described in Table 2. For generating $\boldsymbol{B}^\star$, we first generate two random orthogonal matrices $\boldsymbol{U} \in \mathbb{R}^{d_1 \times d_1}$ and $\boldsymbol{V} \in \mathbb{R}^{d_2 \times d_2}$, then generate $d_1 \times d_2$ diagonal matrix $\boldsymbol{\Lambda}$ by first sampling r locations from $[\min\{d_1, d_2\}]$ and further



(a) Gaussian for $\boldsymbol{\beta}_1^\star$ (b) Gaussian for $\boldsymbol{\beta}_{d_2/2}^\star$ (c) Gaussian for $\boldsymbol{\beta}_{d_2}^\star$

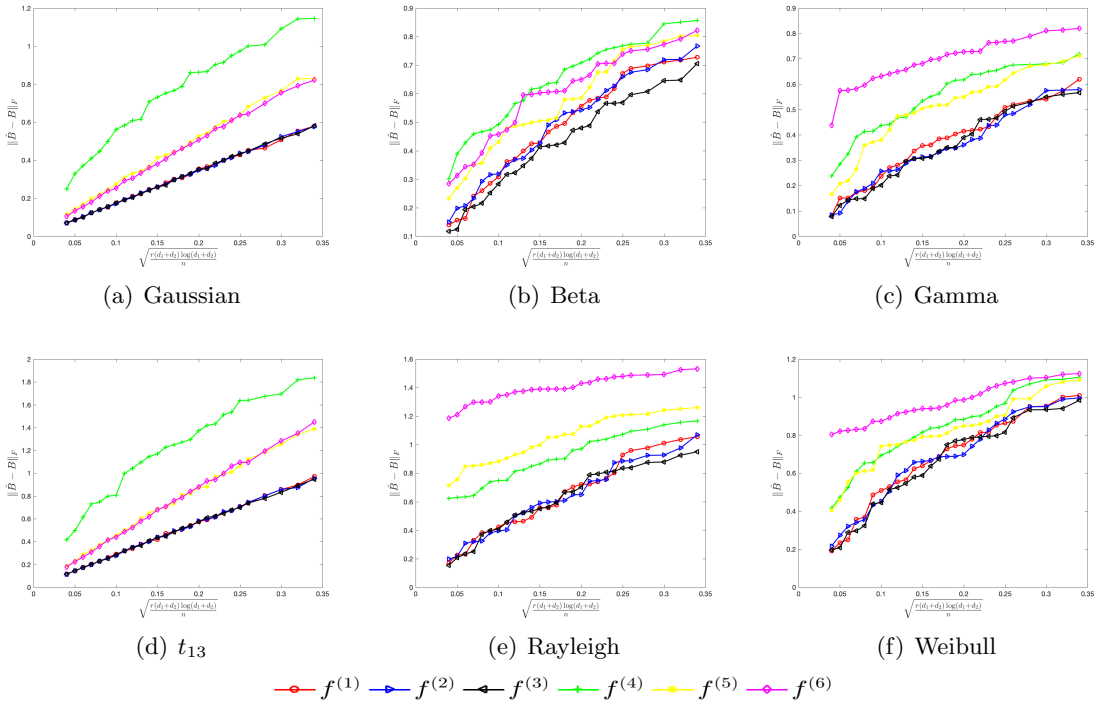
(d) Beta for $\boldsymbol{\beta}_1^\star$ (e) Beta for $\boldsymbol{\beta}_{d_2/2}^\star$ (f) Beta for $\boldsymbol{\beta}_{d_2}^\star$

(g) Gamma for $\boldsymbol{\beta}_1^\star$ (h) Gamma for $\boldsymbol{\beta}_{d_2/2}^\star$ (i) Gamma for $\boldsymbol{\beta}_{d_2}^\star$

$f^{(1)}$ $f^{(2)}$ $f^{(3)}$ $f^{(4)}$ $f^{(5)}$ $f^{(6)}$

Figure 2: Sparse vector estimation plot (I). This figure shows cosine distance trend for error of estimating single sparse parameter in model (1). Six lines indicates six different types of link functions. Each row represents one type of distribution for design.

setting each entry to be $\{-\frac{1}{\sqrt{r}}, \frac{1}{\sqrt{r}}\}$ with equal probability. Finally we let $\boldsymbol{B}^\star = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^T$. We consider both independent and dependent $\boldsymbol{z}$. For the independent case, $\boldsymbol{z}$ is generated as in the previous section. For the dependent case, we generate $\boldsymbol{z}$ as follows. Using the Gaussian copula with the correlation matrix $\boldsymbol{\Sigma} = 0.2 \cdot \mathbf{1}_{d_2}\mathbf{1}_{d_2}^T + 0.8\boldsymbol{I}_{d_2}$, we generate d_2-dimensional random vectors with support in $[0,1]^{d_2}$. Then we apply inverse transformation on each coordinate to make it marginally distributed as t_7. Note that under our setup $\boldsymbol{z}$ has dependent coordinates with each coordinate being t_7 distribution. However, the true covariance matrix of $\boldsymbol{z}$ is not $\boldsymbol{\Sigma}$ anymore and is even unknown. According to Theorem 9, we



(a) Gaussian (b) Beta (c) Gamma

(d) t_{13} (e) Rayleigh (f) Weibull

$f^{(1)}$ $f^{(2)}$ $f^{(3)}$ $f^{(4)}$ $f^{(5)}$ $f^{(6)}$

Figure 3: Low-rank matrix estimation plot. This figure shows $\|\hat{\boldsymbol{B}}-\tilde{\boldsymbol{B}}\|_F$ error of estimating low-rank parameter matrix in model (1). Six lines indicates six different types of link functions. Components of $\boldsymbol{z}$ are independent.

set $\kappa_1 = 2\sqrt{\log(d_1+d_2)/n(d_1+d_2)}$ and $\lambda = 12\sqrt{(d_1+d_2)\log(d_1+d_2)/n}$. The precision matrix estimator we use is defined in (14) with $\kappa_2 = 2\sqrt{\log d_2/nd_2}$, suggested by Lemma 12.

Figure 3 and 4 summarize our results. We observe a linear trend for all link functions, though the quadratic link functions result in a larger error overall. Also, for each distribution, dependent or independent $\boldsymbol{z}$ have very similar error trends among all six types of link functions. This observation verifies our theoretical results in Theorem 9 and Lemma 12 that estimating precision matrix only contributes higher order error.

7.3. Sparse Matrix

For the sparse matrix recovery, we only consider estimating fully sparse $\boldsymbol{B}^\star$ with dependent covariate $\boldsymbol{z}$. We let $d_1 = 100$, $d_2 = 50$, $s = 10$. The considered distributions for design $\boldsymbol{x}$ are in Table 2. $\boldsymbol{B}^\star$ is generated by first generating the support of nonzero entries and then setting each entry on the support to be $\{-\frac{1}{\sqrt{s}}, \frac{1}{\sqrt{s}}\}$ with equal probability. For generating dependent $\boldsymbol{z}$, we define the sparse precision matrix $\Theta = (\theta_{ij})_{i,j=1}^{d_2}$ as follows

$$\theta_{ij} = \begin{cases} 1 & \text{if } i = j, \\ 0.2 & \text{if } |i-j| = 1, \\ 0 & \text{otherwise.} \end{cases}$$

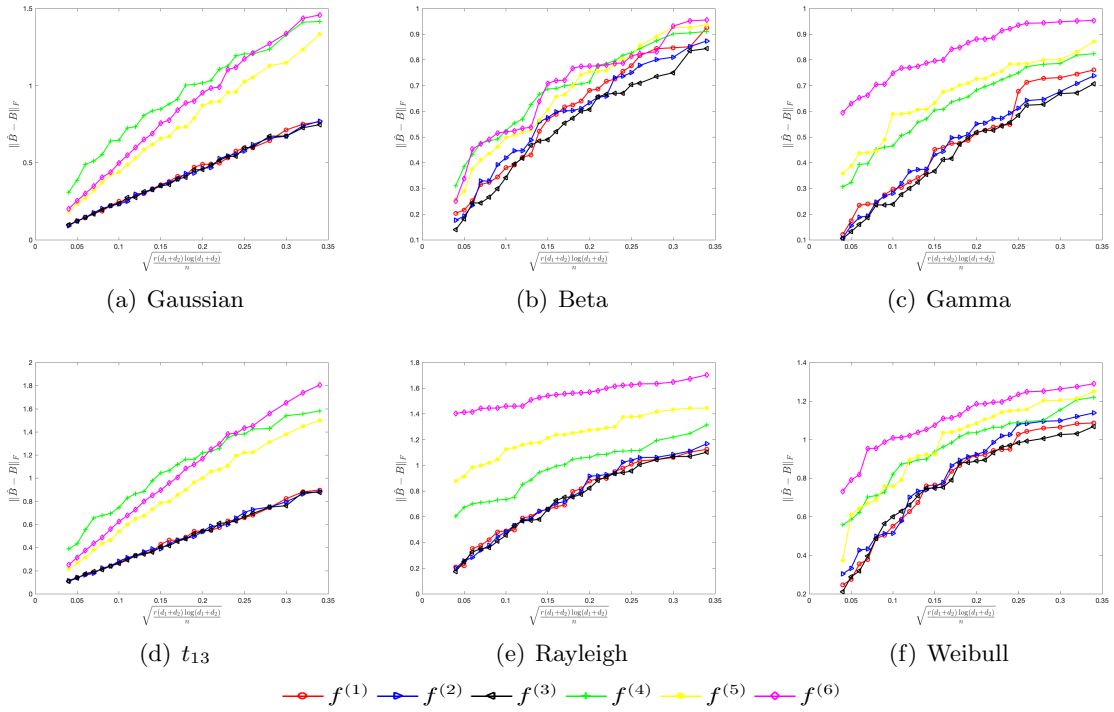


Figure 4: Low-rank matrix estimation plot. This figure shows $\|\hat{\boldsymbol{B}}-\tilde{\boldsymbol{B}}\|_F$ error of estimating low-rank parameter matrix in model (1). Six lines indicates six different types of link functions. Components of $\boldsymbol{z}$ are dependent.

Then we calculate the corresponding covariance matrix Θ^{-1} and convert it to correlation matrix. Following the same steps as in the previous section, we use Gaussian copula to derive dependent covariates $\boldsymbol{z}$, whose marginal distribution for each coordinate is t_7. The precision matrix of $\boldsymbol{z}$ has the same sparsity structure as Θ, even though it is not equal to Θ. According to Theorem 13, we set $\tau = 2(n/\log d_1 d_2)^{1/6}$ and $\lambda = 10\sqrt{\log d_1 d_2/n}$. The sparse precision matrix estimator is defined in (A.1) and (A.2) with truncation threshold $2(n/\log d_2)^{1/4}$ and $\gamma = 10\sqrt{\log d_2/n}$. We use default settings in CVX package (Grant and Boyd, 2008, 2012) to solve (A.1) efficiently. The error plot is shown in Figure 5 and linear trends for errors appear again, as explained by Theorem 13.

7.4. Real Data Application

In this subsection, we will further illustrate the proposed method by analyzing a *Coffea canephora* genetic data set.[3] The data set collects three important traits (phenotypes): production of coffee beans (y_1), leaf rust incidence (y_2) and yield of green beans (y_3), from two recurrent selection populations (x_1) of *Coffea canephora* and each one of traits is evaluated at two locations (x_2). For each individual sample, the single nucleotide polymor-

3. The data set was obtained from Dryad at https://datadryad.org/resource/doi:10.5061/dryad.1139fm7.

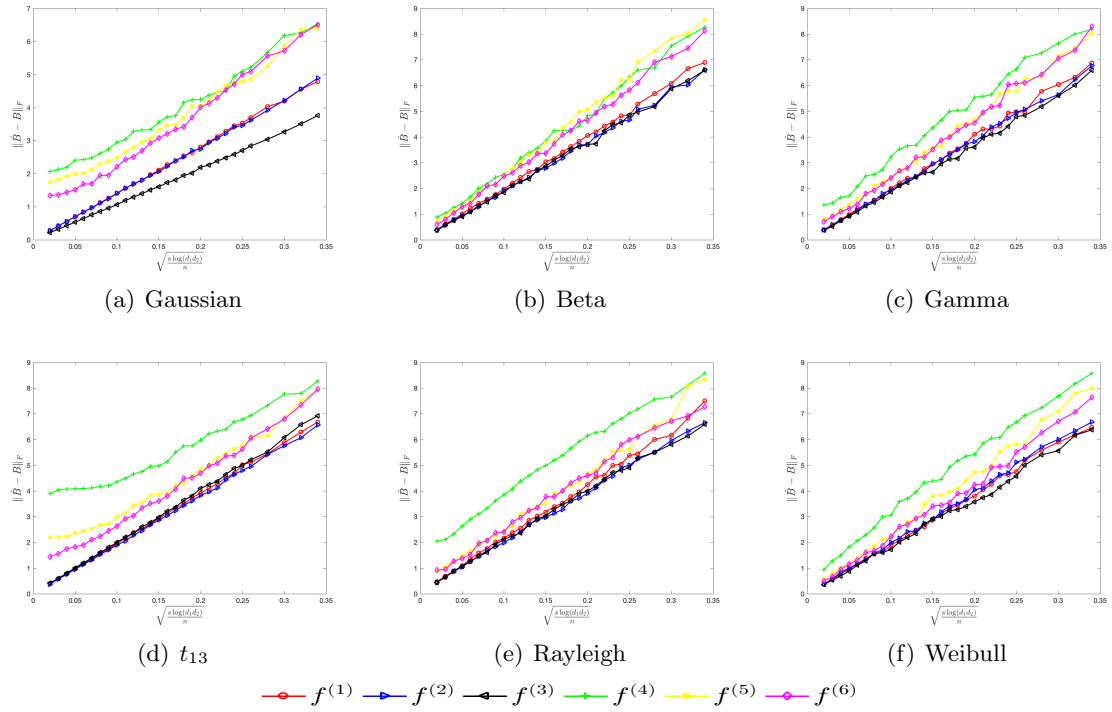


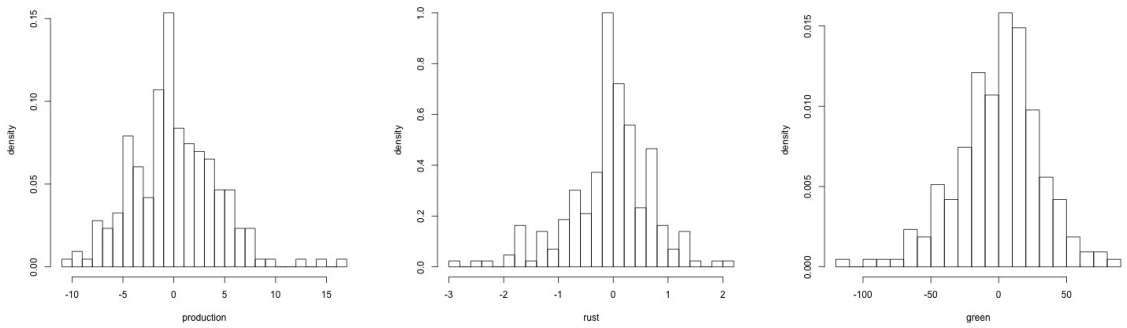
Figure 5: Sparse matrix estimation plot. This figure shows $\|\hat{\boldsymbol{B}} - \tilde{\boldsymbol{B}}\|_F$ error of estimating sparse parameter matrix in model (1) with dependent $\boldsymbol{z}$. Six lines indicates six different types of link functions. All above simulation results are consistent with Theorem 13.

phisms (SNPs)—genotype $\{z_j\}$—are identified by Genotyping-by-Sequencing. In particular, the population group (x_1) and the evaluation location (x_2) are two confounders which may modify the effect of each SNP (z_j) on traits (y_1, y_2, y_3). To study this higher-level effect from confounders, we consider the varying index coefficient model in (1), which in this case is written as

$$y_i = \sum_{j=1}^{d_2} z_j \cdot f_j\big((\beta_j^{1,i})^\star x_1 + (\beta_j^{2,i})^\star x_2\big) + \epsilon, \quad \forall i = 1, 2, 3, \tag{17}$$

and focus on estimating parameter matrices $\boldsymbol{B}_i^\star = \big((\boldsymbol{\beta}^{1,i})^\star, (\boldsymbol{\beta}^{2,i})^\star\big)^T \in \mathbb{R}^{2\times d_2}$.

We follow the preprocessing steps described in Ferrão et al. (2018). The data set has $n = 215$ samples evaluated at two locations in total, where $n_1 = 119$ of them are collected from the first population group with 45748 SNPs measured, while $n_2 = 96$ of them are collected from the second population group with 59332 SNPs measured. There are 38106 SNPs in common and we select $d_2 = 250$ from them uniformly at random. To make individuals independent from each other, in each group we only use the data evaluated at the first location for the first half of individuals and the data evaluated at the second location for the second half of individuals. Continuous distributions are used for the confounders x_1



(a) Production (b) Leaf rust (c) Green beans

Figure 6: Histograms for three traits: production of coffee beans, leaf rust incidence and yield of green beans.

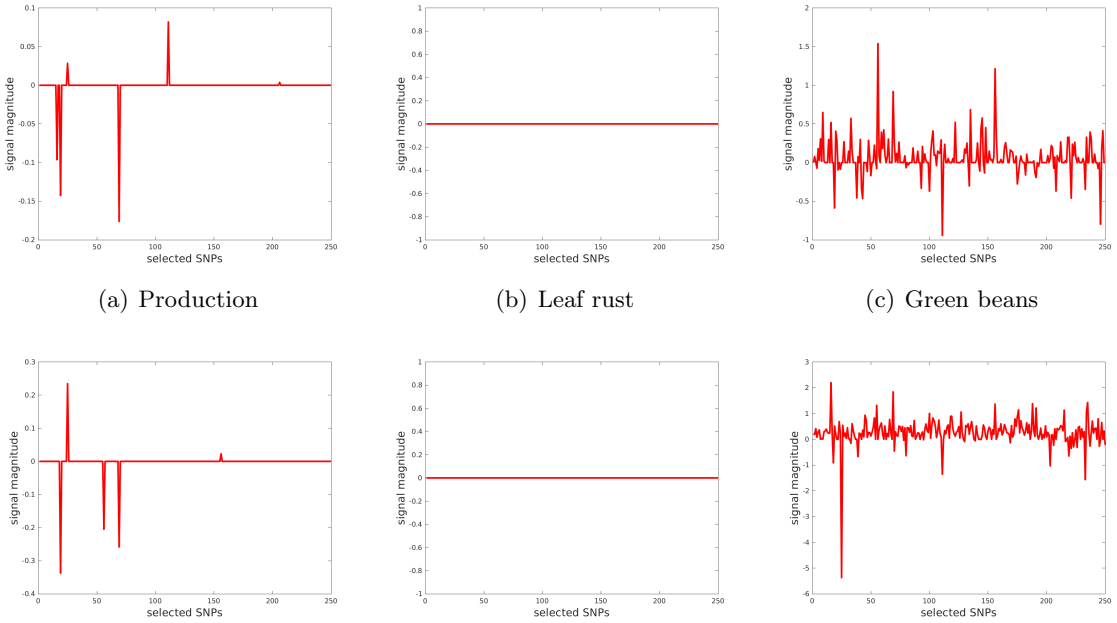
and x_2. In particular, we let $x_1 \sim N(0,1)$, if the observation is from the first population group and $x_1 \sim N(50,1)$, if it is from the second population group. Analogously, we have $x_2 \sim t_{13}$, if the observation is evaluated at the first location and $x_2 \sim 50 + t_{13}$, if it is evaluated at the second location. Here t_{13} refers to the t distribution with degree of freedom 13. Under our setup, the confounders come from the following mixture distribution:

$$x_1 \sim \frac{n_1}{n} N(0,1) + \frac{n_2}{n} N(50,1), \quad x_2 \sim \frac{1}{2} t_{13} + \frac{1}{2}(50 + t_{13}). \tag{18}$$

The corresponding score functions are computed based on the above distributions. We note that the data set is a high-dimensional one, since $n < d_2$, and $\boldsymbol{B}_i^\star$ cannot be estimated using many of the related works, where typically $d_2 < 5$ (Huang and Zhang, 2013; Guo et al., 2016; Zhao et al., 2017). The histograms for three traits are provided in Figure 6. Note that they have been normalized as in Ferrão et al. (2018). We also refer to Ferrão et al. (2018) for other basic statistical analysis and detailed preprocessing steps on this data set.

For each $i = 1, 2, 3$, we assume that $\boldsymbol{B}_i^\star$ is fully sparse. We compute the sparse matrix estimator via (16) with $\tau = \big(n/\log d_1 d_2\big)^6$ and $\lambda = \sqrt{\log d_1 d_2 / n}$. The sparse precision matrix is estimated by conducting CLIME procedure with $\gamma = 5\sqrt{\log d_2 / n}$. Throughout the experiment, we apply the hard truncation on confounders only since other variables are light-tailed already (for example, SNPs take value in $\{-1, 0, 1\}$). The estimated signal trajectories are drawn in Figure 7. From the plots, we observe that when fitting y_1 under the model (17), both confounders may cause different effect sizes for 5 out of 250 selected SNPs, and 3 of them affect common SNPs. However, confounders have no effects on selected SNPs when fitting y_2, while have dense effects when fitting y_3. In the latter case, we should mention the signals with large magnitude are still sparse.

In addition, we provide the 95% confidence interval for each trajectory using a non-parametric bootstrap: we sample the data set with replacement for 100 times and for each data set we estimate $\boldsymbol{B}_i^\star$ using the same method with the same parameters, then compute 0.025 and 0.975 quantiles to construct the confidence interval. The results are shown



(a) Production (b) Leaf rust (c) Green beans

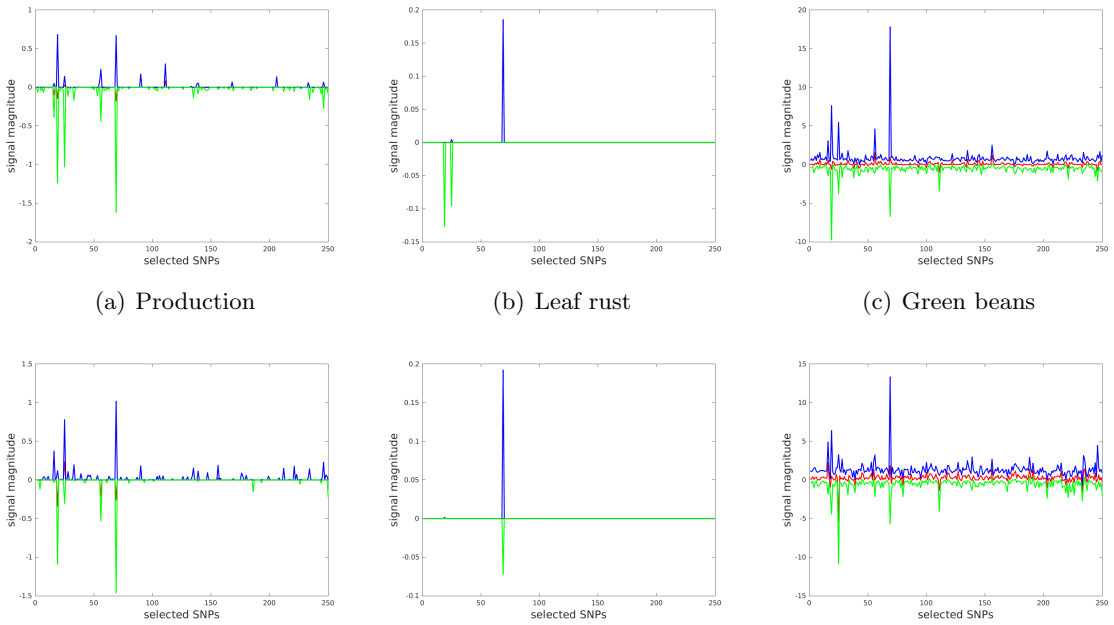
(d) Production (e) Leaf rust (f) Green beans

Figure 7: Signal trajectories. Each plot corresponds to a pair of confounders and responses, i.e. (l, i) plot shows the trajectory of $(\boldsymbol{\beta}^{l,i})^\star$ where $l = 1, 2$ indexes the confounder and $i = 1, 2, 3$ indexes the response.

in Figure 8 where the width of confidence interval indicates the variance of our estimators. We see some of identified signals have large variance but all unidentified signals have small variance. As a complement of analysis, we also conduct parametric bootstrapping: we generate new confounders x_1, x_2 from distribution in (18) for 10000 times, and for each new data set we estimate $\boldsymbol{B}_i^\star$ with the same setup, then use quantiles to construct confidence interval. The plots for parametric bootstrapping are provided in Figure 13 in Appendix D and similar observations appear again. Our code is available for download at: `https://github.com/senna1128/Varying-Index-Coefficient-Models`.

8. Conclusion

In this paper, we proposed new estimators based on Stein's identity for varying index coefficient models. By utilizing score function, we can either estimate a single sparse vector or a low-rank/sparse parameter matrix. The estimator can handle dependent covariates $\boldsymbol{z}$ through estimation of the precision matrix and can achieve optimal convergence rate in sparse estimation and near optimal rate in low-rank estimation. In all cases, the estimators we proposed have closed form and are easy to implement. Instead of assuming that covariates $\boldsymbol{x}$ follow an elliptical distribution, we only require certain finite moment assumption



(a) Production (b) Leaf rust (c) Green beans

(d) Production (e) Leaf rust (f) Green beans

Figure 8: Signal trajectories for 95% confidence interval (nonparametric bootstrapping). The (l, i) plot shows the confidence interval trajectory of $(\boldsymbol{\beta}^{l,i})^\star$ where $l = 1, 2$ indexes the confounder and $i = 1, 2, 3$ indexes the response. The blue line indicates the upper bound, the red line indicates the estimator, and the green line indicates the lower bound.

on response y, covariates $\boldsymbol{z}$, and score variable $S(\boldsymbol{x})$. We also conduct extensive numerical experiments to illustrate our result.

There are still many open problems worth exploring. One of future work is about finite moment assumption. Under the general sparsity assumption, we believe that finite 6-th moment condition is mild enough, however whether it is possible to relax it further is not clear. Furthermore, we note that all first-order Stein's estimators suffer from $\mu_k = \mathbb{E}[f_k'(\langle \boldsymbol{x}, \boldsymbol{\beta}_k^\star \rangle)] = 0$. Therefore developing a second-order Stein's estimator is of practical interest and will be explored elsewhere.

Acknowledgments

This work was completed in part with resources provided by the University of Chicago Research Computing Center. We thank Kenji Fukumizu and two anonymous reviewers for helpful comments.

Appendix A. Estimate of Sparse Precision Matrix

We propose an approach to estimate a high-dimensional sparse precision matrix for heavy-tailed variable. Suppose $\boldsymbol{z}$ has finite 4th moment and $\boldsymbol{\Omega}^\star = (\boldsymbol{\Sigma}^\star)^{-1}$, with $\boldsymbol{\Sigma}^\star = \mathbb{E}[\boldsymbol{z}\boldsymbol{z}^T]$, is column sparse. In particular, we assume that $\boldsymbol{\Omega}^\star \in \mathcal{F}_w^{K4}$ for some w and K. In this setting, we estimate the precision matrix using the CLIME procedure (Cai et al., 2011)

$$\begin{aligned} &\min \|\boldsymbol{\Omega}\|_{1,1}, \\ &\text{s.t.} \quad \|\hat{\boldsymbol{\Sigma}}\boldsymbol{\Omega} - \boldsymbol{I}_{d_2}\|_{\max} \leqslant \gamma \end{aligned} \tag{A.1}$$

but with

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n} \widetilde{\boldsymbol{Z}}_i \widetilde{\boldsymbol{Z}}_i^T \tag{A.2}$$

being a thresholded estimator of the covariance matrix for some threshold $\tau > 0$, and γ is a tuning parameter. The linear program in (A.1) is the same as in Cai et al. (2011), with the difference that we use an estimator of $\boldsymbol{\Sigma}^\star$ that is more suitable for heavy-tailed data.

Lemma 15 *If* $\tau = (M_4 n/\log d_2)^{1/4}/2$ *and* $\gamma = 12\|\boldsymbol{\Omega}^\star\|_1\sqrt{M_4 \log d_2/n}$, *the estimator* (A.1) *satisfies*

$$P\left(\|\hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}^\star\|_2 \leqslant 96\|\boldsymbol{\Omega}^\star\|_1^2 w\sqrt{M_4 \log d_2/n}\right) \geqslant 1 - \frac{2}{d_2^2},$$

and

$$P\left(\|\hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}^\star\|_{\max} \leqslant 48\|\boldsymbol{\Omega}^\star\|_1^2\sqrt{M_4 \log d_2/n}\right) \geqslant 1 - \frac{2}{d_2^2}.$$

From above lemma, we see the setting for γ in (A.1) is oracle in the sense that $\|\boldsymbol{\Omega}^\star\|_1$ is unknown. Cai et al. (2011) showed a detailed discussion on this aspect and this dependence could be removed by using a self-calibrated estimator, similar to scaled lasso (Sun and Zhang, 2013). We should also mention that (A.1) achieves the optimal rate (Cai et al., 2016).

Appendix B. Proofs of Lemmas

Throughout the proof, we frequently utilize the Bernstein's inequality presented in Corollary 2.11 in Boucheron et al. (2013). To simplify subsequent presentation, we define a function to denote the common upper bound:

$$\varphi(t, a, b) = \exp(-\frac{t^2/2}{a + b \cdot t/3}).$$

As shown in Bernstein's inequality, usually a measures the total variance and b is bound for a single variable. We also use M as the substitute of M_p (p is certain moment) for simplicity. We summarize all structures we used in the paper for future reference.

4. See definition in Section 5.

Assumption 16 (Column-wise sparse) *We assume* $\|\boldsymbol{\beta}_k^\star\|_0 \leqslant s$, $\forall k \in [d_2]$.

Assumption 17 (Fully sparse) *We assume* $\boldsymbol{B}^\star$ *is* s*-sparse:* $\|\boldsymbol{B}^\star\|_{0,1} = |supp(\boldsymbol{B}^\star)| \leqslant s$.

Assumption 18 (Low-rank) *We assume* $\boldsymbol{B}^\star$ *satisfies* $rank(\boldsymbol{B}^\star) \leqslant r$.

Assumption 19 (Independence) *We assume* $\boldsymbol{z}$ *satisfies* $\mathbb{E}[z_i z_j] = 0$, $\forall i \neq j \in [d_2]$.

Assumption 20 (Precision matrix restriction) *Define* $\boldsymbol{\Sigma}^\star = \mathbb{E}[\boldsymbol{z}\boldsymbol{z}^T]$ *and let* $\boldsymbol{\Omega}^\star = (\boldsymbol{\Sigma}^\star)^{-1}$, *we assume*

$$\boldsymbol{\Omega}^\star \in \mathcal{F}_w^K = \left\{ \boldsymbol{\Omega} \in \mathbb{R}^{d_2 \times d_2} : \|\boldsymbol{\Omega}\|_{0,\infty} \leqslant w, \|\boldsymbol{\Omega}\|_2 \leqslant K, \|\boldsymbol{\Omega}^{-1}\|_2 \leqslant K \right\}$$

for some w *and* K.

B.1. Proof of Lemma 4

Under Assumption 2, we can get from (7) that

$$\nabla \hat{L}_k(\tilde{\boldsymbol{\beta}}_k) = 2\tilde{\boldsymbol{\beta}}_k - \frac{2}{n}\sum_{i=1}^n y_i Z_{ik} \boldsymbol{X}_i \overset{(5)}{=} 2(\mathbb{E}[y z_k \cdot \boldsymbol{x}] - \frac{1}{n}\sum_{i=1}^n y_i Z_{ik} \boldsymbol{X}_i).$$

So, for any fixed $j \in [d_1]$, we have

$$[\nabla \hat{L}_k(\tilde{\boldsymbol{\beta}}_k)]_j = 2(\mathbb{E}[y z_k \cdot x_j] - \frac{1}{n}\sum_{i=1}^n y_i Z_{ik} X_{ij}). \tag{B.1}$$

Note that $z_k x_j$ is a sub-exponential random variable with

$$\|z_k x_j\|_{\psi_1} \leqslant \|z_k\|_{\psi_2} \|x_j\|_{\psi_2} \leqslant \Upsilon_{\boldsymbol{z}} \Upsilon_{\boldsymbol{x}}, \tag{B.2}$$

where $\Upsilon_{\boldsymbol{x}}$ is ψ_2-norm of a standard Gaussian variable. So we see $\{y_i, Z_{ik} X_{ij}\}_{i\in[n]}$ are n independent copies of y and $z_k x_j$. Based on Lemma C.4 in Yang et al. (2019) and (B.2), let $\gamma = \max(\Upsilon_y, \Upsilon_{\boldsymbol{x}} \Upsilon_{\boldsymbol{z}})$ and we get

$$P(|\frac{1}{n}\sum_{i=1}^n y_i Z_{ik} X_{ij} - \mathbb{E}[y z_k \cdot x_j]| > \Upsilon_\gamma \sqrt{\frac{\log n}{n}}) < \frac{1}{n^2}$$

where $\Upsilon_\gamma > 0$ only depends on γ. According to equation (B.1), we can take union bound and further have

$$P(\|\nabla \hat{L}_k(\tilde{\boldsymbol{\beta}}_k)\|_\infty > 2\Upsilon_\gamma \sqrt{\frac{\log n}{n}}) < \frac{d_1}{n^2},$$

which concludes the proof.

B.2. Proof of Lemma 8

Based on equation (8), we know

$$\nabla \bar{L}_k(\tilde{\boldsymbol{\beta}}_k) = 2\tilde{\boldsymbol{\beta}}_k - \frac{2}{n}\sum_{i=1}^{n} \breve{y}_i \widetilde{Z_{ik}} \widetilde{S(\boldsymbol{X}_i)}.$$

Under Assumption 2 and 19, we know $\tilde{\boldsymbol{\beta}}_k = \mathbb{E}[yz_k \cdot S(\boldsymbol{x})]$. So we can separate above gradient into two parts

$$\begin{aligned}
&\|\nabla \bar{L}_k(\tilde{\boldsymbol{\beta}}_k)\|_\infty = 2\|\tilde{\boldsymbol{\beta}}_k - \frac{1}{n}\sum_{i=1}^{n} \breve{y}_i \widetilde{Z_{ik}} \widetilde{S(\boldsymbol{X}_i)}\|_\infty \\
\leqslant & 2\| \underbrace{\mathbb{E}[yz_k \cdot S(\boldsymbol{x})] - \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\breve{y}_i \widetilde{Z_{ik}} \widetilde{S(\boldsymbol{X}_i)}]}_{\mathcal{I}_1} \|_\infty + 2\| \underbrace{\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\breve{y}_i \widetilde{Z_{ik}} \widetilde{S(\boldsymbol{X}_i)}] - \frac{1}{n}\sum_{i=1}^{n} \breve{y}_i \widetilde{Z_{ik}} \widetilde{S(\boldsymbol{X}_i)}}_{\mathcal{I}_2} \|_\infty.
\end{aligned} \tag{B.3}$$

We will provide a deterministic bound for $\mathcal{I}_1$ and probabilistic bound for $\mathcal{I}_2$. Let's deal with $\mathcal{I}_1$ first. For any $j \in [d_1]$, we know

$$\begin{aligned}
\mathcal{I}_{1j} =& \mathbb{E}[yz_k \cdot S(\boldsymbol{x})_j] - \mathbb{E}[\breve{y}\breve{z}_k \cdot \widetilde{S(\boldsymbol{x})}_j] \\
=& \mathbb{E}\big[yz_k \cdot S(\boldsymbol{x})_j \cdot \mathbf{1}_{|y|>\tau \text{ or } |z_k|>\tau \text{ or } |S(\boldsymbol{x})_j|>\tau}\big] \\
\leqslant& \sqrt{\mathbb{E}[y^2 \boldsymbol{z}_k^2 S(\boldsymbol{x})_j^2] \cdot \big(P(|y|>\tau) + P(|\boldsymbol{z}_k|>\tau) + P(|S(\boldsymbol{x})_j|>\tau)\big)} \\
\leqslant& \sqrt[4]{\mathbb{E}[y^4]\mathbb{E}[z_k^4]\mathbb{E}[S(\boldsymbol{x})_j^4]} \frac{\sqrt{3}M^{1/2}}{\tau^3} \\
\leqslant& \frac{2M}{\tau^3}.
\end{aligned} \tag{B.4}$$

Here, the third inequality is from Cauchy-Schwarz inequality; the fourth inequality is Chebyshev inequality; the last inequality is due to Assumption 6 ($p = 6$). So from (B.4), we know

$$\|\mathcal{I}_1\|_\infty \leqslant 2M/\tau^3. \tag{B.5}$$

For the $\mathcal{I}_2$ term in equation (B.3), we apply Bernstein's inequality. We have $\forall j \in [d_1]$,

$$\begin{aligned}
&-\tau^3 \leqslant \breve{y}_i \widetilde{Z_{ik}} \widetilde{S(\boldsymbol{X}_i)}_j \leqslant \tau^3 \Longrightarrow C = 2\tau^3, \\
&V_n = \sum_{i=1}^{n} \mathrm{Var}(\breve{y}_i \widetilde{Z_{ik}} \widetilde{S(\boldsymbol{X}_i)}_j) \leqslant \sum_{i=1}^{n} \mathbb{E}[\breve{y}_i^2 \widetilde{Z_{ik}}^2 \widetilde{S(\boldsymbol{X}_i)}_j^2] \leqslant nM.
\end{aligned} \tag{B.6}$$

Thus, based on (B.6), we have $\forall t > 0$,

$$P\big(\big|\mathbb{E}[\breve{y}\breve{z}_k \cdot \widetilde{S(\boldsymbol{x})}_j] - \frac{1}{n}\sum_{i=1}^{n} \breve{y}_i \widetilde{Z_{ik}} \widetilde{S(\boldsymbol{X}_i)}_j\big| > t\big) \leqslant 2\varphi(nt, nM, 2\tau^3). \tag{B.7}$$

Then we take union bound for (B.7) and get

$$P(\|\mathcal{I}_2\|_\infty > t) \leqslant 2d_1\varphi(nt, nM, 2\tau^3) \tag{B.8}$$

Combine (B.5) and (B.8) together, and take union bound over k, we have $\forall t, \tau > 0$,

$$P(\|\nabla \bar{L}_k(\tilde{\boldsymbol{\beta}}_k)\|_\infty \leqslant \frac{4M}{\tau^3} + 2t, \quad \forall k \in [d_2]) \geqslant 1 - 2d_1d_2\exp(-\frac{nt^2}{2M + 2\tau^3 t}). \tag{B.9}$$

Suppose, for some positive constant c_1, c_2, we let

$$t = c_1\sqrt{\log d_1 d_2/n} \quad \text{and} \quad \tau = c_2^{1/3}(n/\log d_1 d_2)^{1/6}. \tag{B.10}$$

Then plug (B.10) into the right hand side of (B.9) and get

$$2d_1d_2\exp(-\frac{nt^2}{2M + 2\tau^3 t}) = 2d_1d_2\exp(-\frac{c_1^2\log d_1 d_2}{2M + 2c_1c_2}) \leqslant 2/d_1^2d_2^2, \tag{B.11}$$

if

$$\frac{c_1^2}{2M + 2c_1c_2} \geqslant 3 \Longrightarrow c_1^2 - 6c_1c_2 - 6M \geqslant 0. \tag{B.12}$$

We can set $c_1 = 3\sqrt{M}$ and $c_2 = \sqrt{M}/8$, which satisfies condition in (B.12) naturally. Further we know (B.11) holds. Plug this setting in (B.10) and (B.9) and we get

$$\|\nabla \bar{L}_k(\tilde{\boldsymbol{\beta}}_k)\|_\infty \leqslant 38\sqrt{M\log d_1 d_2/n}, \quad \forall k \in [d_2] \tag{B.13}$$

with probability at least $1 - 2/d_1^2d_2^2$. This finishes the proof.

B.3. Proof of Lemma 10

Define $\mathcal{I}_3 = \frac{1}{n\kappa_1}\sum_{i=1}^n \Phi(\kappa_1 y_i \cdot S(\boldsymbol{X}_i)\boldsymbol{Z}_i^T) - \mathbb{E}[y \cdot S(\boldsymbol{x})\boldsymbol{z}^T]$, we will apply Corollary 3.1 in Minsker (2018). Let's first bound the variance. Under Assumption 2, we know $\mathbb{E}[S(\boldsymbol{x})_j] = 0, \forall j \in [d_1]$. So for any unit vector $\boldsymbol{v} \in \mathbb{R}^{d_1}$, we have

$$\begin{aligned}
&\mathbb{E}[y^2 \cdot \boldsymbol{v}^T S(\boldsymbol{x})\boldsymbol{z}^T\boldsymbol{z}S(\boldsymbol{x})^T\boldsymbol{v}] = \mathbb{E}[y^2 \cdot \boldsymbol{z}^T\boldsymbol{z} \cdot (S(\boldsymbol{x})^T\boldsymbol{v})^2] \leqslant \sqrt{\mathbb{E}[y^4]\mathbb{E}[(\boldsymbol{z}^T\boldsymbol{z})^2]\mathbb{E}[(S(\boldsymbol{x})^T\boldsymbol{v})^4]} \\
&\leqslant M^{1/2}\sqrt{\mathbb{E}[d_2(\boldsymbol{z}_1^4 + ... + \boldsymbol{z}_{d_2}^4)]}\sqrt{\mathbb{E}[\sum_{i_1=1}^{d_1}\sum_{i_2=1}^{d_1} S(\boldsymbol{x})_{i_1}^2 S(\boldsymbol{x})_{i_2}^2 \boldsymbol{v}_{i_1}^2\boldsymbol{v}_{i_2}^2]} \\
&\leqslant d_2M\sqrt{\sum_{i_1=1}^{d_1}\sum_{i_2=1}^{d_1}\mathbb{E}[S(\boldsymbol{x})_{i_1}^2 S(\boldsymbol{x})_{i_2}^2]\boldsymbol{v}_{i_1}^2\boldsymbol{v}_{i_2}^2} \leqslant d_2M\sqrt{\sum_{i_1=1}^{d_1}\sum_{i_2=1}^{d_1}\sqrt{\mathbb{E}[S(\boldsymbol{x})_{i_1}^4]}\sqrt{\mathbb{E}[S(\boldsymbol{x})_{i_2}^4]}\boldsymbol{v}_{i_1}^2\boldsymbol{v}_{i_2}^2} \\
&\leqslant d_2M^{3/2}.
\end{aligned} \tag{B.14}$$

The second inequality uses Cauchy-Schwarz inequality; the third inequality uses Assumption 2. From (B.14) we can get

$$\|\mathbb{E}[y^2 \cdot S(\boldsymbol{x})\boldsymbol{z}^T\boldsymbol{z}S(\boldsymbol{x})^T]\|_2 \leqslant d_2M^{3/2}. \tag{B.15}$$

Follow the exactly same derivation in (B.14) we can also have

$$\|\mathbb{E}[y^2 \cdot \boldsymbol{z}S(\boldsymbol{x})^T S(\boldsymbol{x})\boldsymbol{z}^T]\|_2 \leqslant d_1 M^{3/2}. \tag{B.16}$$

Thus, combine (B.15) and (B.16) together, we have $\forall t > 0$

$$P(\|\mathcal{I}_3\|_2 \geqslant t) \leqslant 2(d_1+d_2)\exp(-n\kappa_1 t + \frac{n(d_1+d_2)M^{3/2}\kappa_1^2}{2}). \tag{B.17}$$

In above tail bound (B.17), we let $t = 2M^{3/4}\sqrt{\frac{2(d_1+d_2)\log(d_1+d_2)}{n}}$ and $\kappa_1 = \sqrt{\frac{2\log(d_1+d_2)}{n(d_1+d_2)M^{3/2}}}$ and have

$$P\bigg(\|\mathcal{I}_3\| \leqslant 2M^{3/4}\sqrt{\frac{2(d_1+d_2)\log(d_1+d_2)}{n}}\bigg) \geqslant 1 - \frac{2}{(d_1+d_2)^2}. \tag{B.18}$$

This is consistent with argument of lemma.

B.4. Proof of Lemma 12

Let's first get concentration rate for $\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^\star\|_2 = \|\frac{1}{n\kappa_2}\Phi(\kappa_2 Z_i Z_i^T) - \mathbb{E}[\boldsymbol{z}\boldsymbol{z}^T]\|_2$. We have $\forall \boldsymbol{v} \in \mathbb{R}^{d_2}$ such that $\|\boldsymbol{v}\|_2 = 1$,

$$\mathbb{E}[\boldsymbol{v}^T\boldsymbol{z}\boldsymbol{z}^T\boldsymbol{v}] = \mathbb{E}[(\boldsymbol{v}^T\boldsymbol{z})^2] \leqslant \mathbb{E}[\|\boldsymbol{z}\|_2^2] \leqslant d_2\sqrt{M}.$$

Based on Corollary 3.1 in Minsker (2018), we know $\forall t > 0$,

$$P(\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^\star\|_2 \geqslant t) \leqslant 2d_2\exp(-n\kappa_2 t + \frac{nd_2\sqrt{M}\kappa_2^2}{2}). \tag{B.19}$$

In above (B.19), we let $t = 2M^{1/4}\sqrt{\frac{2d_2\log d_2}{n}}$ and $\kappa_2 = \sqrt{\frac{2\log d_2}{nd_2M^{1/2}}}$, and have

$$P\bigg(\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^\star\|_2 \leqslant 2M^{1/4}\sqrt{\frac{2d_2\log d_2}{n}}\bigg) \geqslant 1 - \frac{2}{d_2^2}. \tag{B.20}$$

We use matrix perturbation analysis to give bound for $\hat{\boldsymbol{\Omega}}$. As shown in Chapter III Theorem 2.5 in Stewart and Sun (1990), when

$$\|\boldsymbol{\Omega}^\star(\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^\star)\|_2 \leqslant \|\boldsymbol{\Omega}^\star\|_2\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^\star\|_2 \leqslant \|\boldsymbol{\Omega}^\star\|_2 4M^{1/4}\sqrt{d_2\log d_2/n} \leqslant 1/2,$$

we know $\hat{\boldsymbol{\Omega}}$ is perforce invertible and satisfies

$$\|\hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}^\star\|_2 \leqslant 2\|\boldsymbol{\Omega}^\star\|_2^2\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^\star\|_2 \leqslant 8\|\boldsymbol{\Omega}^\star\|_2^2 M^{1/4}\sqrt{d_2\log d_2/n}, \tag{B.21}$$

with probability at least $1 - 2/d_2^2$. Therefore we finish the proof.

B.5. Proof of Lemma 14

We define $\mathcal{I}_7 = \mathbb{E}[y \cdot S(\boldsymbol{x})\boldsymbol{z}^T] - \frac{1}{n}\sum_{i=1}^n \breve{y}_i \cdot \widetilde{S(\boldsymbol{X}_i)}\breve{\boldsymbol{Z}}_i^T$. For any $j \in [d_1], k \in [d_2]$, we know

$$|(\mathcal{I}_7)_{jk}| \leqslant |\frac{1}{n}\sum_{i=1}^n \breve{y}_i \widetilde{Z_{ik}} \widetilde{S(\boldsymbol{X}_i)}_j - \mathbb{E}[\breve{y}_i \widetilde{Z_{ik}} \widetilde{S(\boldsymbol{X}_i)}_j]| + |\frac{1}{n}\sum_{i=1}^n \mathbb{E}[\breve{y}_i \widetilde{Z_{ik}} \widetilde{S(\boldsymbol{X}_i)}_j] - \mathbb{E}[y \cdot z_k \cdot S(\boldsymbol{x})_j]|.$$

Use the same technique as in (B.4)-(B.8), we can get that $\forall t, \tau > 0$

$$P(\|\mathcal{I}_7\|_{\max} > t + \frac{2M}{\tau^3}) \leqslant 2d_1 d_2 \exp(-\frac{nt^2}{2M + 2\tau^3 t}).$$

We let $t = 3\sqrt{M \log d_1 d_2 / n}$, $\tau = (Mn/\log d_1 d_2)^{1/6}/2$ and have

$$P\left(\|\mathcal{I}_7\|_{\max} \leqslant 19\sqrt{\frac{M \log d_1 d_2}{n}}\right) \geqslant 1 - 2/d_1^2 d_2^2. \tag{B.22}$$

So we finish the proof of lemma.

B.6. Proof of Lemma 15

We first prove a concentration bound for truncated empirical covariance $\hat{\boldsymbol{\Sigma}}$. $\forall j, k \in [d_2]$,

$$\begin{aligned} |\hat{\Sigma}_{jk} - \Sigma^\star_{jk}| &= |\frac{1}{n}\sum_{i=1}^n \breve{Z}_{ij}\breve{Z}_{ik} - \mathbb{E}[z_j z_k]| \\ &\leqslant \underbrace{|\frac{1}{n}\sum_{i=1}^n (\breve{Z}_{ij}\breve{Z}_{ik} - \mathbb{E}[\breve{Z}_{ij}\breve{Z}_{ik}])|}_{\mathcal{I}_5} + \underbrace{|\frac{1}{n}\sum_{i=1}^n \mathbb{E}[\breve{Z}_{ij}\breve{Z}_{ik}] - \mathbb{E}[z_j z_k]|}_{\mathcal{I}_6}. \end{aligned} \tag{B.23}$$

Use Bernstein's inequality for $\mathcal{I}_5$, we have

$$\begin{aligned} &-\tau^2 \leqslant \breve{Z}_{ij}\breve{Z}_{ik} \leqslant \tau^2, \\ &V_n = \sum_{i=1}^n \mathrm{Var}(\breve{Z}_{ij}\breve{Z}_{ik}) \leqslant \sum_{i=1}^n \mathbb{E}[\breve{Z}_{ij}^2 \breve{Z}_{ik}^2] \leqslant nM. \end{aligned}$$

Note that the above last inequality holds no matter whether $j = k$ or not. Thus, $\forall t > 0$, we have

$$P(|\mathcal{I}_5| > t) \leqslant 2\varphi(nt, nM, 2\tau^2). \tag{B.24}$$

For the term $\mathcal{I}_6$, we know

$$|\mathcal{I}_6| = \mathbb{E}[z_j z_k \cdot \mathbf{1}_{\{|z_j|>\tau \text{ or } |z_k|>\tau\}}] \leqslant \sqrt{\mathbb{E}[z_j^2 z_k^2] \cdot (P(|z_j| > \tau) + P(|z_k| > \tau))} \leqslant \frac{2M}{\tau^2}. \tag{B.25}$$

Combine (B.23), (B.24) and (B.25) together, we get

$$|\hat{\Sigma}_{jk} - \Sigma^\star_{jk}| \leqslant t + \frac{2M}{\tau^2}, \tag{B.26}$$

with probability at least $1-2\exp(-\frac{nt^2}{2M+2\tau^2 t})$. Take union bound for (B.26) we have

$$P(\|\hat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}^\star\|_{\max}\leqslant t+\frac{2M}{\tau^2})\geqslant 1-2d_2^2\varphi(nt,nM,2\tau^2).$$

Let $t=4\sqrt{M\log d_2/n}$, $\tau=(Mn/\log d_2)^{1/4}/2$, we have

$$P\left(\|\hat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}^\star\|_{\max}\leqslant 12\sqrt{\frac{M\log d_2}{n}}\right)\geqslant 1-\frac{2}{d_2^2}. \tag{B.27}$$

Based on this bound, we deal with convex problem (A.1). Suppose $\hat{\boldsymbol{\Omega}}=(\hat{\boldsymbol{\omega}}_1,...,\hat{\boldsymbol{\omega}}_{d_2})$, we will show each $\hat{\boldsymbol{\omega}}_j$ is also a solution to following problem:

$$\begin{aligned}&\min_{\boldsymbol{l}_j}\ \|\boldsymbol{l}_j\|_1,\\ &\text{s.t.}\ \ \|\hat{\boldsymbol{\Sigma}}\boldsymbol{l}_j-\boldsymbol{e}_j\|_\infty\leqslant\gamma.\end{aligned} \tag{B.28}$$

In fact, it's easy to see $\hat{\boldsymbol{\omega}}_j$ is a feasible point for problem (B.28), so $\|\hat{\boldsymbol{l}}_j\|_1\leqslant\|\hat{\boldsymbol{\omega}}_j\|_1$. Further we know $\|(\hat{\boldsymbol{l}}_1,...,\hat{\boldsymbol{l}}_{d_2})\|_{1,1}\leqslant\|\hat{\boldsymbol{\Omega}}\|_{1,1}$. On the other hand, $\|\hat{\boldsymbol{\omega}}_j\|_1\leqslant\|\hat{\boldsymbol{l}}_j\|_1$ for sure. Otherwise $(\hat{\boldsymbol{\omega}}_1,...,\hat{\boldsymbol{l}}_j,...,\hat{\boldsymbol{\omega}}_{d_2})$ satisfies condition of (A.1) but with smaller objective value. In this case, we know each $\hat{\boldsymbol{\omega}}_j$ can also be solved from (B.28). Note for $\boldsymbol{\Omega}^\star=(\boldsymbol{\omega}_1^\star,...,\boldsymbol{\omega}_{d_2}^\star)\in\mathbb{R}^{d_2\times d_2}$, we have

$$\|\hat{\boldsymbol{\Sigma}}\boldsymbol{\Omega}^\star-\boldsymbol{I}_{d_2}\|_{\max}=\|(\hat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}^\star+\boldsymbol{\Sigma}^\star)\boldsymbol{\Omega}^\star-\boldsymbol{I}_{d_2}\|_{\max}=\|(\hat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}^\star)\boldsymbol{\Omega}^\star\|_{\max}\leqslant\|\hat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}^\star\|_{\max}\|\boldsymbol{\Omega}^\star\|_1.$$

So when $\|\hat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}^\star\|_{\max}\|\boldsymbol{\Omega}^\star\|_1\leqslant\gamma$, we know $\boldsymbol{\Omega}^\star$ is feasible for problem (A.1) and $\boldsymbol{\omega}_j^\star$ is feasible for problem (B.28). So we know

$$\|\hat{\boldsymbol{\Omega}}\|_{1,1}\leqslant\|\boldsymbol{\Omega}^\star\|_{1,1}\ \ \text{and}\ \ \|\hat{\boldsymbol{\omega}}_j\|_1\leqslant\|\boldsymbol{\omega}_j^\star\|_1. \tag{B.29}$$

Based on (B.29), we know $\|\hat{\boldsymbol{\Omega}}\|_1\leqslant\|\boldsymbol{\Omega}^\star\|_1$. Further, we have

$$\begin{aligned}\|\boldsymbol{\Sigma}^\star(\hat{\boldsymbol{\Omega}}-\boldsymbol{\Omega}^\star)\|_{\max}\leqslant&\|(\boldsymbol{\Sigma}^\star-\hat{\boldsymbol{\Sigma}})(\hat{\boldsymbol{\Omega}}-\boldsymbol{\Omega}^\star)\|_{\max}+\|\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\Omega}}-\boldsymbol{\Omega}^\star)\|_{\max}\\ \leqslant&\|\boldsymbol{\Sigma}^\star-\hat{\boldsymbol{\Sigma}}\|_{\max}\|\hat{\boldsymbol{\Omega}}-\boldsymbol{\Omega}^\star\|_1+\|\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{\Omega}}-\boldsymbol{I}_{d_2}\|_{\max}+\|\hat{\boldsymbol{\Sigma}}\boldsymbol{\Omega}^\star-\boldsymbol{I}_{d_2}\|_{\max}\\ \leqslant&\|\boldsymbol{\Sigma}^\star-\hat{\boldsymbol{\Sigma}}\|_{\max}(\|\hat{\boldsymbol{\Omega}}\|_1+\|\boldsymbol{\Omega}^\star\|_1)+2\gamma\\ \leqslant&2\|\boldsymbol{\Sigma}^\star-\hat{\boldsymbol{\Sigma}}\|_{\max}\|\boldsymbol{\Omega}^\star\|_1+2\gamma\\ \leqslant&4\gamma.\end{aligned} \tag{B.30}$$

Based on the bound (B.30), we have

$$\|\hat{\boldsymbol{\Omega}}-\boldsymbol{\Omega}^\star\|_{\max}=\|\boldsymbol{\Omega}^\star\boldsymbol{\Sigma}^\star(\hat{\boldsymbol{\Omega}}-\boldsymbol{\Omega}^\star)\|_{\max}\leqslant\|\boldsymbol{\Omega}^\star\|_\infty\|\boldsymbol{\Sigma}^\star(\hat{\boldsymbol{\Omega}}-\boldsymbol{\Omega}^\star)\|_{\max}\leqslant 4\gamma\|\boldsymbol{\Omega}^\star\|_1. \tag{B.31}$$

For the last inequality in (B.31), we use $\|\boldsymbol{\Omega}^\star\|_1=\|\boldsymbol{\Omega}^\star\|_\infty$ because $\boldsymbol{\Omega}^\star$ is symmetric matrix. To proceed, let's derive the cone condition. Define $\Delta_j=\hat{\boldsymbol{\omega}}_j-\boldsymbol{\omega}_j^\star$ and $s_j=\text{supp}(\boldsymbol{\omega}_j^\star)$. From equation (B.29) we know

$$\|\boldsymbol{\omega}_j^\star\|_1\geqslant\|\hat{\boldsymbol{\omega}}_j\|_1=\|\Delta_j+\boldsymbol{\omega}_j^\star\|_1=\|(\Delta_j+\boldsymbol{\omega}_j^\star)_{s_j}\|_1+\|(\Delta_j)_{s_j^c}\|_1\Longrightarrow\|(\Delta_j)_{s_j}\|_1\geqslant\|(\Delta_j)_{s_j^c}\|_1. \tag{B.32}$$

Combine the cone condition in (B.32) with (B.31) and Assumption 20, we know

$$\|\Delta_j\|_1 \leqslant 2\|(\Delta_j)_{s_j}\|_1 \leqslant 2w\|\Delta_j\|_\infty = 2w\|\hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}^\star\|_{\max} \leqslant 8\|\boldsymbol{\Omega}^\star\|_1 w\gamma. \tag{B.33}$$

Therefore, if $\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^\star\|_{\max}\|\boldsymbol{\Omega}^\star\|_1 \leqslant \gamma$, then

$$\|\hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}^\star\|_2 \leqslant \sqrt{\|\hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}^\star\|_1\|\hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}^\star\|_\infty} = \|\hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}^\star\|_1 \leqslant 8\|\boldsymbol{\Omega}^\star\|_1 w\gamma. \tag{B.34}$$

Based on (B.27) and (B.34), we can choose $\gamma = 12\|\boldsymbol{\Omega}^\star\|_1\sqrt{M\log d_2/n}$, then with probability at least $1 - 2/d_2^2$, we have

$$\|\hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}^\star\|_2 \leqslant 96\|\boldsymbol{\Omega}^\star\|_1^2 w\sqrt{M\log d_2/n}. \tag{B.35}$$

Further, from (B.31), we know with probability at least $1 - 2/d_2^2$

$$\|\hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}^\star\|_{\max} \leqslant 4\|\boldsymbol{\Omega}^\star\|_1\gamma \leqslant 48\|\boldsymbol{\Omega}^\star\|_1^2\sqrt{M\log d_2/n}. \tag{B.36}$$

This concludes the proof.

Appendix C. Proofs of Theorems and Corollaries

C.1. Proof of Theorem 3

Let's fix $k \in [d_2]$ first. Based on the definition of $\hat{\boldsymbol{\beta}}_k$ in (7), we have following basic inequality

$$\hat{L}_k(\hat{\boldsymbol{\beta}}_k) + \lambda_k\|\hat{\boldsymbol{\beta}}_k\|_1 \leqslant \hat{L}_k(\tilde{\boldsymbol{\beta}}_k) + \lambda_k\|\tilde{\boldsymbol{\beta}}_k\|_1. \tag{C.1}$$

We define $\boldsymbol{\theta}_k = \hat{\boldsymbol{\beta}}_k - \tilde{\boldsymbol{\beta}}_k$ and have

$$\begin{aligned}\hat{L}_k(\hat{\boldsymbol{\beta}}_k) - \hat{L}_k(\tilde{\boldsymbol{\beta}}_k) &= \|\boldsymbol{\theta}_k\|_2^2 + 2\langle\tilde{\boldsymbol{\beta}}_k, \boldsymbol{\theta}_k\rangle - \frac{2}{n}\sum_{i=1}^n y_i Z_{ik}\langle\boldsymbol{X}_i, \boldsymbol{\theta}_k\rangle \\ &= \|\boldsymbol{\theta}_k\|_2^2 + \langle\nabla\hat{L}_k(\tilde{\boldsymbol{\beta}}_k), \boldsymbol{\theta}_k\rangle. \end{aligned}\tag{C.2}$$

Given a vector $\boldsymbol{v} \in \mathbb{R}^d$ and an index set $\mathcal{I} \subset [d]$, we define $\boldsymbol{v}_\mathcal{I} \in \mathbb{R}^d$ to be $\boldsymbol{v}$ restricted on $\mathcal{I}$ as $[\boldsymbol{v}_\mathcal{I}]_i = \boldsymbol{v}_i$ if $i \in \mathcal{I}$ and 0 otherwise. Suppose S_k is the support of $\tilde{\boldsymbol{\beta}}_k$, which is the same as $\boldsymbol{\beta}_k^\star$, combine (C.1) and (C.2) together and we get

$$\begin{aligned}\|\boldsymbol{\theta}_k\|_2^2 &\leqslant -\langle\nabla\hat{L}_k(\tilde{\boldsymbol{\beta}}_k), \boldsymbol{\theta}_k\rangle + \lambda_k\|\tilde{\boldsymbol{\beta}}_k\|_1 - \lambda_k\|\hat{\boldsymbol{\beta}}_k\|_1 \\ &= -\langle\nabla\hat{L}_k(\tilde{\boldsymbol{\beta}}_k), \boldsymbol{\theta}_k\rangle + \lambda_k\|(\tilde{\boldsymbol{\beta}}_k)_{S_k}\|_1 - \lambda_k\|(\hat{\boldsymbol{\beta}}_k)_{S_k}\|_1 - \lambda_k\|(\hat{\boldsymbol{\beta}}_k)_{S_k^C}\|_1 \\ &\leqslant -\langle\nabla\hat{L}_k(\tilde{\boldsymbol{\beta}}_k), \boldsymbol{\theta}_k\rangle + \lambda_k\|(\boldsymbol{\theta}_k)_{S_k}\|_1 - \lambda_k\|(\boldsymbol{\theta}_k)_{S_k^C}\|_1 \\ &\leqslant \|\nabla\hat{L}_k(\tilde{\boldsymbol{\beta}}_k)\|_\infty\|\boldsymbol{\theta}_k\|_1 + \lambda_k\|(\boldsymbol{\theta}_k)_{S_k}\|_1 - \lambda_k\|(\boldsymbol{\theta}_k)_{S_k^C}\|_1, \end{aligned}\tag{C.3}$$

where the third inequality is from triangle inequality and the last is based on Hölder's inequality. If we set $\lambda_k = 4\Upsilon_\gamma\sqrt{\log n/n}$, based on Lemma 4, we have

$$\|\nabla\hat{L}(\tilde{\boldsymbol{\beta}}_k)\|_\infty \leqslant \frac{\lambda_k}{2} \tag{C.4}$$

with probability at least $1 - d_1/n^2$. Combine (C.3) and (C.4), we know with probability at least $1 - d_1/n^2$,

$$\|\boldsymbol{\theta}_k\|_2^2 \leqslant \frac{3\lambda_k}{2}\|(\boldsymbol{\theta}_k)_{S_k}\|_1 - \frac{\lambda_k}{2}\|(\boldsymbol{\theta}_k)_{S_k^C}\|_1. \tag{C.5}$$

From (C.5) we further get the following cone condition:

$$\|(\boldsymbol{\theta}_k)_{S_k^C}\|_1 \leqslant 3\|(\boldsymbol{\theta}_k)_{S_k}\|_1. \tag{C.6}$$

Also from equation (C.5) and sparsity condition we know

$$\|\boldsymbol{\theta}_k\|_2^2 \leqslant \frac{3}{2}\lambda_k\|(\boldsymbol{\theta}_k)_{S_k}\|_1 \leqslant \frac{3}{2}\lambda_k\sqrt{s}\|(\boldsymbol{\theta}_k)_{S_k}\|_2 \leqslant \frac{3}{2}\lambda_k\sqrt{s}\|\boldsymbol{\theta}_k\|_2.$$

So we have with probability at least $1 - d_1/n^2$,

$$\|\boldsymbol{\theta}_k\|_2 \leqslant \frac{3}{2}\sqrt{s}\lambda_k.$$

Further by cone condition in (C.6) we can get l_1-norm convergence rate as

$$\|\boldsymbol{\theta}_k\|_1 = \|(\boldsymbol{\theta}_k)_{S_k}\|_1 + \|(\boldsymbol{\theta}_k)_{S_k^C}\|_1 \leqslant 4\|(\boldsymbol{\theta}_k)_{S_k}\|_1 \leqslant 4\sqrt{s}\|(\boldsymbol{\theta}_k)_{S_k}\|_2 \leqslant 6s\lambda_k.$$

By taking the union bound, it's easy to have

$$P(\|\boldsymbol{\theta}_k\|_2 \leqslant \frac{3}{2}\sqrt{s}\lambda_k \text{ and } \|\boldsymbol{\theta}_k\|_2 \leqslant 6s\lambda_k, \forall k) \geqslant 1 - \frac{d_2 d_1}{n^2}.$$

C.2. Proof of Corollary 5

We still fix $k \in [d_2]$ first and then take union bound. Denote $\mu = \min_{j\in[d_2]}|\mu_j|$, from Theorem 3, we know there exists $N(s,\mu)$ such that whenever $n \geqslant N$, we have

$$\|\hat{\boldsymbol{\beta}}_k\|_2 \geqslant \mu - \|\hat{\boldsymbol{\beta}}_k - \tilde{\boldsymbol{\beta}}_k\|_2 \geqslant \mu - 6\Upsilon\sqrt{s\log n/n} \geqslant \mu/2. \tag{C.7}$$

with probability at least $1 - d_1/n^2$. For either l_2-norm or l_1-norm, use (C.7) and we can get

$$\begin{aligned}
\|\frac{\hat{\boldsymbol{\beta}}_k}{\|\hat{\boldsymbol{\beta}}_k\|_2} - \frac{\tilde{\boldsymbol{\beta}}_k}{|\mu_k|}\| =& \frac{\|\hat{\boldsymbol{\beta}}_k - \|\hat{\boldsymbol{\beta}}_k\|_2/|\mu_k|\cdot\tilde{\boldsymbol{\beta}}_k\|}{\|\hat{\boldsymbol{\beta}}_k\|_2} \leqslant \frac{\|\hat{\boldsymbol{\beta}}_k - \tilde{\boldsymbol{\beta}}_k\| + \big||\mu_k| - \|\hat{\boldsymbol{\beta}}_k\|_2\big|\|\boldsymbol{\beta}_k^\star\|}{\|\hat{\boldsymbol{\beta}}_k\|_2} \\
\leqslant& \frac{2}{\mu}\|\hat{\boldsymbol{\beta}}_k - \tilde{\boldsymbol{\beta}}_k\| + \frac{2}{\mu}\|\boldsymbol{\beta}_k^\star\|\cdot\big|\|\tilde{\boldsymbol{\beta}}_k\|_2 - \|\hat{\boldsymbol{\beta}}_k\|_2\big| \\
\leqslant& \frac{2}{\mu}\|\hat{\boldsymbol{\beta}}_k - \tilde{\boldsymbol{\beta}}_k\| + \frac{2}{\mu}\|\boldsymbol{\beta}_k^\star\|\cdot\|\hat{\boldsymbol{\beta}}_k - \tilde{\boldsymbol{\beta}}_k\|_2.
\end{aligned} \tag{C.8}$$

Thus, combine (C.8) with Theorem 3, we know with probability $1 - d_1/n^2$

$$\begin{aligned}
\|\frac{\hat{\boldsymbol{\beta}}_k}{\|\hat{\boldsymbol{\beta}}_k\|_2} - \frac{\tilde{\boldsymbol{\beta}}_k}{|\mu_k|}\|_2 \leqslant& \frac{4}{\mu}\|\hat{\boldsymbol{\beta}}_k - \tilde{\boldsymbol{\beta}}_k\|_2 \lesssim \sqrt{s\log n/n}/\mu, \\
\|\frac{\hat{\boldsymbol{\beta}}_k}{\|\hat{\boldsymbol{\beta}}_k\|_2} - \frac{\tilde{\boldsymbol{\beta}}_k}{|\mu_k|}\|_1 \leqslant& \frac{2}{\mu}\|\hat{\boldsymbol{\beta}}_k - \tilde{\boldsymbol{\beta}}_k\|_1 + \frac{2\sqrt{s}}{\mu}\|\hat{\boldsymbol{\beta}}_k - \tilde{\boldsymbol{\beta}}_k\|_2 \lesssim s\sqrt{\log n/n}/\mu.
\end{aligned}$$

Note that under identifiability condition (2) we have $\boldsymbol{\beta}_k^\star = \text{sign}(\tilde{\beta}_{k1}) \cdot \frac{\tilde{\boldsymbol{\beta}}_k}{|\mu_k|}$, hence there exists $M(s, N, \min_{j\in[d_2]} \beta_{j1}^\star)$, such that $n \geqslant M$, $\text{sign}(\hat{\beta}_{k1}) = \text{sign}(\tilde{\beta}_{k1})$. So we can get for either l_2-norm or l_1-norm

$$\|\frac{\hat{\boldsymbol{\beta}}_k}{\|\hat{\boldsymbol{\beta}}_k\|_2} - \frac{\tilde{\boldsymbol{\beta}}_k}{|\mu_k|}\| = \|\text{sign}(\hat{\beta}_{k1})\frac{\hat{\boldsymbol{\beta}}_k}{\|\hat{\boldsymbol{\beta}}_k\|_2} - \text{sign}(\hat{\beta}_{k1})\frac{\tilde{\boldsymbol{\beta}}_k}{|\mu_k|}\| = \|\text{sign}(\hat{\beta}_{k1})\frac{\hat{\boldsymbol{\beta}}_k}{\|\hat{\boldsymbol{\beta}}_k\|_2} - \boldsymbol{\beta}_k^\star\|.$$

By taking the union bound, we can get the conclusion. Particularly, in the worst case, we have

$$\|\hat{\boldsymbol{B}} - \boldsymbol{B}^\star\|_F^2 = \sum_{j=1}^{d_2} \|\text{sign}(\hat{\beta}_{k1})\frac{\hat{\boldsymbol{\beta}}_k}{\|\hat{\boldsymbol{\beta}}_k\|_2} - \boldsymbol{\beta}_k^\star\|_2^2 \lesssim d_2 s \log n/n\mu^2.$$

So, we know $\|\hat{\boldsymbol{B}} - \boldsymbol{B}^\star\|_F \lesssim \frac{1}{\mu}\sqrt{\frac{d_2 s \log n}{n}}$. This concludes the proof.

C.3. Proof of Theorem 7

We still fix $k \in [d_2]$ first. Start from the definition of $\hat{\boldsymbol{\beta}}_k$ in (8) and basic inequality, then follow the same steps as in (C.1), (C.2), and (C.3), we finally get

$$\|\boldsymbol{\theta}_k\|_2^2 \leqslant \|\nabla \bar{L}_k(\tilde{\boldsymbol{\beta}}_k)\|_\infty \|\boldsymbol{\theta}_k\|_1 + \lambda_k \|(\boldsymbol{\theta}_k)_{\mathcal{S}_k}\|_1 - \lambda_k \|(\boldsymbol{\theta}_k)_{\mathcal{S}_k^C}\|_1. \tag{C.9}$$

Based on Lemma 8, we can let $\lambda_k = 76\sqrt{M \log d_1 d_2/n}$ and have $\|\nabla \bar{L}_k(\tilde{\boldsymbol{\beta}}_k)\|_\infty \leqslant \lambda_k/2$. Plug into (C.9) and follow the derivation in equation (C.5) and (C.6), we can finally get

$$\|\hat{\boldsymbol{\beta}}_k - \tilde{\boldsymbol{\beta}}_k\|_2 \leqslant \frac{3}{2}\sqrt{s}\lambda_k \quad \text{and} \quad \|\hat{\boldsymbol{\beta}}_k - \tilde{\boldsymbol{\beta}}_k\|_1 \leqslant 6s\lambda_k.$$

Note that above error bound holds uniformly over $k \in [d_2]$ and we finish the proof.

C.4. Proof of Theorem 9

To make notation consistent, let's denote the loss function without penalty defined in (13) by $\hat{L}(\boldsymbol{B})$ and define $\mathcal{I}_4 = \hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}^\star$, then we have

$$\begin{aligned}
\nabla \hat{L}(\tilde{\boldsymbol{B}}) =& 2\tilde{\boldsymbol{B}} - \frac{2}{n\kappa_1}\sum_{i=1}^n \Phi(\kappa_1 y_i \cdot S(X_i) Z_i^T)\hat{\boldsymbol{\Omega}} \\
\overset{(9)}{=}& 2\left(\mathbb{E}[y \cdot S(\boldsymbol{x})\boldsymbol{z}^T]\boldsymbol{\Omega}^\star - \frac{1}{n\kappa_1}\sum_{i=1}^n \Phi(\kappa_1 y_i \cdot S(X_i) Z_i^T)\hat{\boldsymbol{\Omega}}\right).
\end{aligned} \tag{C.10}$$

Based on equation (C.10), we use triangle inequality and get

$$\begin{aligned}
&\|\nabla \hat{L}(\tilde{\boldsymbol{B}})\|_2 \qquad\qquad (\text{C.11}) \\
&\leqslant 2\underbrace{\|\mathbb{E}[y \cdot S(\boldsymbol{x})\boldsymbol{z}^T] - \frac{1}{n\kappa_1}\sum_{i=1}^n \Phi(\kappa_1 y_i \cdot S(X_i) Z_i^T)\|_2}_{\mathcal{I}_3}\|\hat{\boldsymbol{\Omega}}\|_2 + 2\|\mathbb{E}[y \cdot S(\boldsymbol{x})\boldsymbol{z}^T]\|_2 \|\underbrace{\hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}^\star}_{\mathcal{I}_4}\|_2 \\
&\leqslant 2\|\mathcal{I}_3\|_2\|\mathcal{I}_4\|_2 + \underbrace{2\|\boldsymbol{\Omega}^\star\|_2\|\mathcal{I}_3\|_2 + 2\|\mathbb{E}[y \cdot S(\boldsymbol{x})\boldsymbol{z}^T]\|_2\|\mathcal{I}_4\|_2}_{\text{dominant term}}. \qquad (\text{C.12})
\end{aligned}$$

Note that

$$\|\mathbb{E}[y \cdot S(\boldsymbol{x})\boldsymbol{z}^T]\|_2 = \|\tilde{\boldsymbol{B}}\boldsymbol{\Sigma}^\star\|_2 \leqslant \max_{j\in[d_2]} |\mu_j| \cdot \|\boldsymbol{B}^\star\|_2 \|\boldsymbol{\Sigma}^\star\|_2. \tag{C.13}$$

So combine (B.18), (C.11), (C.13) together and drop off smaller order term, we can get

$$\begin{aligned} P\bigg(\|\nabla \hat{L}(\tilde{\boldsymbol{B}})\|_2 \leqslant & 8KM^{3/4}\sqrt{\frac{(d_1+d_2)\log(d_1+d_2)}{n}} + 2K \max_{j\in[d_2]} |\mu_j| \cdot \|\boldsymbol{B}^\star\|_2 \mathcal{H}(n,d_2)\bigg) \\ & \geqslant 1 - \frac{2}{(d_1+d_2)^2} - \mathcal{P}(n,d_2). \end{aligned} \tag{C.14}$$

So we know under the setup of λ as in theorem, we have $\|\nabla \hat{L}(\tilde{\boldsymbol{B}})\|_2 \leqslant \lambda/2$ with probability at least $1 - 2/(d_1+d_2)^2 - \mathcal{P}(n,d_2)$. On the other side, start from the definition of $\hat{\boldsymbol{B}}$ in (13), we have following basic inequality

$$\hat{L}(\hat{\boldsymbol{B}}) + \lambda\|\hat{\boldsymbol{B}}\|_* \leqslant \hat{L}(\tilde{\boldsymbol{B}}) + \lambda\|\tilde{\boldsymbol{B}}\|_*. \tag{C.15}$$

Define $\boldsymbol{\Theta} = \hat{\boldsymbol{B}} - \tilde{\boldsymbol{B}}$, we have

$$\begin{aligned} \hat{L}(\hat{\boldsymbol{B}}) - \hat{L}(\tilde{\boldsymbol{B}}) &= \|\hat{\boldsymbol{B}}\|_F^2 - \|\tilde{\boldsymbol{B}}\|_F^2 - \frac{2}{n\kappa_1}\sum_{i=1}^n \langle \Phi(\kappa_1 y_i \cdot S(\boldsymbol{X}_i)\boldsymbol{Z}_i^T)\hat{\boldsymbol{\Omega}}, \hat{\boldsymbol{B}} - \tilde{\boldsymbol{B}}\rangle \\ &= \|\boldsymbol{\Theta}\|_F^2 + 2\langle \tilde{\boldsymbol{B}}, \boldsymbol{\Theta}\rangle - \frac{2}{n\kappa_1}\sum_{i=1}^n \langle \Phi(\kappa_1 y_i \cdot S(\boldsymbol{X}_i)\boldsymbol{Z}_i^T)\hat{\boldsymbol{\Omega}}, \boldsymbol{\Theta}\rangle \\ &= \langle \nabla \hat{L}(\tilde{\boldsymbol{B}}), \boldsymbol{\Theta}\rangle + \|\boldsymbol{\Theta}\|_F^2. \end{aligned} \tag{C.16}$$

Combine (C.15) and (C.16) together, we have

$$\begin{aligned} \|\boldsymbol{\Theta}\|_F^2 &= -\langle \nabla \hat{L}(\tilde{\boldsymbol{B}}), \boldsymbol{\Theta}\rangle + \hat{L}(\hat{\boldsymbol{B}}) - \hat{L}(\tilde{\boldsymbol{B}}) \\ &\leqslant -\langle \nabla \hat{L}(\tilde{\boldsymbol{B}}), \boldsymbol{\Theta}\rangle + \lambda\|\tilde{\boldsymbol{B}}\|_* - \lambda\|\hat{\boldsymbol{B}}\|_* \\ &\leqslant \|\nabla \hat{L}(\tilde{\boldsymbol{B}})\|_2 \|\boldsymbol{\Theta}\|_* + \lambda\|\tilde{\boldsymbol{B}}\|_* - \lambda\|\hat{\boldsymbol{B}}\|_*. \end{aligned} \tag{C.17}$$

Under Assumption 2 and 18, we know $r = \mathrm{rank}(\boldsymbol{B}^\star) = \mathrm{rank}(\tilde{\boldsymbol{B}})$. We let $\tilde{\boldsymbol{B}} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^T$ be its singular value decomposition where diagonal matrix $\boldsymbol{\Lambda} \in \mathbb{R}^{d_1\times d_2}$ can be expressed as $\begin{pmatrix} \boldsymbol{\Lambda}_{11} & 0 \\ 0 & 0 \end{pmatrix}$ for $\boldsymbol{\Lambda}_{11} \in \mathbb{R}^{r\times r}$. We define

$$\boldsymbol{T} = \boldsymbol{U}^T\boldsymbol{\Theta}\boldsymbol{V} = \boldsymbol{T}^{(1)} + \boldsymbol{T}^{(2)}$$

where $\boldsymbol{T}^{(1)} = \begin{pmatrix} 0 & 0 \\ 0 & \boldsymbol{T}_{22} \end{pmatrix}$ and $\boldsymbol{T}^{(2)} = \begin{pmatrix} \boldsymbol{T}_{11} & \boldsymbol{T}_{12} \\ \boldsymbol{T}_{21} & 0 \end{pmatrix}$ have the same corresponding block size as $\boldsymbol{\Lambda}$. Then we get

$$\begin{aligned} \|\hat{\boldsymbol{B}}\|_* =& \|\tilde{\boldsymbol{B}} + \boldsymbol{\Theta}\|_* = \|\boldsymbol{U}(\boldsymbol{\Lambda} + \boldsymbol{T})\boldsymbol{V}^T\|_* = \|\boldsymbol{\Lambda} + \boldsymbol{T}\|_* \\ \geqslant& \|\boldsymbol{\Lambda} + \boldsymbol{T}^{(1)}\|_* - \|\boldsymbol{T}^{(2)}\|_* = \|\tilde{\boldsymbol{B}}\|_* + \|\boldsymbol{T}^{(1)}\|_* - \|\boldsymbol{T}^{(2)}\|_*. \end{aligned} \tag{C.18}$$

The last equality is because of the block diagonal structure of $\mathbf{\Lambda}$ and $\boldsymbol{T}^{(1)}$ and $\|\tilde{\boldsymbol{B}}\|_* = \|\mathbf{\Lambda}\|_*$. Combine (C.18) and (C.17), we have

$$\|\mathbf{\Theta}\|_F^2 \leqslant \frac{3\lambda}{2}\|\boldsymbol{T}^{(2)}\|_* - \frac{\lambda}{2}\|\boldsymbol{T}^{(1)}\|_*. \tag{C.19}$$

According to (C.19), we have following cone condition

$$\|\boldsymbol{T}^{(1)}\|_* \leqslant 3\|\boldsymbol{T}^{(2)}\|_*. \tag{C.20}$$

Also form (C.19) and using Assumption 18, we get

$$\|\mathbf{\Theta}\|_F^2 \leqslant \frac{3\lambda}{2}\|\boldsymbol{T}^{(2)}\|_* \leqslant \frac{3\lambda}{2}\sqrt{\operatorname{rank}(\boldsymbol{T}^{(2)})}\|\boldsymbol{T}^{(2)}\|_F \leqslant 3\lambda\sqrt{r}\|\boldsymbol{T}^{(2)}\|_F \leqslant 3\lambda\sqrt{r}\|\mathbf{\Theta}\|_F.$$

Combining with (C.20), we know with probability at least $1 - 2/(d_1+d_2)^2 - \mathcal{P}(n, d_2)$,

$$\begin{aligned}\|\mathbf{\Theta}\|_F \leqslant& 3\sqrt{r}\lambda,\\ \|\mathbf{\Theta}\|_* \leqslant& \|\boldsymbol{T}^{(1)}\|_* + \|\boldsymbol{T}^{(2)}\|_* \leqslant 4\|\boldsymbol{T}^{(2)}\|_* \leqslant 24r\lambda.\end{aligned}$$

This concludes the theorem.

C.5. Proof of Theorem 13

Let's denote the loss function without penalty defined in (16) by $\hat{L}(\boldsymbol{B})$ and let $\mathcal{I}_4 = \hat{\mathbf{\Omega}} - \mathbf{\Omega}^\star$. We know

$$\begin{aligned}\nabla\hat{L}(\tilde{\boldsymbol{B}}) =& 2\tilde{\boldsymbol{B}} - \frac{2}{n}\sum_{i=1}^n \breve{y}_i \cdot \widetilde{S(\boldsymbol{X}_i)}\breve{\boldsymbol{Z}}_i^T\hat{\mathbf{\Omega}} \overset{(9)}{=} 2\mathbb{E}[y\cdot S(\boldsymbol{x})\boldsymbol{z}^T]\mathbf{\Omega}^\star - \frac{2}{n}\sum_{i=1}^n \breve{y}_i \cdot \widetilde{S(\boldsymbol{X}_i)}\breve{\boldsymbol{Z}}_i^T\hat{\mathbf{\Omega}}\\ =& 2\bigg(\mathbb{E}[y\cdot S(\boldsymbol{x})\boldsymbol{z}^T](\mathbf{\Omega}^\star - \hat{\mathbf{\Omega}}) + (\mathbb{E}[y\cdot S(\boldsymbol{x})\boldsymbol{z}^T] - \frac{2}{n}\sum_{i=1}^n \breve{y}_i \cdot \widetilde{S(\boldsymbol{X}_i)}\breve{\boldsymbol{Z}}_i^T)\hat{\mathbf{\Omega}}\bigg).\end{aligned} \tag{C.21}$$

From (C.21), we have

$$\|\nabla\hat{L}(\tilde{\boldsymbol{B}})\|_{\max} \leqslant 2\|\mathcal{I}_7\|_{\max}\|\mathcal{I}_4\|_1 + \underbrace{2\|\mathcal{I}_7\|_{\max}\|\mathbf{\Omega}^\star\|_1 + 2\|\mathcal{I}_4\|_{\max}\|\mathbb{E}[y\cdot S(\boldsymbol{x})\boldsymbol{z}^T]\|_\infty}_{\text{dominant term}}.$$

Note that

$$\|\mathbb{E}[y\cdot S(\boldsymbol{x})\boldsymbol{z}^T]\|_\infty = \|\tilde{\boldsymbol{B}}\mathbf{\Sigma}^\star\|_\infty \leqslant \max_{j\in[d_2]}|\mu_j|\cdot\|\boldsymbol{B}^\star\mathbf{\Sigma}^\star\|_\infty. \tag{C.22}$$

Combine (C.22) with (B.22) and drop off the intersection term (smaller order), we know

$$\begin{aligned}P\bigg(\|\nabla\hat{L}(\tilde{\boldsymbol{B}})\|_{\max} > 38\|\mathbf{\Omega}^\star\|_1\sqrt{\frac{M\log d_1 d_2}{n}} + 2\max_{j\in[d_2]}|\mu_j|\cdot\|\boldsymbol{B}^\star\mathbf{\Sigma}^\star\|_\infty\tilde{\mathcal{H}}(n,d_2)\bigg)\\ \leqslant \frac{2}{d_1^2 d_2^2} + \tilde{\mathcal{P}}(n,d_2).\end{aligned}$$

So under the setup in theorem we have $\|\nabla \hat{L}(\tilde{\boldsymbol{B}})\|_{\max} \leqslant \lambda/2$. On the other side, based on definition of $\hat{\boldsymbol{B}}$ in (16), we have following basic inequality

$$\hat{L}(\hat{\boldsymbol{B}}) + \lambda\|\hat{\boldsymbol{B}}\|_{1,1} \leqslant \hat{L}(\tilde{\boldsymbol{B}}) + \lambda\|\tilde{\boldsymbol{B}}\|_{1,1}. \tag{C.23}$$

Define $\boldsymbol{\Theta} = \hat{\boldsymbol{B}} - \tilde{\boldsymbol{B}}$, same with (C.16) we have

$$\hat{L}(\hat{\boldsymbol{B}}) - \hat{L}(\tilde{\boldsymbol{B}}) = \langle \nabla \hat{L}(\tilde{\boldsymbol{B}}), \boldsymbol{\Theta}\rangle + \|\boldsymbol{\Theta}\|_F^2. \tag{C.24}$$

Combine (C.23) and (C.24), and define $S = \mathrm{supp}(\tilde{\boldsymbol{B}})$, we have

$$\begin{aligned}
\|\boldsymbol{\Theta}\|_F^2 \leqslant& -\langle \nabla \hat{L}(\tilde{\boldsymbol{B}}), \boldsymbol{\Theta}\rangle + \lambda\|\tilde{\boldsymbol{B}}\|_{1,1} - \lambda\|\hat{\boldsymbol{B}}\|_{1,1} \\
\leqslant& \|\nabla \hat{L}(\tilde{\boldsymbol{B}})\|_{\max}\|\boldsymbol{\Theta}\|_{1,1} + \lambda\|\tilde{\boldsymbol{B}}_S\|_{1,1} - \lambda\|\hat{\boldsymbol{B}}_S\|_{1,1} - \lambda\|\hat{\boldsymbol{B}}_{S^C}\|_{1,1} \\
\leqslant& \|\nabla \hat{L}(\tilde{\boldsymbol{B}})\|_{\max}\|\boldsymbol{\Theta}\|_{1,1} + \lambda\|\boldsymbol{\Theta}_S\|_{1,1} - \lambda\|\boldsymbol{\Theta}_{S^C}\|_{1,1}.
\end{aligned} \tag{C.25}$$

So, based on (C.25), we know, with probability at least $1 - 2/d_1d_2 - \tilde{\mathcal{P}}(n, d_2)$,

$$\|\boldsymbol{\Theta}\|_F^2 \leqslant \frac{3\lambda}{2}\|\boldsymbol{\Theta}_S\|_{1,1} \leqslant \frac{3\lambda}{2}\sqrt{sd_2}\|\boldsymbol{\Theta}_S\|_F \Longrightarrow \|\boldsymbol{\Theta}\|_F \leqslant 2\sqrt{sd_2}\lambda. \tag{C.26}$$

Similarly, we have

$$\|\boldsymbol{\Theta}\|_{1,1} \leqslant 4\|\boldsymbol{\Theta}_S\|_{1,1} \leqslant 8sd_2\lambda.$$

This concludes the proof.

Appendix D. Additional simulation results and plots

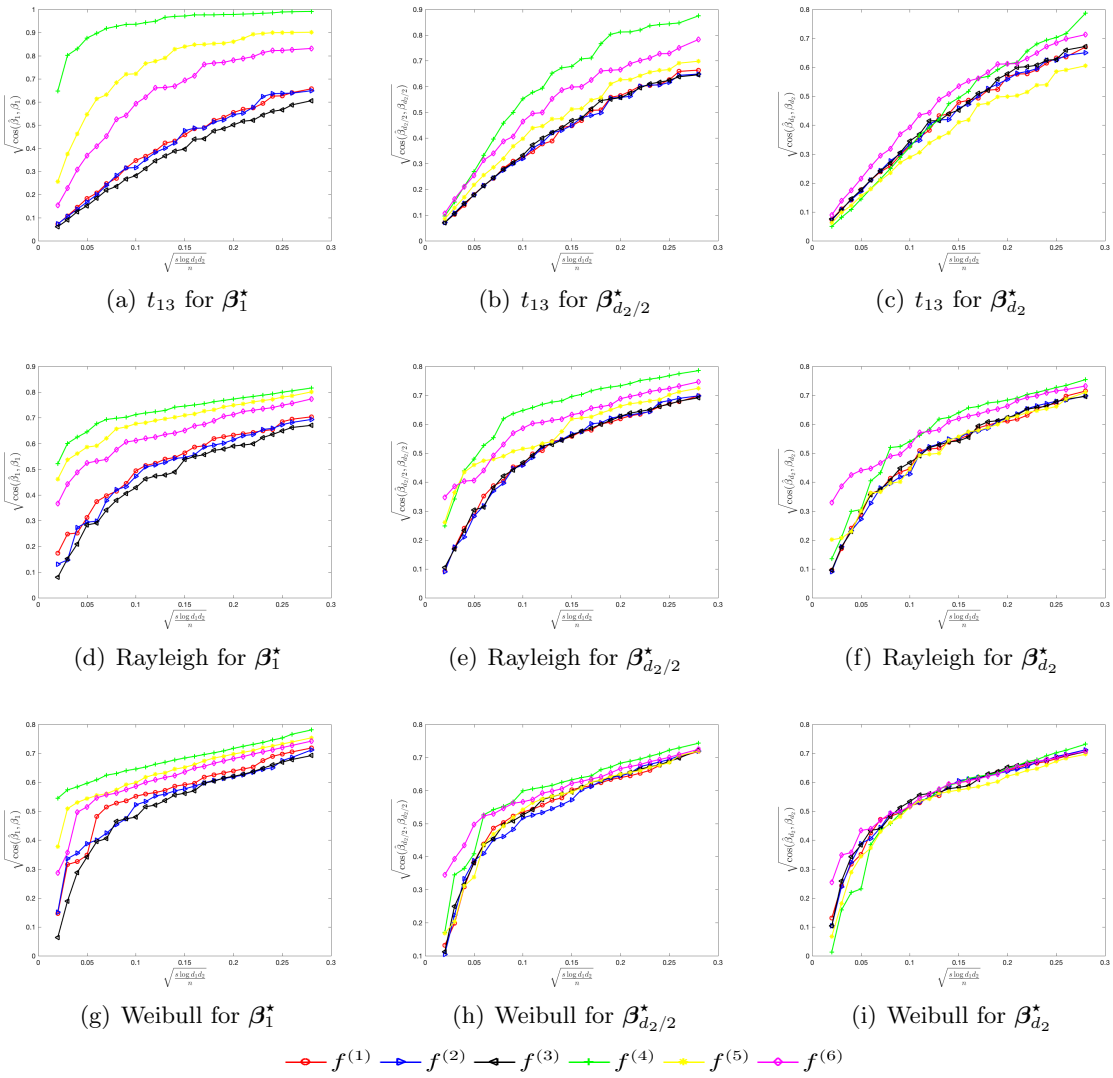
We present additional simulations results for the setting discussed in Section 7.1 and 7.4. Figure 9 illustrates the error trend for estimating $\boldsymbol{\beta}_1^\star$, $\boldsymbol{\beta}_{d_2/2}^\star$ and $\boldsymbol{\beta}_{d_2}^\star$ under different link functions and $\boldsymbol{x}$ following one of the following distributions: t_{13}, Rayleigh or Weibull. Figure 10–12 plot the error trend for all d_2 parameters. From the plots we observe that the linear trend will be more and more obvious when k varies from 1 to d_2, especially for quadratic link functions. Note also that the number of samples needed to get comparable results decrease dramatically as k increases. Figure 13 shows 95% confidence interval for the signal trajectory of the real dataset we analyzed in Section 7.4, which is obtained by doing parametric bootstrapping.

References

D. Babichev and F. Bach. Slice inverse regression with score functions. *Electron. J. Stat.*, 12(1):1507–1543, 2018.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities*. Oxford University Press, Oxford, 2013. A nonasymptotic theory of independence, With a foreword by Michel Ledoux.

P. Bühlmann and S. A. van de Geer. *Statistics for high-dimensional data*. Springer Series in Statistics. Springer, Heidelberg, 2011. Methods, theory and applications.



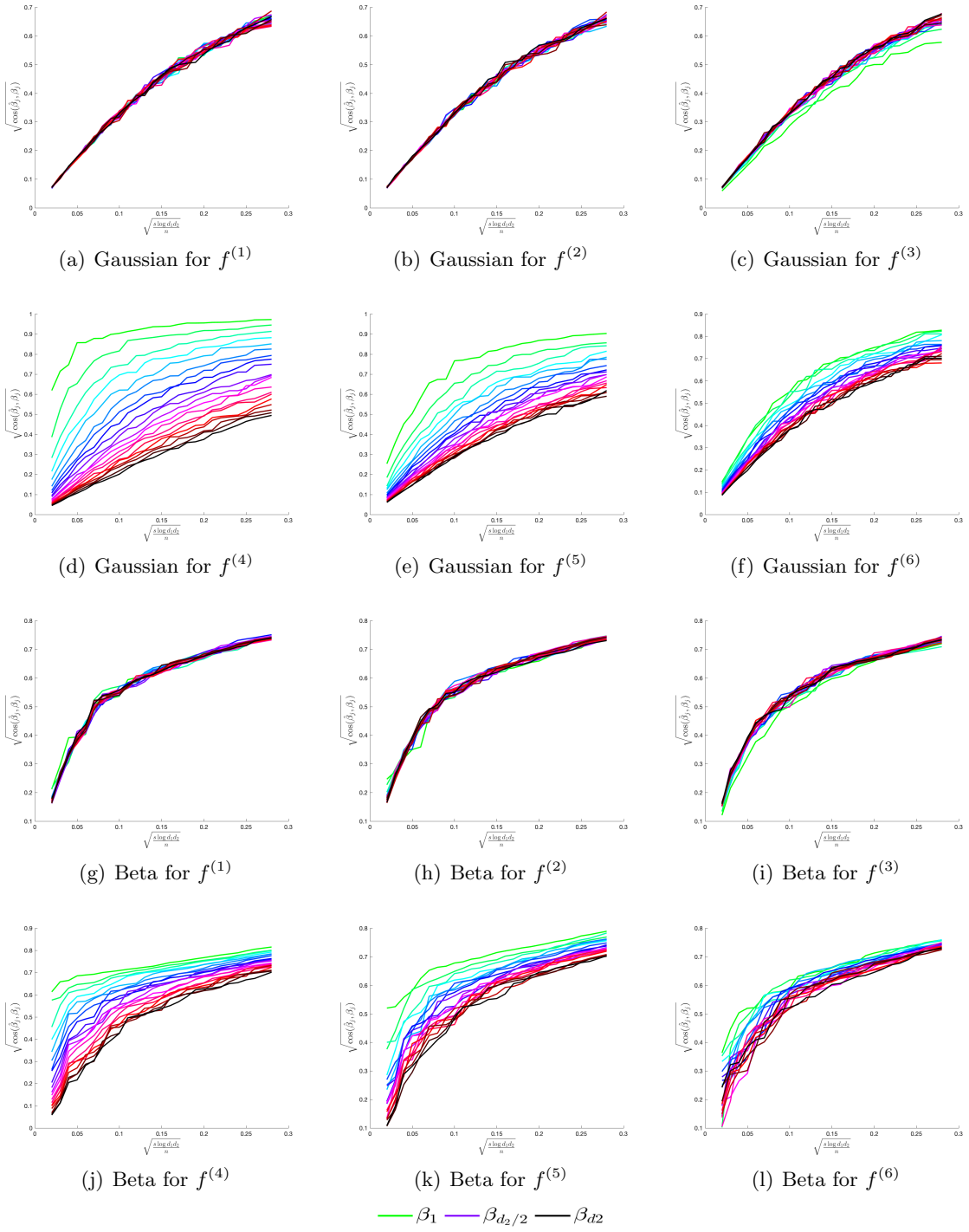
(a) t_{13} for $\boldsymbol{\beta}_1^\star$ (b) t_{13} for $\boldsymbol{\beta}_{d_2/2}^\star$ (c) t_{13} for $\boldsymbol{\beta}_{d_2}^\star$

(d) Rayleigh for $\boldsymbol{\beta}_1^\star$ (e) Rayleigh for $\boldsymbol{\beta}_{d_2/2}^\star$ (f) Rayleigh for $\boldsymbol{\beta}_{d_2}^\star$

(g) Weibull for $\boldsymbol{\beta}_1^\star$ (h) Weibull for $\boldsymbol{\beta}_{d_2/2}^\star$ (i) Weibull for $\boldsymbol{\beta}_{d_2}^\star$

$f^{(1)}$ $f^{(2)}$ $f^{(3)}$ $f^{(4)}$ $f^{(5)}$ $f^{(6)}$

Figure 9: Sparse vector estimation plot (II). This figure shows cosine distance trend for error of estimating single sparse parameter in model (1). Six lines indicates six different types of link functions. Each row represents one type of distribution for design. We only plot the error trend for the first, the middle, the last parameter. All above simulation results are consistent with Theorem 7.



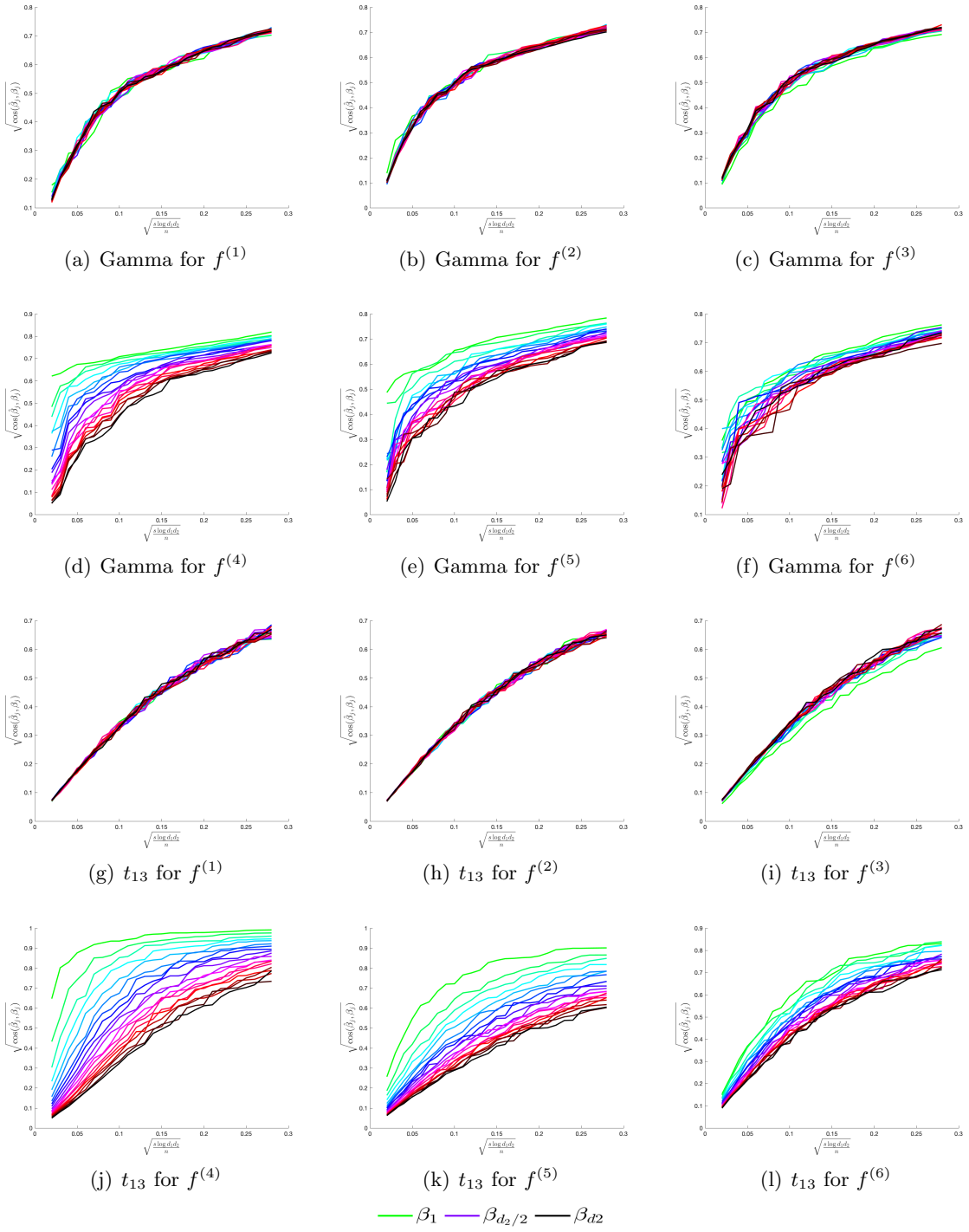
(a) Gaussian for $f^{(1)}$ (b) Gaussian for $f^{(2)}$ (c) Gaussian for $f^{(3)}$

(d) Gaussian for $f^{(4)}$ (e) Gaussian for $f^{(5)}$ (f) Gaussian for $f^{(6)}$

(g) Beta for $f^{(1)}$ (h) Beta for $f^{(2)}$ (i) Beta for $f^{(3)}$

(j) Beta for $f^{(4)}$ (k) Beta for $f^{(5)}$ (l) Beta for $f^{(6)}$

Figure 10: Sparse vector estimation for Gaussian and Beta design. This figure shows cosine distance trend for error of estimating single sparse parameter in model (1). The color varies from green to black when k varies from 1 to d_2. The first and third row correspond to the linear link function, while the second and fourth row correspond to the quadratic link function.



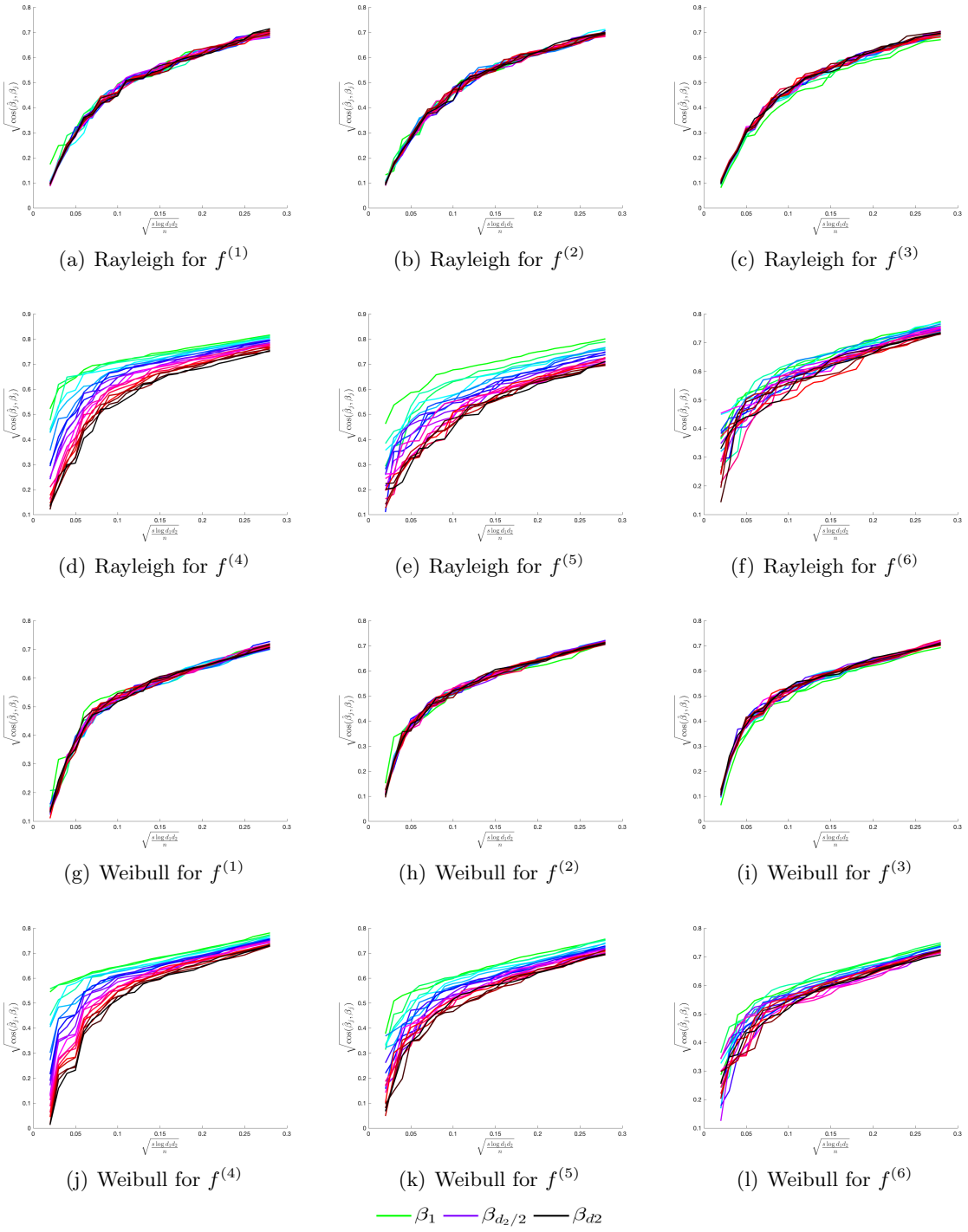
(a) Gamma for $f^{(1)}$ (b) Gamma for $f^{(2)}$ (c) Gamma for $f^{(3)}$

(d) Gamma for $f^{(4)}$ (e) Gamma for $f^{(5)}$ (f) Gamma for $f^{(6)}$

(g) t_{13} for $f^{(1)}$ (h) t_{13} for $f^{(2)}$ (i) t_{13} for $f^{(3)}$

(j) t_{13} for $f^{(4)}$ (k) t_{13} for $f^{(5)}$ (l) t_{13} for $f^{(6)}$

Figure 11: Sparse vector estimation for Gamma and t design. This figure shows cosine distance trend for error of estimating single sparse parameter in model (1). The color varies from green to black when k varies from 1 to d_2. The first and third row correspond to the linear link function, while the second and fourth row correspond to the quadratic link function.



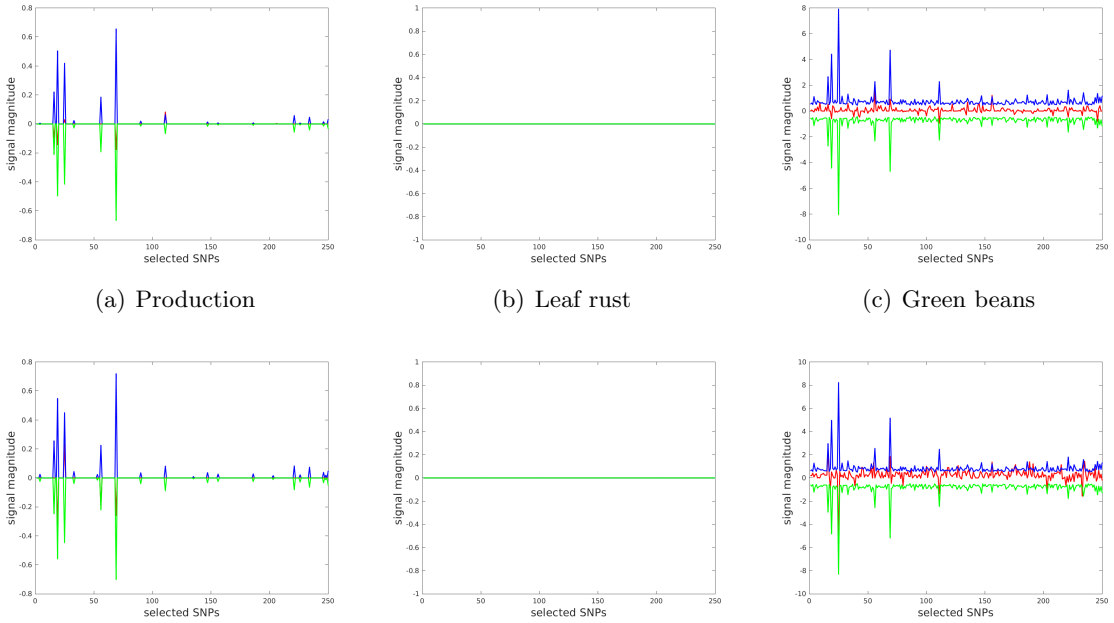
(a) Rayleigh for $f^{(1)}$ (b) Rayleigh for $f^{(2)}$ (c) Rayleigh for $f^{(3)}$

(d) Rayleigh for $f^{(4)}$ (e) Rayleigh for $f^{(5)}$ (f) Rayleigh for $f^{(6)}$

(g) Weibull for $f^{(1)}$ (h) Weibull for $f^{(2)}$ (i) Weibull for $f^{(3)}$

(j) Weibull for $f^{(4)}$ (k) Weibull for $f^{(5)}$ (l) Weibull for $f^{(6)}$

Figure 12: Sparse vector estimation for Rayleigh and Weibull design. This figure shows cosine distance trend for error of estimating single sparse parameter in model (1). The color varies from green to black when k varies from 1 to d_2. The first and third row correspond to the linear link function, while the second and fourth row correspond to the quadratic link function.



(a) Production (b) Leaf rust (c) Green beans

(d) Production (e) Leaf rust (f) Green beans

Figure 13: Signal trajectories for 95% confidence interval (parametric bootstrapping). The (l, i) plot shows the confidence interval trajectory of $(\boldsymbol{\beta}^{l,i})^\star$ where $l = 1, 2$ indexes the confounder and $i = 1, 2, 3$ indexes the response. The blue line indicates the upper bound, the red line indicates the estimator, and the green line indicates the lower bound.

T. T. Cai, W. Liu, and X. Luo. A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *J. Am. Stat. Assoc.*, 106(494):594–607, 2011.

T. T. Cai, W. Liu, and H. H. Zhou. Estimating sparse precision matrix: optimal rates of convergence and adaptive estimation. *Ann. Statist.*, 44(2):455–488, 2016.

R. J. Carroll, J. Fan, I. Gijbels, and M. P. Wand. Generalized partially linear single-index models. *J. Amer. Statist. Assoc.*, 92(438):477–489, 1997.

O. Catoni. Challenging the empirical mean and empirical variance: a deviation study. *Ann. Inst. Henri Poincaré Probab. Stat.*, 48(4):1148–1185, 2012.

H. Chen. Estimation of a projection-pursuit type regression model. *Ann. Statist.*, 19(1): 142–157, 1991.

L. H. Y. Chen, L. Goldstein, and Q.-M. Shao. *Normal approximation by Stein's method*. Probability and its Applications (New York). Springer, Heidelberg, 2011.

Y. Chen and R. J. Samworth. Generalized additive and index models with shape constraints. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 78(4):729–754, 2016.

K. Chwialkowski, H. Strathmann, and A. Gretton. A kernel test of goodness of fit. In M. F. Balcan and K. Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2606–2615, New York, New York, USA, 2016. PMLR.

W. S. Cleveland, E. Grosse, and W. M. Shyu. Local regression models. In J. M. Chambers and T. J. Hastie, editors, *Statistical Models in S*, pages 309–376, 1991.

J. Fan and W. Zhang. Statistical methods with varying coefficient models. *Statistics and its Interface*, 1(1):179–195, 2008.

J. Fan, Q. Yao, and Z. Cai. Adaptive varying-coefficient linear models. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 65(1):57–80, 2003.

L. F. V. Ferrão, R. G. Ferrão, M. A. G. Ferrão, A. Fonseca, P. Carbonetto, M. Stephens, and A. A. F. Garcia. Accurate genomic prediction of coffea canephora in multiple environments using whole-genome statistical models. *Heredity*, pages 1–15, 2018.

L. Goldstein, S. Minsker, and X. Wei. Structured signal recovery from non-linear and heavy-tailed measurements. *IEEE Trans. Inform. Theory*, 64(8):5513–5530, 2018.

M. C. Grant and S. P. Boyd. cvx: Matlab software for disciplined convex programming, version 2.0 beta. `http://cvxr.com/cvx`, 2012.

M. Grant and S. Boyd. Graph implementations for nonsmooth convex programs. In V. Blondel, S. Boyd, and H. Kimura, editors, *Recent Advances in Learning and Control*, Lecture Notes in Control and Information Sciences, pages 95–110. Springer-Verlag Limited, 2008. `http://stanford.edu/~boyd/graph_dcp.html`.

C. Guo, H. Yang, and J. Lv. Generalized varying index coefficient models. *J. Comput. Appl. Math.*, 300:1–17, 2016.

T. J. Hastie and R. J. Tibshirani. Varying-coefficient models. *J. R. Stat. Soc. B*, 55(4): 757–796, 1993.

Z. Huang and R. Zhang. Profile empirical-likelihood inferences for the single-index-coefficient regression model. *Stat. Comput.*, 23(4):455–465, 2013.

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

K.-C. Li. Sliced inverse regression for dimension reduction. *J. Amer. Statist. Assoc.*, 86 (414):316–342, 1991. With discussion and a rejoinder by the author.

Q. Lin, X. Li, D. Huang, and J. S. Liu. On the optimality of sliced inverse regression in high dimensions. *ArXiv e-prints: 1701.06009*, 2017.

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

H. Liu, Y. Feng, Y. Mao, D. Zhou, J. Peng, and Q. Liu. Action-dependent control variates for policy optimization via stein identity. In *International Conference on Learning Representations*, 2018.

Q. Liu and D. Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 2378–2386. Curran Associates, Inc., 2016.

Q. Liu, J. Lee, and M. Jordan. A kernelized stein discrepancy for goodness-of-fit tests. In M. F. Balcan and K. Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 276–284, New York, New York, USA, 2016. PMLR.

S. Ma and P. X.-K. Song. Varying index coefficient models. *J. Amer. Statist. Assoc.*, 110 (509):341–356, 2015.

S. Minsker. Sub-Gaussian estimators of the mean of a random matrix with heavy-tailed entries. *Ann. Statist.*, 46(6A):2871–2903, 2018.

S. Na and M. Kolar. High-dimensional index volatility models via stein's identity. *arXiv preprint arXiv:1811.10790*, 2018.

M. Neykov, J. S. Liu, and T. Cai. L1-regularized least squares for support recovery of high dimensional single index models with gaussian designs. *Journal of Machine Learning Research*, 17(87):1–37, 2016.

Y. Plan, R. Vershynin, and E. Yudovina. High-dimensional estimation with geometric constraints. *Inf. Inference*, 6(1):1–40, 2017.

Y. Plan and R. Vershynin. The generalized Lasso with non-linear observations. *IEEE Trans. Inform. Theory*, 62(3):1528–1537, 2016.

A. Rohde and A. B. Tsybakov. Estimation of high-dimensional low-rank matrices. *Ann. Statist.*, 39(2):887–930, 2011.

H. Sedghi, M. Janzamin, and A. Anandkumar. Provable tensor methods for learning mixtures of generalized linear models. In A. Gretton and C. C. Robert, editors, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 1223–1231, Cadiz, Spain, 2016. PMLR.

C. Stein. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. pages 583–602, 1972.

C. Stein, P. Diaconis, S. Holmes, and G. Reinert. Use of exchangeable pairs in the analysis of simulations. In *Stein's method: expository lectures and applications*, volume 46 of *IMS Lecture Notes Monogr. Ser.*, pages 1–26. Inst. Math. Statist., Beachwood, OH, 2004.

G. W. Stewart and J. G. Sun. *Matrix Perturbation Theory*. Computer Science and Scientific Computing. Academic Press Inc., Boston, MA, 1990.

T. Sun and C.-H. Zhang. Sparse matrix inversion with scaled lasso. *J. Mach. Learn. Res.*, 14:3385–3418, 2013.

C. Thrampoulidis, E. Abbasi, and B. Hassibi. Lasso with non-linear measurements is equivalent to one with linear measurements. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 3420–3428. Curran Associates, Inc., 2015.

R. J. Tibshirani. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B*, 58 (1):267–288, 1996.

R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In *Compressed sensing*, pages 210–268. Cambridge Univ. Press, Cambridge, 2012.

T. Wang, J. Zhang, H. Liang, and L. Zhu. Estimation of a groupwise additive multiple-index model and its applications. *Statist. Sinica*, 25(2):551–566, 2015.

Y. Xia and W. K. Li. On single-index coefficient regression models. *J. Amer. Statist. Assoc.*, 94(448):1275–1285, 1999.

L. Xue and Q. Wang. Empirical likelihood for single-index varying-coefficient models. *Bernoulli*, 18(3):836–856, 2012.

Z. Yang, K. Balasubramanian, and H. Liu. High-dimensional non-Gaussian single index models via thresholded score function estimation. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3851–3860, International Convention Centre, Sydney, Australia, 2017. PMLR.

Z. Yang, L. F. Yang, E. X. Fang, T. Zhao, Z. Wang, and M. Neykov. Misspecified nonconvex statistical optimization for sparse phase retrieval. *Math. Program.*, 176(1-2, Ser. B):545–571, 2019.

M. Yuan. On the identifiability of additive index models. *Statist. Sinica*, 21(4):1901–1911, 2011.

J. Zhang, X. Chen, and W. Zhou. High dimensional elliptical sliced inverse regression in non-gaussian distributions. *ArXiv e-prints 1801.01950*, 2017.

W. Zhao, H. Lian, and H. Liang. Quantile regression for the single-index coefficient model. *Bernoulli*, 23(3):1997–2027, 2017.

L. Zhu, B. Miao, and H. Peng. On sliced inverse regression with high-dimensional covariates. *J. Amer. Statist. Assoc.*, 101(474):630–643, 2006.