Statistical Analysis for Complex Data by Generalized Indirect Dependency Learning and Slack Empirical Likelihood

Peng Zhao

FLORIDA STATE UNIVERSITY

COLLEGE OF ARTS AND SCIENCES

STATISTICAL ANALYSIS FOR COMPLEX DATA BY GENERALIZED INDIRECT DEPENDENCY LEARNING AND SLACK EMPIRICAL LIKELIHOOD

By

PENG ZHAO

A Dissertation submitted to the
Department of Statistics
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

2020

Peng Zhao defended this dissertation on July 7, 2020.
The members of the supervisory committee were:

Yiyuan She

Professor Directing Dissertation

Zhenghao Zhang

University Representative

Fred W. Huffer

Committee Member

Jonathan R. Bradley

Committee Member

The Graduate School has verified and approved the above-named committee members, and certifies that the dissertation has been approved in accordance with university requirements.

To my parents.

ACKNOWLEDGMENTS

First of all, I would like to express my sincere gratitude to my advisors Dr. Yiyuan She. His immense knowledge and deep views on statistics and optimization strongly support my Ph.D. study. Beyond that, he helped me tremendously on my English writings and presentations. In addition, his personalities, including great diligence, critical thinking and conscientiousness not only inspire me in my research, but also raise me up on my future career pursuit. Without his outstanding help, there is no possibility for me to finish my Ph.D. study.

I would also like to thank my committee members: Dr. Zhenghao Zhang, Dr. Fred Huffer and Dr. Bradley. Their instructive comments and strong encouragement are great motivations for me to write this dissertations during such a hard outbreaking period. Special thanks to Dr. Yun Yang, who led me to the word of statistical research and taught me a lot of knowledge in nonparametric Bayesian and implicit regularizations.

Moreover, I would also like to thank the people who helped me during my years of graduate study. Special thanks my peers Wenjing, Xiaoyang and Yuqing, the time we spent together is an unforgettable experience during my life. Thanks my friends Jiahui, Jingze and Dongrui, I still remember every time we hung out together for dinners. I would like to thank my roommates Yanpeng, Zihan and Zhiji for taking care of me during my daily life. Thanks Guanyu and Hou-Cheng for help in my research and postdoc applications. I would also thank to Dr. Niu and Pam for their support in building such a harmonious department.

Finally, I want express my deepest appreciation to my parents, who always provide me their love through my entire life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Dependence is one of the most important concepts in probability and statistics. To detect and measure dependency between response variables and predictors, various models have been constructed, from simple models like least square regressions to complex ones like deep neural networks. However, less literature focuses on detecting the dependency structures among responses and incorporating this structure information to facilitate other analyses. Dependency among responses can appear when the responses are multivariate or the data are observed as groups. Markov Random Field (MRF) and Generalized Estimating Equations (GEE) are proposed to detecting the dependency structures and improve the efficiency of analysis of mean effect under these circumstances. However, these methods may not perform well when data is complex, such as non-Gaussian, heavy-tailed or skewed. In Chapter 1, we consider a generalized indirect learning with dependency (GIDL) framework to detect and apply the dependent structure between responses in both multivariate and grouped dependencies, when data are non-Gaussian, heavy-tailed or skewed. We focus on statistical analysis that covers asymptotic distribution and signal selection of dependency structures and asymptotic distribution of the coefficients of predictors with the assistant of structure information.

In Chapter 2, we develop a slack empirical likelihood (SEL) inference framework that is able to handle non-Gaussian, heavy-tailed or skewed types of data. Empirical likelihood is powerful because an inference problem is transformed into an optimization one. Besides, fewer distributional assumptions are required than traditional likelihood-based inference methods. Modern statistical models often implicitly put nonsmooth constraints through regularizations on the possible solutions for parameter estimations, like the restricted cone induced by ℓ_1 penalized regressions. Traditional empirical likelihood can not handle such nonsmooth constraints induced by regularizations when the solutions could appear at the boundary of constrained regions because the estimating equation is not sample additive. By carefully examining the directional derivatives at the nonsmooth region and introducing slack variables such that the modified estimating equations are sample additive, we extend the empirical likelihood framework such that the nonsmooth constraints can be handled by a joint optimization on parameters and the slack variable. We show some examples that our framework works for some traditional constrained empirical likelihood problems such that the

correct asymptotic distribution can be derived. Besides, some explorations on modern statistical problems including high dimensional regression with regularizations are provided.

CHAPTER 1

INFERENCE FOR GENERALIZED INDIRECT LEARNING WITH DEPENDENCY (GIDL)

1.1 Introduction

Big data arising in modern signal processing applications call for the need of analyzing multi-response systems, such as gene regulatory networks, multi-label classification, natural language processing, and stock markets, in the areas of biology, computer vision, neuroscience, economics and others. We assume a supervised setup with n samples of m responses and p predictors: $\boldsymbol{Y} = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m] \in \mathbb{R}^{n\times m}$, $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p] \in \mathbb{R}^{n\times p}$. A common linear model minimizes $\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}\|_F^2$ over $\boldsymbol{B} \in \mathbb{R}^{p\times m}$, where $\|\cdot\|_F$ denotes the Frobenius norm.

Instead of naively treating all responses as independent random variables, this article studies how we capture the *between-response dependency* that cannot be ignored in real applications, and incorporate such information into a given loss.

Most studies focus on the type of regularizer in the presence of a number of responses. But we are concerned about how to make a more reasonable loss to include the multivariate dependency. Most matrix estimation problems are formulated using an *additive* loss, like squared Frobenius norm or KL divergence, as in matrix completion, recommender systems and nonnegative matrix factorization. From a statistical perspective, this amounts to assuming all responses are independent, a simplified treatment, but a true risk of distorting the data and a loss of information. How to capture various dependencies for *non-Gaussian* systems is quite a pressing challenge. The conventional probabilistic Markov random field models, in addition to being computationally intractable, have shown severe limitations in including both positive and negative correlations on say discrete data [29, 30, 7]. Given the losses to measure the accuracy on each response, which may or may not be the same, if simply adding them together is crude and inaccurate, this article studies how to make a suitable aggregation of marginal losses to incorporate higher-order statistics into the criterion.

Denote the systematic (mean) component by $\boldsymbol{M} = \boldsymbol{XB}$ with $\boldsymbol{m}_k$ the kth ($1 \leq k \leq m$) column. As aforementioned, many multivariate methods simply utilize an additive loss to measure the discrepancy between $\boldsymbol{Y}$ and $\boldsymbol{M}$: $l_0(\boldsymbol{M}; \boldsymbol{Y}) = \sum_{k=1}^{m} l_k(\boldsymbol{m}_k, \boldsymbol{y}_k)$, where l_k is the loss for the kth response (an instance is the use of a uniform loss $\sum_{i,k} l(m_{i,k}, y_{i,k})$ like squared distance or KL divergence). Though convenient, the additive form is problematic either from the perspective of joint entropy or statistical modeling, since it implicitly assumes independence between all $\boldsymbol{y}_k$, while for large m, real life multi-response systems are almost surely dependent. For example, in multi-label classification problem, like in [25], where the authors considered a multi-label classification problem of music into emotions. Since a piece of music can belong to more than one class of emotions, the existence of correlation between different emotions is obvious. Even when the labels of emotions have already been clustered based on some psychological models [24], the labels left still enjoy some correlation: for example, a happy song is unlikely to be sad, while a relaxing song can make people feel quiet, too. Estimating with incorporation of the correlation structure can usually improve the efficiency. Another important points in multi-label classification is the mislabeled issue: since emotions are abstract and the same songs may provide inconsistent emotions to different people. The mislabeled data may cause a harmful effect to modeling if the loss functions are not properly chosen.

In longitudinal and survival analysis, it is very common to use the patients' information repeatedly collected over time (e.g., CD4 counts, blood pressures) to predict responses like severity of diseases, probability of recovery, time for death and so on. With the repeated observations over time, the correlation among values of the same response variable must be considered in the modeling to estimate efficiently. In addition, many responses have heavy-tails in distribution, for instance, the costs of health care are usually heavy-tailed [11], since people with rare disease tend to have a high cost in treatments.

There are mainly two kinds of methods to handle correlation structure in response variables when modeling the first order effect. The first one is the well-known Generalized Estimating Equations (GEE) [9]: Usually for non-Gaussian data, it is hard to model the joint distribution of the repeated observations, and GEE only assumes estimating equations for marginal distributions, avoiding the usage of joint distribution. GEE adds a working correlation matrix into the diagonal covariance matrix of the responses, and even if the working correlation matrix is misspecified, the

estimation of coefficients can still be consistent with improvement of the efficiency. However. GEE can be limited sometimes, the first problem of GEE is its non-robust to outliers, since all marginal losses for GEE are derived from exponential families. For example, by directly applying the logistic loss in the classification problem, the model could suffer a lot from the mislabeled issues. Another problem for GEE is the large sample size requirement for high dimensional estimation problems, for example, for penalized GEE proposed in [27], the number of variables p can only be in the polynomial order of the sample size n.

The other way that statisticians deal with the issue is by defining a joint Markov random field (MRF) and limiting the loss to the exponential family. This has met quite some obstacles in recent years. First, it has been realized MRF has severe limitations in modeling data dependencies. In particular, on discrete counts (prevalent in text, genomic sequencing, site-visit problems) the associated Poisson MRF can pick only negative conditional dependencies [29, 30]. Second, the partition function in MRF is often computationally intractable, and so people have to resort to various approximations of the joint likelihood [17, 6, 28]. Third, the exponential family places stringent restrictions on the marginal loss and becomes inappropriate when the data are fat-tailed, skewed, or truncated. For example, hinge loss, savage loss [12], Tukey/Hampel's robust losses [5], and many other useful losses simply do not correspond to any probability distribution (let alone one in the exponential family).

To tackle the challenges, an intriguing and ambitious topic would be to aggregate customer-provided marginal losses in a clever way to incorporate and learn dependencies. This is different from defining an entire distribution for a large number of variables which may easily break down in applications due to limited modeling capability and data imperfections. One natural question is whether there is a unique framework to generalize the weighting strategy to losses which are not necessarily derived from probability distributions. It turns out our proposed iterative weighted algorithm can achieve this goal as long as the gradient of the loss function is Lipschitz continuous.

We study generalized indirect dependent learning framework that allows for customizing marginal losses. First, since our methods are based on marginal models, there is no explicit restrictions on how to insert the correlation structure in our framework, unlike the MRF in Poisson case. In addition, by allowing to apply robust type of losses into the marginal models, our framework can handle mislabeled or extreme data. Finally, with explicit loss functions, sparse induced penalty

can be added together, which makes can our framework easily adapted to the high dimensional estimation problems.

The work is structured as follows: in Section 1.3, we provide analysis on asymptotic properties for the dependency matrix; Section 1.4 describes analysis on asymptotic properties for the coefficients of predictors; in section 1.5, we provide a computation framework for our approach.

We will use the following notations and symbols in the rest of the paper. For a matrix $\boldsymbol{A} \in \mathbb{R}^{n\times m}$, let $\|\boldsymbol{A}\|_2$ and $\|\boldsymbol{A}\|_F$ denote its Frobenius norm and spectral norm respectively. We denote vec as the vectorization operation, for example, vec $(\boldsymbol{A})$ will convert a $n \times m$ matrix to a nm dimensional column vector keeping all the consecutive orders of the column vectors of $\boldsymbol{A}$. Define the operator $\text{vec}^{-1} : \mathbb{R}^{mn} \to \mathbb{R}^{m\times n}$ such that

$$\begin{aligned} &\text{vec}^{-1}(\text{vec}(\boldsymbol{A})) = \boldsymbol{A} \text{ for all } \boldsymbol{A} \in \mathbb{R}^{m\times n} \\ &\text{vec}(\text{vec}^{-1}(\boldsymbol{\alpha})) = \boldsymbol{\alpha} \text{ for all } \boldsymbol{\alpha} \in \mathbb{R}^{mn}. \end{aligned} \tag{1.1}$$

Let $\otimes$ be the standard tensor product. Let $\mathbb{S}^m_{++}$ denote the set of all positive definite matrices of dimension $m \times m$. The inner product of two matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$ (of the same size) is defined as $\langle \boldsymbol{X}, \boldsymbol{Y}\rangle = \text{Tr}(\boldsymbol{X}^T\boldsymbol{Y})$. We use $1_{\mathcal{A}}(t)$ to denote the indicator function, i.e., $1_{\mathcal{A}}(t) = 1$ if $t \in \mathcal{A}$ and 0 otherwise. We use $\boldsymbol{X}[\mathcal{I}, \mathcal{J}]$ to denote a submatrix of $\boldsymbol{X}$ with rows and columns indexed by $\mathcal{I}, \mathcal{J}$ respectively, and abbreviate $\boldsymbol{X}[\mathcal{I}, 1 : m]$ and $\boldsymbol{X}[1 : p, \mathcal{J}]$ to $\boldsymbol{X}[\mathcal{I},]$ and $\boldsymbol{X}[, \mathcal{J}]$, respectively. When $\mathcal{I} = \{i\}$ and $\mathcal{J} = \{j\}$, we also use X_{ij} to stand for $X[\mathcal{I}, \mathcal{J}]$ for abbreviation. For a matrix $\boldsymbol{B} = [\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p]^\top$, let $\mathcal{J}(\boldsymbol{B}) = \{j : \boldsymbol{\beta}_j \neq 0\}$ and $J(\boldsymbol{B}) = |\mathcal{J}(\boldsymbol{B})| = \|\boldsymbol{B}\|_{2,0}$.

1.2 Generalized Indirect Dependent Learning

Given an observation matrix $\boldsymbol{Y} \in \mathbb{R}^{n\times m}$, the target is to learn the column dependency structure $\boldsymbol{W} \in \mathcal{S}^m_{++}$ when the pre-specified loss for the kth variable is l_k for $k = 1, 2, ..., m$, which are not necessary the same. Let $\boldsymbol{C} \in \mathbb{R}^{n\times m}, \boldsymbol{M} \in \mathbb{R}^{n\times m}$ where $\boldsymbol{M} = \boldsymbol{1}\boldsymbol{\alpha}^T$ is the mean and already known, $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_m]^T$. [23] considers the following generalized multivariate learning with dependency framework

$$\begin{aligned} \min_{(\boldsymbol{W},\boldsymbol{C})\in\mathcal{Z}} &\phi^{-1}\bar{l}(\boldsymbol{M} + \boldsymbol{C}(\boldsymbol{I} - \phi\boldsymbol{W})^{1/2}; \boldsymbol{Y}) + \frac{1}{2}\,\text{Tr}\{\boldsymbol{C}\boldsymbol{W}\boldsymbol{C}^T\} \\ &- \frac{n}{2}\log\det\boldsymbol{W} + P_W(\boldsymbol{W}; \lambda_W), \end{aligned} \tag{1.2}$$

where $\mathcal{Z} = \mathcal{S}_{++}^m(\phi) \times \mathbb{R}^{n\times m}$ with $\mathcal{S}_{++}^m(\phi) = \{\boldsymbol{W} \in \mathcal{S}_{++}^m : \boldsymbol{W} \preceq \boldsymbol{I}/\phi\}$ and $\bar{l} = \sum_{k=1}^m l_k$. Let $\boldsymbol{\Theta} = [\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_m] = \boldsymbol{C}(\boldsymbol{I} - \phi\boldsymbol{W})^{1/2}$. Assume that the $l_k(\cdot)$ is scaled by $l_k''(\alpha_k)$, such that we have

$$\nabla\bar{l}(\boldsymbol{\Theta}) = [\nabla l_1(\boldsymbol{\theta}_1), \ldots, \nabla l_m(\boldsymbol{\theta}_m)] = \nabla\bar{l}(\boldsymbol{M}) + \boldsymbol{C}\boldsymbol{Z}^{1/2}\boldsymbol{D} + \boldsymbol{\Delta}(\boldsymbol{\Theta}), \tag{1.3}$$

with $\boldsymbol{D} = \boldsymbol{I}$ and $\boldsymbol{\Delta}(\boldsymbol{\Theta})$ is the reminder term with respect to the expansion of $\nabla\bar{l}(\boldsymbol{\Theta})$.

In addition, the expansion of the loss $\bar{l}$ can also be written as:

$$\bar{l}(\boldsymbol{\Theta}) = \bar{l}(\boldsymbol{M}) + \langle\nabla\bar{l}(\boldsymbol{M}), \boldsymbol{C}\boldsymbol{Z}^{\frac{1}{2}}\rangle + \frac{1}{2}\boldsymbol{C}\boldsymbol{Z}\boldsymbol{C}^T + \delta(\boldsymbol{\Theta}), \tag{1.4}$$

where $\delta(\boldsymbol{\Theta})$ is the reminder term with respect to the expansion of $\bar{l}(\boldsymbol{\Theta})$.

Define

$$\hat{\boldsymbol{\Sigma}}^n = \frac{(\nabla\bar{l}(\boldsymbol{M}) + \boldsymbol{\Delta}^*)^T(\nabla\bar{l}(\boldsymbol{M}) + \boldsymbol{\Delta}^*)}{n}, \tag{1.5}$$

where $\boldsymbol{\Delta}^* = \boldsymbol{\Delta}(\boldsymbol{\Theta}^*)$. Let $\boldsymbol{\Sigma}^*$ be the inverse of the true dependency structure: $\boldsymbol{\Sigma}^* = \boldsymbol{W}^{*-1}$. Suppose that the true dependency structure $\boldsymbol{W}^*$ is sparse with nonzero index set $\mathcal{J}_W^*$ such that $\boldsymbol{W}_{\mathcal{J}_W^{*c}} = \boldsymbol{0}$ and $W_{ij} \neq 0$ for all $(i,j) \in \mathcal{J}_W^*$.

Some remarkable results are shown in [23]: first, as shown in Theorem 1 in this paper, when the loss is the squared loss, then the high order term $\boldsymbol{\Delta}^* = \boldsymbol{0}$ such that the optimization problem for optimization (1.2) degenerates to Gaussian graph leaning. In addition, $\hat{\boldsymbol{\Sigma}}^n$ plays a role as the sample covariance matrix, where the higher order term $\boldsymbol{\Delta}^*$ is incorporated. Finally, unlike the GEE which often fails in high large p learning, as stated in Theorem 4 in [23], even for the case $p >> n$, the above optimization can still deliver a reasonable estimating results.

1.3 Asymptotic Analysis for the Dependency Matrix

1.3.1 Main Results

In this part, we show our main theorem. To be more specific, we show that under certain regularity conditions, the signal part of the true dependency matrix can be identified, in addition, the signals follow a multivariate normal distribution, which can be directly applied into inference via Hotelling's T-squared statistics. The following conditions are necessary in our main theorem:

1. Effective noise with respect to the mean: $\mathbb{E}(-\nabla\bar{l}(\boldsymbol{M})) = \boldsymbol{0}$;
2. Effective noise with respect to $\boldsymbol{W}^*$: $\mathbb{E}\{-\hat{\boldsymbol{\Sigma}}^n + \boldsymbol{W}^{*-1}\}$=$\boldsymbol{0}$;

3. The second order moment of the effective noise with respect to the true signal part of $\boldsymbol{W}^*$:

$$\mathbb{E}\left\{(\hat{\boldsymbol{\Sigma}}^n_{\mathcal{J}^*_W} - \boldsymbol{\Sigma}^*_{\mathcal{J}^*_W}) \otimes (\hat{\boldsymbol{\Sigma}}^n_{\mathcal{J}^*_W} - \boldsymbol{\Sigma}^*_{\mathcal{J}^*_W})\right\}$$

is positive-definite and finite;

4. Let $\sigma_{\max}(\boldsymbol{W}^*)$ and $\sigma_{\min}(\boldsymbol{W}^*)$ be the maximal and minimal eigenvalues of $\boldsymbol{W}^*$, then there exist positive constants $\gamma_{\min}$ and $\gamma_{\max}$ such that $\gamma_{\min} < \sigma_{\min}(\boldsymbol{W}^*) < \sigma_{\max}(\boldsymbol{W}^*) < \gamma_{\max}$;

5. There exists an open set $\Omega \subset \mathcal{Z}$ that contains true $(\boldsymbol{W}^*, \boldsymbol{C}^*)$ that for almost all $\boldsymbol{y}$ all $(\boldsymbol{W}, \boldsymbol{C}) \in \Omega$, we have $\|\boldsymbol{\Delta}(\boldsymbol{\Theta})\|_F = o_p(1)$ and $\|\delta(\boldsymbol{\Theta})/\mathrm{d}\boldsymbol{W}\|_F = o_p(n^{1/2})$.

6. Properties of the penalty:

$$\begin{aligned}
\lim_{n\to\infty} \max(P''_W(\boldsymbol{W}^*_{\mathcal{J}^*_W}; \lambda_W))/n = 0;\\
\lim_{n\to\infty} \lim_{t\to 0+} P'_W(t; \lambda_W)/\lambda_W > 0;\\
\max(P'_W(\boldsymbol{W}^*_{\mathcal{J}^*_W}; \lambda_W))/n = O_p(n^{-1/2}); \quad \text{and}
\end{aligned}$$

there exist constants C_1, C_2 such that when $t_1, t_2 > C_1\lambda_n$, we have $|P''_W(t_1; \lambda_n) - P''_W(t_2; \lambda_n)| \leq C_2|t_1 - t_2|$;

7. Choices of the tuning parameter:

$$\lim_{n\to\infty} \lambda_W/n = 0, \quad \lim_{n\to\infty} \lambda_W/\sqrt{n} \to \infty,$$

Then we have the following theorem:

Theorem 1.3.1. *Suppose the above conditions* $1-7$ *provided are satisfied, let* $a_n = \max(P'_W(\boldsymbol{W}^*_{\mathcal{J}^*_W}; \lambda_W)/n)$, *with probability tending to* 1, *we have:*

1. *There is a local minimizer* $\widehat{\boldsymbol{W}}$ *such that* $\|\widehat{\boldsymbol{W}} - \boldsymbol{W}^*\|_F = O(n^{-1/2} + a_n)$;
2. *The above defined local minimizer* $\widehat{\boldsymbol{W}}$ *satisfies* $\widehat{\boldsymbol{W}}_{\mathcal{J}_{\mathcal{W}}{}^{*c}} = \boldsymbol{0}$;
3. *Asymptotic normality:* $\sqrt{n}(vec\, \widehat{\boldsymbol{W}}_{\mathcal{J}^*_W} - vec\, \boldsymbol{W}^*_{\mathcal{J}_W})$ *converges to a normal distribution with mean zero and covariance matrix:*

$$\{\boldsymbol{W}^*_{\mathcal{J}^*_W} \otimes \boldsymbol{W}^*_{\mathcal{J}^*_W}\} \lim_{n\to\infty} \mathrm{Cov}(vec\,[-\sqrt{n}\{\hat{\boldsymbol{\Sigma}}^n_{\mathcal{J}^*_W} - \boldsymbol{\Sigma}^*_{\mathcal{J}^*_W}\}])\{\boldsymbol{W}^*_{\mathcal{J}^*_W} \otimes \boldsymbol{W}^*_{\mathcal{J}^*_W}\}, \tag{1.6}$$

where recall that we define $\hat{\boldsymbol{\Sigma}}^n = (\nabla \bar{l}(\boldsymbol{M}) + \Delta^*)^T(\nabla \bar{l}(\boldsymbol{M}) + \Delta^*)/n$ *and* $\boldsymbol{\Sigma}^* = \boldsymbol{W}^{*-1}$.

1.3.2 Equivalence to Pseudo-likelihood for Ising Models

In this part, we show the equivalence of selection between pseudo-likelihood model and GIDL in the statistical sense when the sample size $n \to \infty$. Consider the Ising model:

$$p(\boldsymbol{y}, \boldsymbol{W}) = \exp\left[\sum_{(s,t)\in E} W_{st} y_s y_t - \Psi(\boldsymbol{W})\right], \tag{1.7}$$

where $\boldsymbol{y} = (y_1, ..., y_m)^T \in \{-1, 1\}^m$, the vertices $\boldsymbol{V} = \{1, 2, \ldots, m\}$, the edges E captures the conditional independence between variables of $\boldsymbol{y}$, and $\Psi(\boldsymbol{W}) = \log \sum_{y_s, y_t \in \{-1,1\}}^{m} \exp\{\sum_{(s,t)\in E} W_{st} y_s y_t\}$ is the normalizing constant. Due to the infeasible computation of the normalizing constant. The pseudo-likelihood approach is often considered [6], which is the direct aggregate of all conditional likelihoods. By conditional likelihood of y_s given all others and normalizing, given the index s, let $\{-s\}$ be the index set of all other response variables except of s, we have:

$$p\left(y_s, \boldsymbol{W} | \mathbf{y}_{\{-s\}}\right) = \frac{\exp\left(\sum_{t\neq s} W_{st} y_s y_t\right)}{\exp\left(\sum_{t\neq s} W_{st} y_t\right) + \exp\left(-\sum_{t\neq s} W_{st} y_t\right)}. \tag{1.8}$$

Then the log-pseudo-likelihood is defined by

$$\begin{aligned} l(\boldsymbol{W}|\mathbf{y}) &= \sum_{s=1}^{m} \log p\left(y_s, \boldsymbol{W} | \mathbf{y}_{\{-s\}}\right) \\ &= \sum_{s=1}^{m} \left[\sum_{t\neq s} y_s y_t W_{st} - \log\cosh(\sum_{t\neq s} W_{st} y_t) - \log 2\right]. \end{aligned} \tag{1.9}$$

Then for data matrix $\boldsymbol{Y} \in \{-1, 1\}^{n\times m}$, we have the log-pseudo-likelihood:

$$l(\boldsymbol{W}|\boldsymbol{Y}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{s=1}^{m}\left[\sum_{t\neq s} Y_{is} Y_{it} W_{st} - \log\cosh(\sum_{t\neq s} W_{st} Y_{it}) - \log 2\right]. \tag{1.10}$$

To minimize $-l(\boldsymbol{W}|\boldsymbol{Y})$, the first order derivative can be calculated:

$$-\frac{\partial l}{\partial W_{sk}} = -\frac{1}{n}\sum_{i=1}^{n}\left[Y_{is}Y_{ik} - Y_{ik}\tanh(\sum_{t\neq s} W_{st} Y_{it})\right]. \tag{1.11}$$

Note that

$$\tanh(\sum_{t\neq s} W_{st} Y_{it}) = \frac{\exp(\sum_{t\neq s} W_{st} Y_{it}) - \exp(-\sum_{t\neq s} W_{st} Y_{it})}{\exp(\sum_{t\neq s} W_{st} Y_{it}) + \exp(-\sum_{t\neq s} W_{st} Y_{it})} = \mathrm{E}(Y_{is}|\{Y_{it}, t\neq s\}) \tag{1.12}$$

is the conditional expectation. Consistency of pseudo-likelihood when $n \to \infty$ can be shown by the fact that Y_{ik} is uncorrelated with $Y_{is} - \mathrm{E}(Y_{is}|\{Y_{it}, t \neq s\})$ and the identity $\mathbb{E}(Y_{is} - \mathbb{E}(Y_{is}|\{Y_{it}, t \neq s\})) = 0$.

Theorem 1.3.2. *Suppose the data matrix* $\boldsymbol{Y} \in \{-1, 1\}^{n\times m}$ *is* n *independent observations generated from density* $P(\boldsymbol{y}, \boldsymbol{W}^*)$ *defined in equation* (1.7). *Let* $\boldsymbol{W}^{pse}$ *be the estimator through pseudo-likelihood approach. If the conditions in Theorem 1.3.1 are satisfied. Then as* $n \to \infty$, *if for any index* s, k, *the following condition holds*

$$\frac{1}{n}\sum_{i=1}^{n}(Y_{ik}\boldsymbol{Y}_{i\{-s\}}(\boldsymbol{Y}_{\{-s\}}^T\boldsymbol{Y}_{\{-s\}})^{-1}\boldsymbol{Y}_{\{-s\}}^T\boldsymbol{Y}_s)^3 \to 0. \tag{1.13}$$

Then there exists a stationary point for our GIDL method, denoted as $\boldsymbol{W}^{gidl}$, *such that*

$$P(nzoff(\boldsymbol{W}^{gidl}) = nzoff(\boldsymbol{W}^{pse})) \to 1, \tag{1.14}$$

where $nzoff(\boldsymbol{W})$ *is the index set for the non-zero components on the off-diagonal components of* $\boldsymbol{W}$.

Remark 1.3.1. *The regularity condition* (1.13) *puts a high order constraint on the graph structure of the Ising model. Similar constraints on graph structure are also required in other literature, for example in [10], the graph with singleton separator sets is required .*

1.4 Generalized Indirect Dependency Learning with Predictors

Although we formulate the problem by introducing the dependency between columns of response matrix, we show that through converting the problem to the vectorized case. The clustered data type of problem can also be handled. Let us first illustrate our motivation from multivariate Gaussian case: for multiple responses case, let $\boldsymbol{y} = \mathrm{vec}\,(\boldsymbol{Y})$, $\boldsymbol{L} = \boldsymbol{I}_{m\times m} \otimes \boldsymbol{X}$ and $\boldsymbol{\beta} = \mathrm{vec}\,(\boldsymbol{B})$, then the multivariate Gaussian distribution can be represented as $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{L}\boldsymbol{\beta}, \boldsymbol{S})$, where $\boldsymbol{S} \in \mathbb{S}_{++}^{mn}$ captures the covariance structure of the column vectors of $\boldsymbol{Y}$. For example, for multiple response problems, $\boldsymbol{S} = \boldsymbol{\Sigma} \otimes \boldsymbol{I}_{n\times n}$, where $\boldsymbol{\Sigma}$ is the covariance structure between response variables. Now suppose the loss is quadratic for $\boldsymbol{y} - \boldsymbol{L}\boldsymbol{\beta}$, then we have the estimating equation $\boldsymbol{L}^T(\boldsymbol{L}\boldsymbol{\beta} - \boldsymbol{y}) = \boldsymbol{0}$. In the Gaussian case we can directly add the weighting matrix into the estimation equations:

$$\boldsymbol{L}^T\boldsymbol{S}^{-1}(\boldsymbol{L}\boldsymbol{\beta} - \boldsymbol{y}) = \boldsymbol{0}. \tag{1.15}$$

Note that when $\boldsymbol{S} = \boldsymbol{\Sigma} \otimes \boldsymbol{I}_{n\times n}$,

$$\begin{aligned}
\boldsymbol{L}^T\boldsymbol{S}^{-1}(\boldsymbol{L}\boldsymbol{\beta} - \boldsymbol{y}) &= (\boldsymbol{I} \otimes \boldsymbol{X}^T)(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{I})(\boldsymbol{L}\boldsymbol{\beta} - \boldsymbol{y}) \\
&= (\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{X}^T)(\boldsymbol{L}\boldsymbol{\beta} - \boldsymbol{y}) \\
&= (\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{X}^T)\,\mathrm{vec}\,(\boldsymbol{X}\boldsymbol{B} - \boldsymbol{Y}) \\
&= \mathrm{vec}\,(\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{B} - \boldsymbol{Y})\boldsymbol{\Sigma}^{-1}) = \boldsymbol{0},
\end{aligned} \tag{1.16}$$

which is reduced to the ordinary estimating equation for multivariate Gaussian case.

For clustered data case, $\boldsymbol{Y} \in \mathbb{R}^{n\times m}$, where n is the number of observation and m is the number of groups. $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ is the predictor matrix and $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ is the coefficient. Now the coefficient is just a vector because there are only measurement errors within each cluster and sharing the same treatment effect $\boldsymbol{\beta}_0$. Following the same idea, let $\boldsymbol{W} \in \mathbb{R}^{n\times n}$ is the dependency matrix for each cluster. By vectorization, let $\boldsymbol{L} = \boldsymbol{I}_{m\times m} \otimes \boldsymbol{X}$ and $\boldsymbol{\beta} = \boldsymbol{I}_{m\times 1} \otimes \boldsymbol{\beta}_0$, then the distribution can also be specified if likelihood is normal: $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{L}\boldsymbol{\beta}, \boldsymbol{S})$, where $\boldsymbol{S} = (\boldsymbol{I}_{m\times m} \otimes \boldsymbol{W})^{-1}$. That is to say, the differences between clustered data case and multiple response case is that for clustered data, the dependency is introduced among rows rather than columns. In addition, if we formulate the coefficient $\boldsymbol{\beta}$ into a $\mathbb{R}^{p\times m}$ matrix, then all columns should be the same. For convenience, we show all the following results in the vectorized multiple responses case, but similar analysis technique can be applied on clustered data.

Now if we have a response data matrix $\boldsymbol{Y}$ which is not multivariate Gaussian, and the loss l we want to evaluate is also not ℓ_2 type, even not necessarily derived from a probability distribution. Like in traditional weighted least squares, we prespecify an estimated inverse covariance matrix $\boldsymbol{W} \in \mathbb{S}^m_{++}$ of $\boldsymbol{Y}$, which captures the conditional independence between the response variables ($\boldsymbol{S}^{-1}$ in multivariate Gaussian case). In order to incorporate the covariance structure into estimation beyond the Gaussian case, we consider an auxiliary vector $\boldsymbol{c}$ as random effects, with a design matrix $(\boldsymbol{I} - \phi \boldsymbol{W} \otimes \boldsymbol{I})^{1/2}$. Given the estimated $\boldsymbol{W}$, let $\boldsymbol{\theta} = \boldsymbol{L}\boldsymbol{\beta} + (\boldsymbol{I} - \phi \boldsymbol{W} \otimes \boldsymbol{I})^{1/2}\boldsymbol{c}$, $\boldsymbol{\Theta} = \text{vec}^{-1}(\boldsymbol{\theta})$, and $\boldsymbol{C} = \text{vec}^{-1}(\boldsymbol{c})$, we first consider the multi-dependency learning framework to without regularizations on the estimated coefficients of predictors:

$$\min_{\boldsymbol{\beta},\boldsymbol{c},\boldsymbol{W}} \phi^{-1} l(\boldsymbol{\theta}; \boldsymbol{y}) + \frac{1}{2}\boldsymbol{c}^T(\boldsymbol{W} \otimes \boldsymbol{I})\boldsymbol{c} - \frac{n}{2}\log\det \boldsymbol{W} + P_W(\boldsymbol{W}; \lambda_W). \tag{1.17}$$

Denote $F(\boldsymbol{\beta}, \boldsymbol{W}, \boldsymbol{c})$ as the objective function, by taking $\nabla_{\boldsymbol{c}} F = \boldsymbol{0}$, $\nabla_{\boldsymbol{\beta}} F = 0$, we can obtain the estimation equation:

$$\phi^{-1}(\boldsymbol{I} - \phi \boldsymbol{W} \otimes \boldsymbol{I})^{1/2} \nabla l(\boldsymbol{\theta}) + (\boldsymbol{W} \otimes \boldsymbol{I})\boldsymbol{c} = \boldsymbol{0}, \tag{1.18}$$

$$\phi^{-1}\boldsymbol{L}^T \nabla l(\boldsymbol{\theta}) = \boldsymbol{0}, \tag{1.19}$$

For the Gaussian case, our GIDL framework can help to estimate $\boldsymbol{\beta}$ with the structure information $\boldsymbol{W}$. In fact, when l in equation (1.17) is the quadratic loss, plugging the loss into the gradient

equation (1.18), we have

$$\phi^{-1}(\boldsymbol{I}-\phi\boldsymbol{W}\otimes\boldsymbol{I})^{1/2}(\boldsymbol{L}\boldsymbol{\beta}-\boldsymbol{y}+(\boldsymbol{I}-\phi\boldsymbol{W}\otimes\boldsymbol{I})^{1/2}\boldsymbol{c})+(\boldsymbol{W}\otimes\boldsymbol{I})\boldsymbol{c}=\boldsymbol{0},$$

where $\boldsymbol{c}=(\boldsymbol{I}-\phi\boldsymbol{W}\otimes\boldsymbol{I})^{1/2}(\boldsymbol{y}-\boldsymbol{L}\boldsymbol{\beta})$ can be obtained. Plugging the form of $\boldsymbol{c}$ back into the optimization problem,

$$\begin{aligned}
&\min_{\boldsymbol{\beta},\boldsymbol{c},\boldsymbol{W}} \phi^{-1}l(\boldsymbol{\theta};\boldsymbol{y})+\frac{1}{2}\boldsymbol{c}^T(\boldsymbol{W}\otimes\boldsymbol{I})\boldsymbol{c}+P_W(\boldsymbol{W};\lambda_W)\\
=&\min_{\boldsymbol{\beta},\boldsymbol{W}} \phi^{-1}\|(\phi\boldsymbol{W}\otimes\boldsymbol{I})(\boldsymbol{y}-\boldsymbol{L}\boldsymbol{\beta})\|_2^2+\frac{1}{2}(\boldsymbol{y}-\boldsymbol{L}\boldsymbol{\beta})^\top(\boldsymbol{I}-\phi\boldsymbol{W}\otimes\boldsymbol{I})(\boldsymbol{W}\otimes\boldsymbol{I})(\boldsymbol{y}-\boldsymbol{L}\boldsymbol{\beta})+P_W(\boldsymbol{W};\lambda_W)\\
=&\min_{\boldsymbol{\beta},\boldsymbol{W}} \frac{1}{2}\{(\boldsymbol{y}-\boldsymbol{L}\boldsymbol{\beta})^\top(\boldsymbol{W}\otimes\boldsymbol{I}^{\frac{1}{2}})(\boldsymbol{W}\otimes\boldsymbol{I}^{\frac{1}{2}})(\boldsymbol{y}-\boldsymbol{L}\boldsymbol{\beta})\}+P_W(\boldsymbol{W};\lambda_W)\\
\overset{(i)}{=}&\min_{\boldsymbol{B},\boldsymbol{W}} \frac{1}{2}\operatorname{vec}((\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{B})\boldsymbol{W}^{\frac{1}{2}})^\top\operatorname{vec}((\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{B})\boldsymbol{W}^{\frac{1}{2}})+P_W(\boldsymbol{W};\lambda_W)\\
=&\min_{\boldsymbol{B},\boldsymbol{W}} \operatorname{Tr}\{(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{B})\boldsymbol{W}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{B})^\top\}+P_W(\boldsymbol{W};\lambda_W),
\end{aligned}$$

where in the first equality we use the fact that $\boldsymbol{I}-\phi\boldsymbol{W}\otimes\boldsymbol{I}$ and $\boldsymbol{W}\otimes\boldsymbol{I}$ are exchangeable, and in (i) we apply the identity $\operatorname{vec}((\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{B})\boldsymbol{W}^{1/2})=(\boldsymbol{W}^{1/2}\otimes\boldsymbol{I})\operatorname{vec}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{B})$. Then the optimization problem in equation (1.17) is equivalent with

$$\operatorname*{argmin}_{\boldsymbol{B},\boldsymbol{W}} \frac{1}{2}\{(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{B})\boldsymbol{W}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{B})^T\}-\frac{n}{2}\log\det\boldsymbol{W}+P_W(\boldsymbol{W};\lambda_W). \tag{1.20}$$

Remark 1.4.1. *We can see the advantages of applying the GIDL framework* (1.17) *to non-Gaussian losses:*

1. *The loss function l can be customized directly while the structure information still plays a role in estimating the first order effect;*
2. *With the explicit appearance of loss functions, we can directly add regularizations on the loss to deal with high dimensional estimation problems, which is more nature comparing with penalization on the estimating equations (e.g. [27]).*

For the non-Gaussian cases, we expand the $\nabla l(\boldsymbol{\theta})$ to the second order term: $\boldsymbol{\delta}(\boldsymbol{\beta};\boldsymbol{W};\boldsymbol{c})=\nabla l(\boldsymbol{\theta})-\nabla l(\boldsymbol{L}\boldsymbol{\beta})-\nabla^2 l(\boldsymbol{L}\boldsymbol{\beta})(\boldsymbol{\theta}-\boldsymbol{L}\boldsymbol{\beta})$. Denote $\boldsymbol{Z}=\boldsymbol{I}-\phi\boldsymbol{W}\otimes\boldsymbol{I}$, $\boldsymbol{D}(\boldsymbol{\beta})=\nabla^2 l(\boldsymbol{L}\boldsymbol{\beta})$, we now try to get rid of $\boldsymbol{c}$ to obtain the estimating equations only for $\boldsymbol{\beta}$ given $\boldsymbol{W}$. Based on the equation (1.18), we have:

$$\nabla l(\boldsymbol{L}\boldsymbol{\beta})+\boldsymbol{\delta}(\boldsymbol{\beta};\boldsymbol{W};\boldsymbol{c})+\boldsymbol{D}(\boldsymbol{\beta})\boldsymbol{Z}^{1/2}\boldsymbol{c}=-\phi\boldsymbol{Z}^{-1/2}(\boldsymbol{W}\otimes\boldsymbol{I})\boldsymbol{c}, \tag{1.21}$$

then rearranging the terms, we can obtain:

$$\begin{aligned}\boldsymbol{Z}^{-1/2}\boldsymbol{c} &= -\boldsymbol{Z}^{-1/2}(\boldsymbol{Z}^{1/2}\boldsymbol{D}(\boldsymbol{\beta})\boldsymbol{Z}^{1/2}+\phi\boldsymbol{W}\otimes\boldsymbol{I})^{-1}\boldsymbol{Z}^{1/2}(\nabla l(\boldsymbol{L}\boldsymbol{\beta})+\boldsymbol{\delta}(\boldsymbol{\beta};\boldsymbol{W};\boldsymbol{c}))\\ &= -[\boldsymbol{D}(\boldsymbol{\beta})\boldsymbol{Z}+\phi\boldsymbol{Z}^{-1/2}(\boldsymbol{W}\otimes\boldsymbol{I})\boldsymbol{Z}^{1/2}]^{-1}(\nabla l(\boldsymbol{L}\boldsymbol{\beta})+\boldsymbol{\delta}(\boldsymbol{\beta};\boldsymbol{W};\boldsymbol{c})) \qquad (1.22)\\ &= -[\boldsymbol{D}(\boldsymbol{\beta})(\boldsymbol{I}-\phi\boldsymbol{W}\otimes\boldsymbol{I})+\phi\boldsymbol{W}\otimes\boldsymbol{I}]^{-1}(\nabla l(\boldsymbol{L}\boldsymbol{\beta})+\boldsymbol{\delta}(\boldsymbol{\beta};\boldsymbol{W};\boldsymbol{c})),\end{aligned}$$

where in the second equality we use the fact that $\boldsymbol{Z}$ and $\boldsymbol{W}\otimes\boldsymbol{I}$ are exchangeable. Plugging the obtained identity to the equation (2.40), we have:

$$\boldsymbol{L}^T(\boldsymbol{W}\otimes\boldsymbol{I})[\boldsymbol{D}(\boldsymbol{\beta})(\boldsymbol{I}-\phi\boldsymbol{W}\otimes\boldsymbol{I})+\phi\boldsymbol{W}\otimes\boldsymbol{I}]^{-1}(\nabla l(\boldsymbol{L}\boldsymbol{\beta})+\boldsymbol{\delta}(\boldsymbol{\beta};\boldsymbol{W};\boldsymbol{c}))=\boldsymbol{0}, \qquad (1.23)$$

Now we incorporate equation (1.18) into the above estimating equation. With a little abuse of notation, denote

$$\boldsymbol{\delta}(\boldsymbol{\beta};\boldsymbol{W})=\boldsymbol{\delta}(\boldsymbol{\beta};\boldsymbol{W};\boldsymbol{c}^\star(\boldsymbol{\beta};\boldsymbol{W})), \qquad (1.24)$$

where $\boldsymbol{c}^\star(\boldsymbol{\beta})$ is a root of $\boldsymbol{c}$ for equation (1.18) given $\boldsymbol{\beta}$ and $\boldsymbol{W}$. Therefore, the final estimating equation becomes:

$$\boldsymbol{L}^T(\boldsymbol{W}\otimes\boldsymbol{I})[\boldsymbol{D}(\boldsymbol{\beta})+\phi(\boldsymbol{W}\otimes\boldsymbol{I})(\boldsymbol{I}-\boldsymbol{D}(\boldsymbol{\beta}))]^{-1}(\nabla l(\boldsymbol{L}\boldsymbol{\beta})+\boldsymbol{\delta}(\boldsymbol{\beta};\boldsymbol{W}))=\boldsymbol{0}. \qquad (1.25)$$

Notice that when l is quadratic loss, $\boldsymbol{D}(\boldsymbol{\beta})=\boldsymbol{I}$, $\boldsymbol{\delta}=\boldsymbol{0}$, then the estimating equation reduces to weighted least squares given $\boldsymbol{W}$ as shown in equation (1.16). When $\boldsymbol{\delta}=0$ and $\phi=0$, it seems we insert a covariance matrix into the traditional GLM estimating equations, just like GEE.

1.4.1 Asymptotic Analysis for Coefficients

Consider the case when $n\to\infty$ while p,m are fixed. Define the loss function as

$$\begin{aligned}\bar{l}(\boldsymbol{L}\boldsymbol{\beta},\boldsymbol{y}) &= \sum_{k=1}^{m} l_k(\boldsymbol{X}\boldsymbol{B}[,k],\boldsymbol{Y}[,k]) = \sum_{i=1}^{n}\sum_{k=1}^{m} l_k(\boldsymbol{X}[i,]\boldsymbol{B}[,k],\boldsymbol{Y}[i,k]) \qquad (1.26)\\ &= \sum_{i=1}^{n}\bar{l}(\boldsymbol{L}[,m(i)]\boldsymbol{\beta}[m(i)],\boldsymbol{y}[m(i)]), \qquad (1.27)\end{aligned}$$

where $m(i)=\{i+jn\}_{j=0,1,\ldots,m-1}$ and l_k is the prespecified loss for the kth variable of $\boldsymbol{Y}$. Define

$$\bar{l}_i(\boldsymbol{L}\boldsymbol{\beta},\boldsymbol{y})=\bar{l}(\boldsymbol{L}[,m(i)]\boldsymbol{\beta}[m(i)],\boldsymbol{y}[m(i)]), \qquad (1.28)$$

then for $i=1,2,\ldots,n$, $\bar{l}_i$ are independent and follow the same distribution, since they are the losses for each independent observations, with each observation including all response variables. Then we have the asymptotic property of the estimator $\hat{\boldsymbol{\beta}}$:

Theorem 1.4.1. *Suppose the all loss functions l_k, $k = 1, \ldots, m$ have bounded second order derivatives and denote $\boldsymbol{\beta}^*$ as the true vectorized coefficients. Assume*

1. *Given $\boldsymbol{\beta}, \boldsymbol{W}$, the solution for axillary variable $\boldsymbol{c}$ exists for equation* (1.18)*;*
2. *Given $\boldsymbol{\beta}$, the dependency matrix $\boldsymbol{W}$ can be estimated $O_p(1/\sqrt{n})$ consistently converging to a constant positive-definite matrix $\overline{\boldsymbol{W}}$, which can be different from the truth;*
3. *The high order term $\boldsymbol{\delta}(\boldsymbol{\beta}^*; \boldsymbol{W})$ satisfies $\sqrt{n}\boldsymbol{\delta}(\boldsymbol{\beta}^*; \boldsymbol{W}) \to \mathbf{0}$ and $\| \frac{\partial \boldsymbol{\delta}}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}^*; \boldsymbol{\delta}(\boldsymbol{\beta}^*, \boldsymbol{W}))\|_2/n \to 0$ for any $\boldsymbol{W}$ if $n \to \infty$.*

Then let $\hat{\boldsymbol{\beta}}$ be the solution of the optimization problem

$$\min_{\boldsymbol{\beta}, \boldsymbol{W}, \boldsymbol{c}} \phi^{-1} \bar{l}(\boldsymbol{L}\boldsymbol{\beta} + (\boldsymbol{I} - \phi \boldsymbol{W} \otimes \boldsymbol{I})^{1/2}\boldsymbol{c}; \boldsymbol{y}) + \frac{1}{2}\boldsymbol{c}^T(\boldsymbol{W} \otimes \boldsymbol{I})\boldsymbol{c} + P_W(\boldsymbol{W}; \lambda_W), \tag{1.29}$$

where $\bar{l}$ are defined in equation (1.26)*, $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)$ is asymptotically multivariate Gaussian with zero mean and covariance matrix:*

$$\begin{aligned}\boldsymbol{V} = \lim_{n\to\infty} n[\boldsymbol{L}^T\overline{\boldsymbol{W}}_T[\boldsymbol{D}(\boldsymbol{\beta}^*) + \phi\overline{\boldsymbol{W}}_T(\boldsymbol{I} - \boldsymbol{D}(\boldsymbol{\beta}^*))]^{-1}\boldsymbol{D}(\boldsymbol{\beta}^*)\boldsymbol{L}]^{-1} \\ [\boldsymbol{L}^T\overline{\boldsymbol{W}}_T[\boldsymbol{D}(\boldsymbol{\beta}^*) + \phi\overline{\boldsymbol{W}}_T(\boldsymbol{I} - \boldsymbol{D}(\boldsymbol{\beta}^*))]^{-1}\mathrm{Cov}(\boldsymbol{y})[\boldsymbol{D}(\boldsymbol{\beta}^*) + \phi\overline{\boldsymbol{W}}_T(\boldsymbol{I} - \boldsymbol{D}(\boldsymbol{\beta}^*))]^{-1}\overline{\boldsymbol{W}}_T\boldsymbol{L}] \\ [\boldsymbol{L}^T\overline{\boldsymbol{W}}_T[\boldsymbol{D}(\boldsymbol{\beta}^*) + \phi\overline{\boldsymbol{W}}_T(\boldsymbol{I} - \boldsymbol{D}(\boldsymbol{\beta}^*))]^{-1}\boldsymbol{D}(\boldsymbol{\beta}^*)\boldsymbol{L}]^{-1},\end{aligned}$$

where $\overline{\boldsymbol{W}}_T = \overline{\boldsymbol{W}} \otimes \boldsymbol{I}$.

Remark 1.4.2. *For the Gaussian case, $\boldsymbol{D}(\boldsymbol{\beta}^*) = \boldsymbol{I}$, then the covariance matrix is reduced to $\boldsymbol{V} = \lim_{n\to\infty} n[\boldsymbol{L}^\top\overline{\boldsymbol{W}}_T\boldsymbol{L}]^{-1}[\boldsymbol{L}^\top\overline{\boldsymbol{W}}_T\mathrm{Cov}(\boldsymbol{y})\overline{\boldsymbol{W}}_T\boldsymbol{L}][\boldsymbol{L}^\top\overline{\boldsymbol{W}}_T\boldsymbol{L}]^{-1}$. In addition, when $\overline{\boldsymbol{W}}$ is correctly specified as the truth of inverse covariance matrix, then $\boldsymbol{V} = \lim_{n\to\infty} n[\boldsymbol{L}^\top\overline{\boldsymbol{W}}_T\boldsymbol{L}]^{-1}$.*

Remark 1.4.3. *The conditions required in the above theorem are similar with clustered-specific model for GEE where there is also a random effect in GEE [31]. Both the random effect $\boldsymbol{c}$ and the high order terms $\boldsymbol{\delta}$ need to be regularized properly. However, for GEE with a random effect, the distribution of the random effect needs to be known as a priori, while for our model, the random effect $\boldsymbol{c}$ is defined in an optimization way. Our model is computational more efficient since optimization is usually more feasible than integral over a distribution.*

Remark 1.4.4. *When $\phi = 0$, the asymptotic covariance matrix has the similar form with GEE. However, there is a main difference between our estimator and GEE: for finite sample estimation,*

our method incorporates the high order information $\boldsymbol{\delta}(\boldsymbol{\beta}; \boldsymbol{W})$ into estimation. This could be more desirable just like in literature about bootstrap, bootstrapping the entire estimator could produce better results than only the linearized part of the estimator (e.g. [3]).

1.5 Algorithm Design

Construct a surrogate function by linearizing the first term $\phi^{-1} l_0$ only at $\boldsymbol{\Theta} = \boldsymbol{M} + \boldsymbol{C}(\boldsymbol{I} - \phi\boldsymbol{W})^{1/2}$ as a whole:

$$g(\boldsymbol{M}, \boldsymbol{W}, \boldsymbol{C}; \boldsymbol{M}^{[k]}, \boldsymbol{W}^{[k]}, \boldsymbol{C}^{[k]}) = \phi^{-1}\, l_0(\boldsymbol{\Theta}^{[k]}; \boldsymbol{Y}) + \phi^{-1}\langle \nabla l_0(\boldsymbol{\Theta}^{[k]}), \boldsymbol{\Theta} - \boldsymbol{\Theta}^{[k]} \rangle + \frac{\rho}{2\phi}\|\boldsymbol{\Theta} - \boldsymbol{\Theta}^{[k]}\|_F^2 + \frac{1}{2} Tr\{\boldsymbol{C}\boldsymbol{W}\boldsymbol{C}^T\} - \frac{n}{2}\log\det \boldsymbol{W} + P_M(\boldsymbol{M}) + P_W(\boldsymbol{W})$$

Define the next iterate as $(\boldsymbol{M}^{[k+1]}, \boldsymbol{W}^{[k+1]}, \boldsymbol{C}^{[k+1]}) = \mathrm{argmin}_{\boldsymbol{M}\in\mathcal{X}, (\boldsymbol{W},\boldsymbol{C})\in\mathcal{Z}}\, g(\boldsymbol{M}, \boldsymbol{W}, \boldsymbol{C}; \boldsymbol{M}^{[k]}, \boldsymbol{W}^{[k]}, \boldsymbol{C}^{[k]})$. Assume ∇l_0 is L-Lipschitz continuous. Then, as long as $\rho \geq L$, the sequence of iterates guarantees the objective function values to be non-increasing and so convergent. Rewrite the g-optimization

$$\min_{\boldsymbol{M},\boldsymbol{W},\boldsymbol{C}} \frac{\rho}{2\phi}\|\boldsymbol{M} + \boldsymbol{C}(\boldsymbol{I} - \phi\boldsymbol{W})^{1/2} - \boldsymbol{\Xi}^{[k+1]}\|_F^2 + \frac{1}{2} Tr\{\boldsymbol{C}\boldsymbol{W}\boldsymbol{C}^T\} - \frac{n}{2}\log\det(\boldsymbol{W}) + P_M(\boldsymbol{M}) + P_W(\boldsymbol{W}),$$
$$\text{with } \boldsymbol{\Xi}^{[k+1]} = \boldsymbol{\Theta}^{[k]} - \nabla_{\boldsymbol{\Theta}} l_0(\boldsymbol{\Theta}^{[k]})/\rho,$$

and $1/\rho$ amounts to the step size. The problem can be further simplified—we can scale l_0 by L beforehand and set $\rho = 1$. Then we are back to the plain multivariate Gaussian estimation.

Another appealing fact is that the auxiliary matrix $\boldsymbol{C}$ does *not* have to be explicitly computed at all. Forming $\boldsymbol{\Xi}^{[k]}$ avoids the need of computing SVD or matrix square-roots. Concretely, the key quantity $\boldsymbol{\Theta}^{[k]}$ can be conveniently written as a weighted average of $\boldsymbol{M}$ and $\boldsymbol{\Xi}^{[k]}$:

$$\boldsymbol{\Theta}^{[k]} = \boldsymbol{M} + \boldsymbol{C}^{[k]}(\boldsymbol{I} - \phi\boldsymbol{W}^{[k]})^{1/2} = \boldsymbol{M} + (\boldsymbol{\Xi}^{[k]} - \boldsymbol{M})(\boldsymbol{I} - \phi\boldsymbol{W}^{[k]})^{1/2}(\boldsymbol{I} - \phi\boldsymbol{W}^{[k]})^{1/2} = \boldsymbol{M}(\boldsymbol{W}^{[k]}\phi) + \boldsymbol{\Xi}^{[k]}(\boldsymbol{I} - \phi\boldsymbol{W}^{[k]}).$$

Hence if the scaled l_0 is 1-strongly smooth, the main loop of the algorithm consists of only 3 steps:

1) $\boldsymbol{\Xi}^{[k]} \leftarrow \boldsymbol{\Theta}^{[k-1]} - \nabla l_0(\boldsymbol{\Theta}^{[k-1]})$;

2) $(\boldsymbol{M}^{[k]}, \boldsymbol{W}^{[k]}) \leftarrow \mathrm{argmin}_{\boldsymbol{M},\boldsymbol{W}}\, \frac{1}{2} Tr\{(\boldsymbol{\Xi}^{[k]} - \boldsymbol{M})\boldsymbol{W}(\boldsymbol{\Xi}^{[k]} - \boldsymbol{M})^T\} - \frac{n}{2}\log\det(\boldsymbol{W}) + P_M(\boldsymbol{M}) + P_W(\boldsymbol{W})$

3) $\boldsymbol{\Theta}^{[k]} \leftarrow \boldsymbol{\Xi}^{[k]} + \phi(\boldsymbol{M}^{[k]} - \boldsymbol{\Xi}^{[k]})\boldsymbol{W}^{[k]}$

The optimization algorithm has great implementation ease: 1) and 3) involve some basic matrix operations to form the key matrices $\boldsymbol{\Xi}$ and $\boldsymbol{\Theta}$, and 2) just carries out (ordinary) Gaussian learning Correspondingly, GIDL calls multivariate Gaussian estimation *iteratively* to harness a general loss driven dependent learning with guaranteed convergence. (Special case: for $l(\theta, y) = (\theta - y)^2/2$, $\boldsymbol{\Xi}^{[k]} = \boldsymbol{\Theta}^{[k-1]} - (\boldsymbol{\Theta}^{[k-1]} - \boldsymbol{Y}) = \boldsymbol{Y}$ stays fixed, indicating no need to iterate.) The iterative/indirect dependent learning mechanism is innovative to the best of our knowledge, and relies on quadratic distance based learning—any such optimization algorithms can be seamlessly applied in Step 2).

Under high dimensional settings with predictors where p is much larger than n. It is natural and important to assume joint row sparsity on the true coefficient matrix $\boldsymbol{B}^*$ and add penalization on the objective function, which means that all the observations for some variables are zero simultaneously. Let $\boldsymbol{B} = [\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p]^\top$ and for simplicity denote $P_g(\boldsymbol{\beta}; \lambda) := \sum_{i=1}^p P(\|\boldsymbol{\beta}_i\|_2; \lambda)$, where P is a sparse inducing penalty. Then $P_g(\boldsymbol{\beta}; \lambda)$ can be seen as a group penalty for the vectorized coefficient $\boldsymbol{\beta}$. We consider the following optimization problem:

$$\min_{\boldsymbol{\beta}, \boldsymbol{c}, \boldsymbol{W}} \phi^{-1} \bar{l}(\boldsymbol{L}\boldsymbol{\beta} + (\boldsymbol{I} - \phi \boldsymbol{W} \otimes \boldsymbol{I})^{1/2} \boldsymbol{c}; \boldsymbol{y}) + \frac{1}{2} \boldsymbol{c}^T (\boldsymbol{W} \otimes \boldsymbol{I}) \boldsymbol{c} - \frac{n}{2} \log\det \boldsymbol{W} + P_g(\boldsymbol{\beta}; \lambda) + P_W(\boldsymbol{W}; \lambda_W) \tag{1.30}$$

With a bit abuse of notation, we denote the objective function as $F(\boldsymbol{\beta}, \boldsymbol{W}, \boldsymbol{c})$, which can be non-convex. It is not possible to obtain global optimal through computation. Instead, we provide a computational framework and establish the statistical analysis on the fixed points obtained through the algorithm we designed in the next subsection. Following the spirits of traditional iterative computing methods like weighted least squares, we first convert the optimization problem into a Gaussian one through linearization and then apply a weighted type of algorithm to obtain the fixed-point solutions. Given the tth iterate $(\boldsymbol{\beta}^{[k]}, \boldsymbol{W}^{[k]}, \boldsymbol{c}^{[k]})$, let $\boldsymbol{\theta}^{[k]} = \boldsymbol{L}\boldsymbol{\beta}^{[k]} + (\boldsymbol{I} - \phi \boldsymbol{W}^{[k]} \otimes \boldsymbol{I})^\top \boldsymbol{c}^{[k]}$, we construct the following surrogate function

$$\begin{aligned} g(\boldsymbol{\beta}, \boldsymbol{W}, \boldsymbol{c}; \boldsymbol{\beta}^{[k]}, \boldsymbol{W}^{[k]}, \boldsymbol{c}^{[k]}) = \phi^{-1} \bar{l}(\boldsymbol{\theta}^{[k]}; \boldsymbol{y}) + \phi^{-1} \langle \nabla_{\boldsymbol{\theta}} \bar{l}(\boldsymbol{\theta}^{[k]}), \boldsymbol{\theta} - \boldsymbol{\theta}^{[k]} \rangle + \frac{\rho}{2\phi} \|\boldsymbol{\theta} - \boldsymbol{\theta}^{[k]}\|_F^2 \\ - \frac{n}{2} \log\det \boldsymbol{W} + \frac{1}{2} \boldsymbol{c}^\top (\boldsymbol{W} \otimes \boldsymbol{I}) \boldsymbol{c} + P_g(\boldsymbol{\beta}; \lambda) + P_W(\boldsymbol{W}; \lambda_W), \end{aligned} \tag{1.31}$$

where ρ is a step size which can be tuned. Then at time $t+1$ we obtain

$$(\boldsymbol{\beta}^{[k+1]}, \boldsymbol{W}^{[k+1]}, \boldsymbol{c}^{[k+1]}) = \underset{(\boldsymbol{\beta} \in \mathbb{R}^{mp}; \boldsymbol{W} \in \mathcal{S}_{++}^m; \boldsymbol{c} \in \mathbb{R}^{mp})}{\operatorname{argmin}} g(\boldsymbol{\beta}, \boldsymbol{W}, \boldsymbol{c}; \boldsymbol{\beta}^{[k]}, \boldsymbol{W}^{[k]}, \boldsymbol{c}^{[k]}). \tag{1.32}$$

Convergence of our algorithm under the Lipschitz conditions on the gradients of all loss functions are natural: suppose the following Lipschitz conditions hold for all loss l_k,

$$\|\nabla l_k(\boldsymbol{\theta}_1) - \nabla l_k(\boldsymbol{\theta}_2)\|_2 \leq L\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2, \quad \forall \boldsymbol{\theta}_1, \boldsymbol{\theta}_2. \tag{1.33}$$

Then as long as we set $\rho \geq L$ in the above algorithm, the values of objective function F satisfies:

$$F(\boldsymbol{\beta}^{[k+1]}, \boldsymbol{W}^{[k+1]}, \boldsymbol{c}^{[k+1]}) \leq F(\boldsymbol{\beta}^{[k]}, \boldsymbol{W}^{[k]}, \boldsymbol{c}^{[k]}) \tag{1.34}$$

by standard convergence analysis of gradient descent (e.g. see [23]). Instead of directly solving the optimization problem (1.31), we can first apply linearization on $\boldsymbol{\theta}$ to turn the problem into Gaussian and then apply the weighting scheme. Define $\boldsymbol{\xi}^{[k+1]} = \boldsymbol{\theta}^{[k]} - \nabla \bar{l}(\boldsymbol{\theta}^{[k]}; \boldsymbol{y})/\rho$, then the optimization problem (1.31) can be rewritten as

$$\min_{\boldsymbol{\beta}, \boldsymbol{c}, \boldsymbol{W}} \frac{\rho}{2\phi}\|\boldsymbol{L}\boldsymbol{\beta} + (\boldsymbol{I} - \phi \boldsymbol{W} \otimes \boldsymbol{I})^{1/2}\boldsymbol{c} - \boldsymbol{\xi}^{[k+1]}\|_F^2 - \frac{n}{2}\log\det \boldsymbol{W} + \frac{1}{2}\boldsymbol{c}^\top(\boldsymbol{W} \otimes \boldsymbol{I})\boldsymbol{c} + P_g(\boldsymbol{\beta}; \lambda) + P_W(\boldsymbol{W}; \lambda_W). \tag{1.35}$$

In order to obtain the closed form of $\boldsymbol{c}^{[k+1]}$ for Gaussian case, we need to set $\rho = 1$. This can be done by scaling the loss function $\bar{l}$ by the Lipschitz constant L. Given $\boldsymbol{\beta}$ and $\boldsymbol{W}$, we have the explicit form $\boldsymbol{c}^{[k+1]} = (\boldsymbol{I} - \phi \boldsymbol{W} \otimes \boldsymbol{I})^{1/2}(\boldsymbol{\xi}^{[k+1]} - \boldsymbol{L}\boldsymbol{\beta})$. Then plugging the form into the above problem (1.35), we obtain the following optimization problem for $\boldsymbol{\beta}^{[k+1]}$ and $\boldsymbol{W}^{[k+1]}$

$$\min_{\boldsymbol{\beta}, \boldsymbol{W}} \|\boldsymbol{W}^{\frac{1}{2}}\,\mathrm{vec}^{-1}(\boldsymbol{\xi}^{[k+1]} - \boldsymbol{L}\boldsymbol{\beta})\|_F^2 - \frac{n}{2}\log\det \boldsymbol{W} + P_g(\boldsymbol{\beta}; \lambda) + P_W(\boldsymbol{W}; \lambda_W). \tag{1.36}$$

Now the optimization becomes a natural extension of traditional weighted least squares applied in high dimensional problems when given $\boldsymbol{W}$. In this case, efficient gradient descent type algorithm or coordinate descent algorithm can be applied to solve the above optimization for $\boldsymbol{\beta}$ just like the group lasso. While when $\boldsymbol{\beta}$ is given, we are back to a Gaussian graph learning problem, like graphical lasso when the ℓ_1 penalty is used for $\boldsymbol{W}$. Thus a block coordinate descent algorithm can be applied for solving $\boldsymbol{\beta}^{[k+1]}$ and $\boldsymbol{W}^{[k+1]}$. In the inner loop of the BCD algorithm, it can be terminated as long as the function value of the surrogate function decreases. Then $\boldsymbol{\theta}^{[k+1]}$ can also be obtained without explicitly calculating $\boldsymbol{c}^{[k+1]}$:

$$\boldsymbol{\theta}^{[k+1]} = \boldsymbol{L}\boldsymbol{\beta}^{[k+1]} + (\boldsymbol{I} - \phi \boldsymbol{W} \otimes \boldsymbol{I}^{[k+1]})^{\frac{1}{2}}\boldsymbol{c}^{[k+1]} = \phi \boldsymbol{L}\boldsymbol{\beta}^{[k+1]} + (\boldsymbol{I} - \phi \boldsymbol{W} \otimes \boldsymbol{I}^{[k+1]})\boldsymbol{\xi}^{[k+1]}. \tag{1.37}$$

To conclude, the full algorithm is summarized as follows:

Algorithm 1: Multivariate Learning with Dependency

Data: $\boldsymbol{Y} \in \mathbb{R}^{n\times m}$, $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, $\boldsymbol{y} = \mathrm{vec}\,(\boldsymbol{Y})$, $\boldsymbol{L} = \boldsymbol{I}_{m\times m} \otimes \boldsymbol{X}$, $\bar{l}$ such that $\nabla\bar{l}$ is Lipschitz-1 continuous, initial value $\boldsymbol{\xi}^{[0]}$, $\boldsymbol{\beta}^{[0]}$ and $\boldsymbol{W}^{[0]} \in \mathcal{S}^m_{++}$, $\phi \le \|\boldsymbol{W}^{[0]}\|_2^{-1}$.

Initialize iteration number $k \to 0$;

while *not convergence* **do**

  $k \to k+1$;

  $\boldsymbol{\theta}^{[k-1]} = \phi\boldsymbol{L}\boldsymbol{\beta}^{[k-1]} + (\boldsymbol{I} - \phi\boldsymbol{W}\otimes\boldsymbol{I})\boldsymbol{\xi}^{[k-1]}$, $\boldsymbol{\xi}^{[k]} = \boldsymbol{\theta}^{[k-1]} - \nabla\bar{l}(\boldsymbol{\theta}^{[k]}; \boldsymbol{y})$;

  $t \to 0$, $\boldsymbol{\beta}^{(0)} = \boldsymbol{\beta}^{[k-1]}, \boldsymbol{W}^{(0)} = \boldsymbol{W}^{[k-1]}$;

  while $g(\boldsymbol{\beta}^{(t)}, \boldsymbol{W}^{(t)}, \boldsymbol{c}^{(t)}; \boldsymbol{\beta}^{k-1}, \boldsymbol{W}^{k-1}) > g(\boldsymbol{\beta}^{(t-1)}, \boldsymbol{W}^{(t-1)}, \boldsymbol{c}^{(t-1)})$ *if existing* **do**

    $t \to t+1$;

    $\boldsymbol{W}^{(t)} \to \mathrm{argmin}_{\boldsymbol{W}}\, \|\boldsymbol{W}^{\frac{1}{2}}\,\mathrm{vec}^{-1}(\boldsymbol{\xi}^{[k+1]} - \boldsymbol{L}\boldsymbol{\beta}^{(t-1)})\|_F^2 - \frac{n}{2}\log\det\boldsymbol{W} + P_W(\boldsymbol{W}; \lambda_W)$;

    $\boldsymbol{\beta}^{(t)} \to \mathrm{argmin}_{\beta}\, \|(\boldsymbol{W}\otimes\boldsymbol{I}^{(t)\frac{1}{2}})(\boldsymbol{\xi}^{[k+1]} - \boldsymbol{L}\boldsymbol{\beta})\|_F^2 + P_g(\boldsymbol{\beta}; \lambda_W)$

  end

  $\boldsymbol{\beta}^{[k]} \to \boldsymbol{\beta}^{(t)}, \boldsymbol{W}^{[k]} \to \boldsymbol{W}^{(t)}$;

end

Result: Output the final estimate $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}^{[k]}$, $\hat{\boldsymbol{W}} = \boldsymbol{W}^{[k]}$.

Note that for clustered data case, there are two main differences: first, since $\boldsymbol{W}$ capture the dependency among rows, we need change W- optimization into

$$\boldsymbol{W}^{(t)} \to \underset{\boldsymbol{W}}{\mathrm{argmin}}\, \|\,\mathrm{vec}^{-1}(\boldsymbol{\xi}^{[k+1]} - \boldsymbol{L}\boldsymbol{\beta}^{(t-1)})\boldsymbol{W}^{\frac{1}{2}}\|_F^2 - \frac{n}{2}\log\det\boldsymbol{W} + P_W(\boldsymbol{W}; \lambda_W). \tag{1.38}$$

Besides, there should be a constraint on $\boldsymbol{\beta}$ optimization such that $\boldsymbol{\beta}^{(t)}$ should be projected in to the space $\{\boldsymbol{\beta} : \exists\, \boldsymbol{\beta}_0 \in \mathbb{R}^p,\, \boldsymbol{\beta} = \boldsymbol{I}_{m\times 1} \otimes \boldsymbol{\beta}_0\}$.

1.6 Experiments

1.6.1 Simulation Studies

In this section, we provide some simulation results to support our main theorems. Consider the Gaussian graph model $\boldsymbol{Y} = \boldsymbol{M} + \boldsymbol{E}$ where we have $\boldsymbol{M} = \boldsymbol{0}$ and the rows of $\boldsymbol{E}$ are independent distributed as $\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}^*)$ with $\boldsymbol{W}^* = \boldsymbol{\Sigma}^{*-1}$. Suppose $\boldsymbol{W}^*$ is sparse with index $\mathcal{J}^*$ on the upper diagonal parts. The main goal is to perform statistical inference on the nonzero part of true dependency $\boldsymbol{W}^*_{\mathcal{J}^*}$. Given the generated data $\boldsymbol{Y}$, the steps to perform statistical inference of the true dependency matrix $\boldsymbol{W}^*$ are as follows: First, an estimator $\hat{\boldsymbol{W}}$ with nonzero index $\hat{\mathcal{J}}$ can be obtained through our GIDL framework with an adaptive lasso penalty, where we assume $\mathcal{J}^* \subset \hat{\mathcal{J}}$ by choosing proper λ. Then, the sample covariance of $\mathrm{vec}\,[-\sqrt{n}\{\hat{\boldsymbol{\Sigma}}^n_{\mathcal{J}^*_W} - \boldsymbol{\Sigma}^*_{\mathcal{J}^*_W}\}])$ can be calculated

Table 1.1: Converge probability of the confidence interval constructed based on the Hotelling's t-Squared statistics derived from GIDL

%	Case 1	Case 2	Case 3	Case 4
$C = 0.90$	89.00	92.00	89.00	88.67
$C = 0.95$	94.67	96.00	93.00	93.67
$C = 0.99$	97.33	98.67	97.67	97.00

as a plug-in estimate of its population correspondence. Finally, based on the asymptotic normal distribution on the index $\mathcal{J}^*$, we can construct a Hotelling's t-Squared type of statistics to perform the statistical inference. We perform simulations to demonstrate that the above steps can deliver a reasonable result for statistical inference. To generate the true sparse dependency matrix, we set ρ as the percentage of non-zero off-diagonal elements, randomly generate nonzero index for the off diagonal part based on probability ρ and then let $W_{ij} \sim \mathcal{N}(0,1)$ for $(i,j) \in \mathcal{J}^*$. We show the coverage probability of the constructed confidence interval under the following cases of simulation settings at confidence level $C = 0.99, 0.95, 0.90$ by performing Monte Carlo simulations for 300 times:

1. (Case 1:) $n = 100, m = 10, \rho = 0.1$;
2. (Case 2:) $n = 500, m = 10, \rho = 0.1$;
3. (Case 3:) $n = 500, m = 10, \rho = 0.5$;
4. (Case 4:) $n = 500, m = 20, \rho = 0.1$.

The simulation results are shown in table 1.1. From the table, we can see that as long as the sample size n is large enough, the coverage probabilities for all confidence intervals are almost match the corresponding confidence levels. This result implies that our method can perform a valid statistical inference for the nonzero part of the true dependency $\boldsymbol{W}^*_{\mathcal{J}^*}$.

We also perform a comparison between squared loss and Tukey's loss when applying main theorem to calculate confidence intervals and the presence of outliers is a problem. The Tukey's loss is a well-known robust loss which is defined by a ψ-function: $l(\theta, y) = \int_0^{|\theta - y|} \psi(t)\mathrm{d}t$ [5], where Tukey's bisquare is

$$\psi(t) = \begin{cases} t\left[1 - \left(\frac{t}{c}\right)^2\right]^2, & |t| \le c \\ 0, & |t| > c \end{cases} \tag{1.39}$$

Table 1.2: Converge probability and length of the confidence interval comparisons between squared and Tukey's loss by GIDL

	Case 1		Case 2		Case 3		
	Coverage %	Length	Coverage %	Length	Coverage %	Length	C
Gaussian	100	14.99	100	17.35	100	20.06	0.90
Tukey	82.67	0.88	88	0.43	81	0.44	
Gaussian	100	16.47	100	19.65	100	22.42	0.95
Tukey	89.33	1.09	93.67	0.52	89.6667	0.53	
Gaussian	100	19.31	100	23.07	100	25.82	0.99
Tukey	97	1.44	98.33	0.67	97	0.670	

and $c = 4.685\sigma$ is recommended. To generate synthetic data, we use the model $\boldsymbol{Y} = \boldsymbol{M} + \boldsymbol{\Psi} + \boldsymbol{E}$ where the settings of $\boldsymbol{M}$ and $\boldsymbol{E}$ are still the same with the above simulation. In addition, $\boldsymbol{\Psi} \in \mathbb{R}^{n\times m}$ is a sparse matrix that incorporates the outlying effects in $\boldsymbol{Y}$ as its components have 0.95 probability to be zero and 0.05 probability to be 20. In order to compare the length of calculated confidence intervals, we only set one dependent edge on the off-diagonal parts of $\boldsymbol{W}^*$: $|\mathcal{J}^*| = 1$ where its location is random assigned as before. We use the same approach as the above simulations to calculate both the confidence intervals and length of the confidence interval of $\boldsymbol{W}_{\mathcal{J}^*}$ under both squared loss and Tukey's loss. Three simulation cases are considered: Case 1: $n = 100, m = 10$; Case 2: $n = 500, m = 10$; Case 3: $n = 500, m = 20$ and the average coverage probabilities and lengths at confidence level $C = 0.90, 0.95, 0.99$ are reported for 300 Monte Carlo simulations. The simulation result is listed in table 1.2. From the table, some phenomenon can be observed: first, the confidence intervals obtained via squared loss are not informative, since the length of the interval is to large so that the covering probabilities are always 100%; moreover, the coverage probabilities of the intervals obtained via Tukey's loss is less than what we desire, which is reasonable since our model is misspecified for the true data generating process; finally, when n is large enough, we can see the coverage probabilities for the Tukey's loss are approaching to the desired ones. The above conclusions strike us to apply robust loss functions to make statistical inference when outlying effect must be taken into account.

1.6.2 Real Data

We explore the NASDAQ-100 stock market index data from 01/03/2011 to 12/31/2011. The differences between closed price of two consecutive days is the main target, where we want to detect

and make inference on the dependency structures of these differences between the top 98 largest non-financial companies listed on the Nasdaq stock market. Since the differences could be extremely large in some days, it is more reasonable to adopt some robust loss functions when apply our GIDL framework than traditional squared loss. Here we provide some exploration to demonstrate that robust loss is a more reasonable choice by making inference on the first detected dependent edges through both squared loss and Tukey's loss. By tuning the parameter λ of the adaptive lasso penalty, we make sure that only one edge can be detected and then calculate the confidence interval of the weigh of the edge through our main theorem under both kinds of losses. For Gaussian loss, the detected edge is the link between 'FAST' and 'ALXN' with weight -5.47. However, the 95% confidence interval of this edge we calculate is $[-64.8, 53.9]$, which has a unreasonable magnitude and incorrectly contains zero. The two problems imply that assuming Gaussian distribution is even useless in making inference on the dependent edge. Instead, when we apply the Tukey's loss, the detected edge is the link between 'LLTC' and 'TXN' with weight -5.48. The corresponding 95% confidence interval is $[-6.2, -0.6]$, which doesn't contain zero and enjoy a more reasonable length. Therefore, We have reasons to believe that for GIDL, robust loss functions are essential for reasonable inference through our main theorems when the presence of outliers is a problem.

1.7 Summary

In this section, we explore the statistical inference problem for GIDL where we estimate the dependency matrix or incorporate the structure information into estimating the first order effect with the allowance of customizing the marginal losses. Some asymptotic theories are provided to assistant the inference. However, there are still some problems to be answered in the future. It is a great of interest to make inference for high dimensional estimated dependencies and coefficients obtained through our framework. Whether incorporating the structure information can indeed lead to more efficient estimation as the traditional asymptotic cases should be explained in the future.

1.8 Outlines of Proofs

1.8.1 Proof of Theorem 1.3.1

The first Step: Solving $\boldsymbol{C}$ based on expansion of gradient of the loss:

By taking derivatives of the objective function with respective to $\boldsymbol{C}$, we have $\nabla \bar{l}(\boldsymbol{\Theta})\boldsymbol{Z}^{1/2} + \phi \boldsymbol{C}\boldsymbol{W} = \boldsymbol{0}$. Note that the we also have $\nabla \bar{l}(\boldsymbol{\Theta}) = \boldsymbol{\nabla} \bar{l}(\boldsymbol{M}) + \boldsymbol{C}\boldsymbol{Z}^{1/2} + \boldsymbol{\Delta}$. Then $\boldsymbol{C}$ can be solved as

$$\boldsymbol{C} = -(\nabla \bar{l}(\boldsymbol{M}) + \boldsymbol{\Delta})\boldsymbol{Z}^{\frac{1}{2}}(\boldsymbol{Z} + \phi \boldsymbol{W})^{-1} = -(\nabla \bar{l}(\boldsymbol{M}) + \boldsymbol{\Delta})\boldsymbol{Z}^{\frac{1}{2}}. \tag{1.40}$$

The second Step: Plugging $\boldsymbol{C}$ back into the loss, obtaining an approximation Gaussian representation of the loss based through direct expansion:

By plugging the expression of $\boldsymbol{C}$ into the original optimization problem, we have

$$\begin{aligned}
&\bar{l}\,(\boldsymbol{\Theta}) + \frac{\phi}{2}\operatorname{Tr}\left\{\boldsymbol{C}\boldsymbol{W}\boldsymbol{C}^T\right\} - \bar{l}(\boldsymbol{M}) \\
&= \langle \nabla \bar{l}(\boldsymbol{M}) + \frac{1}{2}\boldsymbol{C}\boldsymbol{Z}^{\frac{1}{2}}, \boldsymbol{C}\boldsymbol{Z}^{\frac{1}{2}} \rangle + \frac{\phi}{2}\operatorname{Tr}\left\{\boldsymbol{C}\boldsymbol{W}\boldsymbol{C}^T\right\} + \boldsymbol{\Gamma}(\boldsymbol{\Theta}) \\
&= \langle \nabla \bar{l}(\boldsymbol{M}) + \boldsymbol{\Delta}, \boldsymbol{C}\boldsymbol{Z}^{\frac{1}{2}} \rangle + \frac{1}{2}\langle \boldsymbol{C}\boldsymbol{Z}^{\frac{1}{2}}, \boldsymbol{C}\boldsymbol{Z}^{\frac{1}{2}} \rangle + \langle -\boldsymbol{\Delta}, \boldsymbol{C}\boldsymbol{Z}^{\frac{1}{2}} \rangle + \frac{\phi}{2}\operatorname{Tr}\left\{\boldsymbol{C}\boldsymbol{W}\boldsymbol{C}^T\right\} + \delta(\boldsymbol{\Theta}) \\
&= \langle \nabla \bar{l}(\boldsymbol{M}) + \boldsymbol{\Delta}, \boldsymbol{C}\boldsymbol{Z}^{\frac{1}{2}} \rangle + \frac{1}{2}\operatorname{Tr}\left\{\boldsymbol{C}(\boldsymbol{Z} + \phi\boldsymbol{W})\boldsymbol{C}^T\right\} + \langle -\boldsymbol{\Delta}, \boldsymbol{C}\boldsymbol{Z}^{\frac{1}{2}} \rangle + \delta(\boldsymbol{\Theta}) \\
&= -\frac{1}{2}\operatorname{Tr}\left\{(\nabla \bar{l}(\boldsymbol{M}) + \boldsymbol{\Delta})\boldsymbol{Z}(\nabla \bar{l}(\boldsymbol{M}) + \boldsymbol{\Delta})^T\right\} + \langle \boldsymbol{\Delta}, (\nabla \bar{l}(\boldsymbol{M}) + \boldsymbol{\Delta})\boldsymbol{Z} \rangle + \delta(\boldsymbol{\Theta}) \\
&= -\frac{1}{2}\operatorname{Tr}\left\{\nabla \bar{l}(\boldsymbol{M})\boldsymbol{Z}\nabla \bar{l}(\boldsymbol{M})^T\right\} + \delta(\boldsymbol{\Theta}).
\end{aligned}$$

This results in

$$\begin{aligned}
&\phi^{-1}\bar{l}\,(\boldsymbol{\Theta}) + \frac{1}{2}\operatorname{Tr}\left\{\boldsymbol{C}\boldsymbol{W}\boldsymbol{C}^T\right\} - \phi^{-1}\bar{l}(\boldsymbol{M}) \\
&= \frac{1}{2}\operatorname{Tr}\left\{\nabla \bar{l}(\boldsymbol{M})\boldsymbol{W}\nabla \bar{l}(\boldsymbol{M})^T\right\} - \frac{1}{2\phi}\|\nabla \bar{l}(\boldsymbol{M})\|_F^2 + \phi^{-1}\delta(\boldsymbol{\Theta}).
\end{aligned}$$

Denote the objective function as $F(\boldsymbol{W}; \boldsymbol{C})$ and $F(\boldsymbol{W})$ as the abbreviation for $F(\boldsymbol{W}; \boldsymbol{C}(\boldsymbol{W}))$. Then we have an approximation Gaussian representation of the loss:

$$F(\boldsymbol{W}) = \frac{1}{2}\operatorname{Tr}\left\{\nabla \bar{l}(\boldsymbol{M})\boldsymbol{W}\nabla \bar{l}(\boldsymbol{M})^T\right\} - \frac{1}{2\phi}\|\nabla \bar{l}(\boldsymbol{M})\|_F^2 - \frac{n}{2}\log\det \boldsymbol{W} + P_W\left(\boldsymbol{W}; \lambda_W\right) + \phi^{-1}\delta(\boldsymbol{\Theta}).$$

We need to investigate the terms containing $\boldsymbol{\Delta}(\boldsymbol{\Theta})$ and $\delta(\boldsymbol{\Theta})$ such that the effect of them are always higher order terms in the following analysis. Based on condition 5, we have two useful results:

$$\|\boldsymbol{\Delta}^*\|_F = o_p(1)$$

and

$$\|\frac{\mathrm{d}\delta(\boldsymbol{\Theta})}{\mathrm{d}\boldsymbol{W}}\|_F = o_p(n^{1/2}).$$

The third Step: showing rate consistency of $\widehat{\boldsymbol{W}}$ by drawing a circle: Next we consider to show the rate consistency of $\widehat{\boldsymbol{W}}$. In particular, we want to show that for any given $\epsilon > 0$, there exists a large

constant C_1 such that

$$P\{\inf_{\|\boldsymbol{U}\|_F=C_1} F(\boldsymbol{W}^* + d_n\boldsymbol{U}) - F(\boldsymbol{W}^*) > 0\} \geq 1-\epsilon,$$

where $\boldsymbol{U}$ is invertible and $d_n = n^{-1/2} + a_n$. This implies that with probability at least $1-\epsilon$, there exists a local minimum in the ball $\{\boldsymbol{W} : \|\boldsymbol{W} - \boldsymbol{W}^*\|_F \leq d_n C_1\}$. By Taylor's expansion, there exists a constant $0 < t < 1$, such that

$$\begin{aligned}
&F(\boldsymbol{W}^* + d_N\boldsymbol{U}) - F(\boldsymbol{W}^*) \\
&= \frac{1}{2}\langle d_n\boldsymbol{U}, \nabla\bar{l}(\boldsymbol{M})^T\nabla\bar{l}(\boldsymbol{M})\rangle - \frac{n}{2}(\log\det(\boldsymbol{W}^* + d_n\boldsymbol{U}) - \log\det\boldsymbol{W}^*) + \phi^{-1}[\delta(\boldsymbol{\Theta}) - \delta(\boldsymbol{\Theta}^*)] \\
&+ P_W(\boldsymbol{W}^* + d_n\boldsymbol{U}; \lambda_W) - P_W(\boldsymbol{W}^*; \lambda_W) \\
&\geq \frac{n}{2}\langle d_n\boldsymbol{U}, \hat{\boldsymbol{\Sigma}}^n - \boldsymbol{W}^{*-1}\rangle + \frac{n}{2}\langle d_n\boldsymbol{U}, \frac{1}{2}d_n\boldsymbol{W}^{*-1}\boldsymbol{U}\boldsymbol{W}^{*-1}\rangle - \frac{n}{2}\boldsymbol{R}(d_n\boldsymbol{U}) + \phi^{-1}[\delta(\boldsymbol{\Theta}) - \delta(\boldsymbol{\Theta}^*)] \\
&+ \sum_{(i,j)\in\mathcal{J}^*_{\mathcal{W}}} d_N U_{ij} P'_W(|W^*_{ij}|; \lambda_W)\mathrm{sgn}(|W^*_{ij}|) + \sum_{(i,j)\in\mathcal{J}^*_{\mathcal{W}}} \frac{1}{2}d_N^2 P''_W(W^*_{ij}; \lambda_W)U^2_{ij}(1+o(1)) \\
&- \frac{1}{2}\langle d_n\boldsymbol{U}, 2\boldsymbol{\Delta}^{*T}\nabla\bar{l}(\boldsymbol{M}) + \boldsymbol{\Delta}^{*T}\boldsymbol{\Delta}^*\rangle \\
&\equiv T_1 + T_2 + T_3 + T_4 + T_5 + T_6 + T_7,
\end{aligned}$$

where $\boldsymbol{R}(d_n\boldsymbol{U})$ is the reminder for the log det function with $\boldsymbol{R}(\tilde{\boldsymbol{U}}) = \mathrm{Tr}(\sum_{j=3}^{\infty}(-1)^{j-1}(\tilde{\boldsymbol{U}}\boldsymbol{W}^{*-1})^j/j)$ for any $\tilde{\boldsymbol{U}}$ is based on the identity

$$\log\det(\boldsymbol{W}^* + d_n\boldsymbol{U}) - \log\det\boldsymbol{W}^* = \log\det(\boldsymbol{I} + d_n\boldsymbol{U}\boldsymbol{W}^{*-1}) = \mathrm{Tr}\log(\boldsymbol{I} + d_n\boldsymbol{U}\boldsymbol{W}^{*-1})$$

when $\boldsymbol{I} + d_n\boldsymbol{U}\boldsymbol{W}^{*-1}$ is positive-definite.

For the term T_1, $\mathbb{E}(\boldsymbol{W}^{*-1}) = (\nabla\bar{l}(\boldsymbol{M}) + \boldsymbol{\Delta}^*)^T(\nabla\bar{l}(\boldsymbol{M}) + \boldsymbol{\Delta}^*)/n$, it holds that $|T_1| = O_p(C_1 d_n\sqrt{n})$ by the central limit theorem. For the term T_2, by the positive-definiteness of $\boldsymbol{W}^{*-1} \otimes \boldsymbol{W}^{*-1}$, we have $T_2 \geq C_1^2 n d_n^2/(4\gamma^2_{max})$; moveover, since we have

$$|R(d_n\boldsymbol{U})| \leq \sum_{j=3}^{\infty}(-1)^{j-1}d_n^j\|\boldsymbol{U}\|_F^j\|\boldsymbol{W}^{*-1}\|_F^j/j \leq d_n^3\|\boldsymbol{U}\|_F^3\|\boldsymbol{W}^{*-1}\|_F^3/3,$$

this implies that $|T_3| = O_p(C_1^3 n d_n^3)$. By the mean value theorem, there exists a $\tilde{\boldsymbol{\Theta}}$ between $\boldsymbol{\Theta}$ and $\boldsymbol{\Theta}^*$ such that $|\delta(\boldsymbol{\Theta}) - \delta(\boldsymbol{\Theta}^*)| = \langle \mathrm{d}\delta(\tilde{\boldsymbol{\Theta}})/\mathrm{d}\boldsymbol{W}, d_N\boldsymbol{U}\rangle$, so we have $|T_4| = o_p(C_1\sqrt{n}d_n)$ based on $\|\mathrm{d}\delta(\boldsymbol{\Theta})/\mathrm{d}\boldsymbol{W}\|_F = o_p(1)$. For the terms with respect to the penalties, we have $T_5 \geq -\sqrt{J^*_W}nd_n a_n C_1$ and the term T_6 is bounded below by $-nd_N^2C_1^2\max_{(j,k)\in\mathcal{J}^*}P''_W(W^*_{jk};\lambda_W)$. Finally, the term T_7 is at the order of $o_p(C_1 d_n\sqrt{n})$ based on $\|\nabla\bar{l}(\boldsymbol{M})\|_F = O_p(\sqrt{n})$, $\|\boldsymbol{\Delta}^*\|_F = o_p(1)$. Therefore, by choosing

a large enough C_1, the summation from T_1 to T_7 is dominated by the positiveness of T_2. When on the boundary, all values are greater than $F(\boldsymbol{W}^*)$, then there must be a local minimizer within the ball, denoted as $\widehat{\boldsymbol{W}}$, so $\|\widehat{\boldsymbol{W}}-\boldsymbol{W}^*\|_F = O_p(n^{-1/2}+a_n)$ holds by choosing $\widehat{\boldsymbol{W}}$ such that $\|\widehat{\boldsymbol{W}}-\boldsymbol{W}^*\|_F$ is the smallest among all the roots within the region.

To prove the second part, for any $\boldsymbol{W}$ such that $\|\boldsymbol{W}-\boldsymbol{W}^*\|_F = O_p(n^{-1/2})$ and thus $\max_{j\in\mathcal{J}_{\mathcal{W}}{}^{*c}}|W_j| = O_p(n^{-1/2})$. Then we have

$$\begin{aligned}
&\frac{\partial F(\boldsymbol{W})}{\partial \boldsymbol{W}} = \frac{\nabla\bar{l}(\boldsymbol{M})^T\nabla\bar{l}(\boldsymbol{M})}{2} - \frac{n}{2}\boldsymbol{W}^{-1} + P'_W\left(|\boldsymbol{W}|;\lambda_W\right)\operatorname{sgn}\left(\boldsymbol{W}\right) + \phi^{-1}\frac{\mathrm{d}\delta(\boldsymbol{\Theta})}{\mathrm{d}\boldsymbol{W}}\\
&= \frac{n}{2}\{\hat{\boldsymbol{\Sigma}}^n - \boldsymbol{W}^{*-1}\} + (-\frac{n}{2}\boldsymbol{W}^{-1} + \frac{n}{2}\boldsymbol{W}^{*-1}) + \phi^{-1}\frac{\mathrm{d}\delta(\boldsymbol{\Theta})}{\mathrm{d}\boldsymbol{W}} - \frac{1}{2}(2\boldsymbol{\Delta}^{*T}\nabla\bar{l}(\boldsymbol{M}) + \boldsymbol{\Delta}^{*T}\boldsymbol{\Delta}^*)\\
&+ P'_W\left(|\boldsymbol{W}|;\lambda_W\right)\operatorname{sgn}\left(\boldsymbol{W}\right)\\
&= T'_1 + T'_2 + T'_3 + T'_4 + T'_5.
\end{aligned} \tag{1.41}$$

Since $\|\boldsymbol{W}-\boldsymbol{W}^*\|_F = O_p(n^{-1/2})$, the Frobenius norm of the first and second terms can be bounded by $O_p(\sqrt{n})$. For the term T'_3, we have $\|\mathrm{d}\delta(\tilde{\boldsymbol{\Theta}})/\mathrm{d}\boldsymbol{W}\|_F = o_p(1)$. Based on $\|\nabla\bar{l}(\boldsymbol{M})\|_F = O_p(\sqrt{n})$, $\|\boldsymbol{\Delta}^*\|_F = o_p(1)$, we have $\|T'_4\|_F = o_p(\sqrt{n})$. Finally, for the last term, since $\lim_{n\to\infty}\lim_{t\to 0+}P'_W(t;\lambda_W)/\lambda_W > 0$ and $\lim_{n\to\infty}\lambda_W/\sqrt{n}\to\infty$, all its elements can dominate the Frobenius norm of the other terms. Then for $j\in\mathcal{J}_{\mathcal{W}}{}^{*c}$, we have $\frac{\partial F(\boldsymbol{W})}{\partial \boldsymbol{W}_j} < 0$ for $0 < W_j$ and $\frac{\partial F(\boldsymbol{W})}{\partial \boldsymbol{W}_j} > 0$ for $W_j > 0$. Note that when $n\to\infty$, $a_n\to 0$. Since our defined $\widehat{\boldsymbol{W}}$ is a root n consistent estimator, it satisfies $\widehat{\boldsymbol{W}}_{\mathcal{J}_{\mathcal{W}}{}^{*c}} = 0$.

For the third part, by the above Taylor expansion (2.80), the $\sqrt{n}$ consistency of $\widehat{\boldsymbol{W}}$, we have

$$\begin{aligned}
\mathbf{0} = &\frac{n}{2}\operatorname{vec}\{\hat{\boldsymbol{\Sigma}}^n - \boldsymbol{W}^{*-1}\} + \frac{n}{2}\boldsymbol{W}^{*-1}\otimes\boldsymbol{W}^{*-1}\operatorname{vec}(\widehat{\boldsymbol{W}}-\boldsymbol{W}^*)(1+o_p(1))\\
&+ \operatorname{vec}\{P'_W\left(|\boldsymbol{W}^*|;\lambda_W\right)\operatorname{sgn}\left(\boldsymbol{W}^*\right)\} + \tilde{\boldsymbol{R}},
\end{aligned}$$

where $\tilde{\boldsymbol{R}}$ is a reminder term satisfying $\|\tilde{\boldsymbol{R}}\|_F = o_p(\sqrt{n})$. Therefore, by only considering the submatrix on $\mathcal{J}_{\boldsymbol{W}^*}$ and based on the conditions for penalties, we have

$$\sqrt{n}(\operatorname{vec}\widehat{\boldsymbol{W}}_{\mathcal{J}_W^*} - \operatorname{vec}\boldsymbol{W}^*_{\mathcal{J}_W^*}) = \{\boldsymbol{\Sigma}^*_{\mathcal{J}_W^*}\otimes\boldsymbol{\Sigma}^*_{\mathcal{J}_W^*}\}^{-1}\operatorname{vec}\left[-\sqrt{n}\{\hat{\boldsymbol{\Sigma}}_{\mathcal{J}_W^*} - \boldsymbol{\Sigma}^*_{\mathcal{J}_W^*}\}\right](1+o_p(1)),$$

where recall $\boldsymbol{\Sigma}^* = \boldsymbol{W}^{*-1}$. Finally, the conclusion follows by Slutsky's theorem.

1.8.2 Proof of Theorem 1.3.2

For simplicity, when the mean matrix $\boldsymbol{M}$ is known, we can perform a location transformation such that $\boldsymbol{Y}\leftarrow\boldsymbol{Y}-\boldsymbol{M}$. Then based on the assumption that $\boldsymbol{M} = \mathbf{1}\boldsymbol{\alpha}^T$, we can scale each loss

function $l_k(\cdot)$ by $l_k''(\alpha_k) = 1/4$. Based on the scaling, we have $\boldsymbol{D} = \boldsymbol{I}$. Apply the Theorem 2 in [23], there is a $\boldsymbol{W}_{gidl}$, such that

$$\boldsymbol{W}^{gidl^{-1}} = \frac{(\nabla l(\boldsymbol{M}) + \boldsymbol{\Delta}(\boldsymbol{M}, \boldsymbol{W}, \boldsymbol{C}))^T (\nabla l(\boldsymbol{M}) + \boldsymbol{\Delta}(\boldsymbol{M}, \boldsymbol{W}, \boldsymbol{C}))}{n}. \tag{1.42}$$

Note that due to relocation and scaling, $-\nabla l(\boldsymbol{M}) = 4\boldsymbol{Y}$, thus as the regularity condition for high order term satisfied, we have

$$\boldsymbol{W}^{gidl-1} \to 16\frac{\boldsymbol{Y}^T\boldsymbol{Y}}{n}. \tag{1.43}$$

Denote $\hat{\boldsymbol{W}} = (\boldsymbol{Y}^T\boldsymbol{Y}/n)^{-1}$, we now prove as $n \to \infty$, the stationary condition for pseudo-likelihood can be satisfied when applying the graph structure decided by $\hat{\boldsymbol{W}}$. Note that $\hat{\boldsymbol{W}}_{ii} = 1$ for $i = 1, ..., m$. Based on inverse of block matrix (Schur component), we have

$$\hat{\boldsymbol{W}}[\{-s\}, s]\hat{\boldsymbol{W}}[s, s]^{-1} = -(\boldsymbol{Y}[, \{-s\}]^T\boldsymbol{Y}[, \{-s\}])^{-1}\boldsymbol{Y}[, \{-s\}]^T\boldsymbol{Y}[, s]. \tag{1.44}$$

Then if we consider the graph structure defined by $-\hat{\boldsymbol{W}}$ that is the same with $\hat{\boldsymbol{W}}$, we have the property

$$\begin{aligned} -\frac{1}{n}\sum_{i=1}^{n} Y_{ik}\sum_{t\neq s}\hat{\boldsymbol{W}}_{st}Y_{it} &= -\frac{1}{n}\sum_{i=1}^{n} Y_{ik}\sum_{t\neq s}\hat{\boldsymbol{W}}_{st}Y_{it} \\ &= \frac{1}{n}\boldsymbol{Y}[, k]^T\boldsymbol{Y}[, \{-s\}](\boldsymbol{Y}[, \{-s\}]^T\boldsymbol{Y}[, \{-s\}])^{-1}\boldsymbol{Y}[, \{-s\}]^T\boldsymbol{Y}[, s] \\ &\overset{(i)}{=} \frac{1}{n}\boldsymbol{Y}[, k]^T\boldsymbol{Y}[, s], \end{aligned} \tag{1.45}$$

where (i) is because $\boldsymbol{Y}[, \{-s\}](\boldsymbol{Y}[, \{-s\}]^T\boldsymbol{Y}[, \{-s\}]^{-1}\boldsymbol{Y}[, \{-s\}]^T$ is a projection matrix that can map $\boldsymbol{Y}[, k]$ to itself. Moreover, if we apply the graph structure for $-\hat{\boldsymbol{W}}$ into the gradient for pseudo-likelihood, we have

$$\begin{aligned} -\frac{\partial l}{\partial W_{sk}} &= -\frac{1}{n}\sum_{i=1}^{n}\left[Y_{is}Y_{ik} - Y_{ik}\tanh(\sum_{t\neq s}\hat{\boldsymbol{W}}_{st}Y_{it})\right] \\ &\overset{(i)}{=} -\frac{1}{n}\sum_{i=1}^{n}\left[Y_{is}Y_{ik} - \tanh(\sum_{t\neq s}Y_{ik}\hat{\boldsymbol{W}}_{st}Y_{it})\right] \\ &= -\frac{1}{n}\boldsymbol{Y}[, s]^T\boldsymbol{Y}[k,] + \frac{1}{n}\sum_{i=1}^{n}\tanh(Y_{ik}\boldsymbol{Y}_{i\{-s\}}(\boldsymbol{Y}_{\{-s\}}^T\boldsymbol{Y}_{\{-s\}})^{-1}\boldsymbol{Y}_{\{-s\}}^T\boldsymbol{Y}_s) \\ &\overset{(ii)}{=} \frac{1}{n}\sum_{i=1}^{n}(tY_{ik}\boldsymbol{Y}_{i\{-s\}}(\boldsymbol{Y}_{\{-s\}}^T\boldsymbol{Y}_{\{-s\}})^{-1}\boldsymbol{Y}_{\{-s\}}^T\boldsymbol{Y}_s)^3 \\ &\to 0. \end{aligned} \tag{1.46}$$

where (i) is because y_{ik} can only take $+1$ or -1 and in (ii) we apply the above identity.

1.8.3 Proof of Theorem 1.4.1

Let $\boldsymbol{W}_T = \boldsymbol{W} \otimes \boldsymbol{I}$ and $\overline{\boldsymbol{W}}_T = \overline{\boldsymbol{W}} \otimes \boldsymbol{I}$, then $\hat{\boldsymbol{\beta}}$ is the solution of the estimation problem $U(\boldsymbol{\beta}; \boldsymbol{\delta}(\boldsymbol{\beta}; \boldsymbol{W})) = \boldsymbol{L}^T \boldsymbol{W}_T [\boldsymbol{D}(\boldsymbol{\beta}) + \phi \boldsymbol{W}_T (\boldsymbol{I} - \boldsymbol{D}(\boldsymbol{\beta}))]^{-1} (\nabla l(\boldsymbol{L}\boldsymbol{\beta}) + \boldsymbol{\delta}(\boldsymbol{\beta}; \boldsymbol{W})) = \boldsymbol{0}$. By Taylor expansion:

$$0 = U(\hat{\boldsymbol{\beta}}; \overline{\boldsymbol{W}}; \boldsymbol{\delta}(\hat{\boldsymbol{\beta}}; \overline{\boldsymbol{W}})) = U(\boldsymbol{\beta}^*; \overline{\boldsymbol{W}}; \boldsymbol{\delta}(\boldsymbol{\beta}^*; \overline{\boldsymbol{W}})) + \frac{\mathrm{d}U(\boldsymbol{\beta}^*; \overline{\boldsymbol{W}}; \boldsymbol{\delta}(\boldsymbol{\beta}^*; \overline{\boldsymbol{W}}))}{\mathrm{d}\boldsymbol{\beta}^*} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)(1 + o(1)),$$

thus

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \approx -[\frac{\mathrm{d}U(\boldsymbol{\beta}^*; \overline{\boldsymbol{W}}; \boldsymbol{\delta}(\boldsymbol{\beta}^*; \overline{\boldsymbol{W}}))}{n\, \mathrm{d}\boldsymbol{\beta}^*}]^{-1} \cdot \frac{U(\boldsymbol{\beta}^*; \overline{\boldsymbol{W}}; \boldsymbol{\delta}(\boldsymbol{\beta}^*; \overline{\boldsymbol{W}}))}{\sqrt{n}}$$

Let $U_i(\boldsymbol{\beta}^*; \overline{\boldsymbol{W}}; \boldsymbol{\delta}(\boldsymbol{\beta}^*; \overline{\boldsymbol{W}})) := \boldsymbol{L}^T \overline{\boldsymbol{W}}_T [\boldsymbol{D}(\boldsymbol{\beta}^*) + \phi \overline{\boldsymbol{W}}_T (\boldsymbol{I} - \boldsymbol{D}(\boldsymbol{\beta}^*))]^{-1} (\nabla \bar{l}_i(\boldsymbol{L}\boldsymbol{\beta}^*) + \boldsymbol{\delta}(\boldsymbol{\beta}^*; \overline{\boldsymbol{W}}))$. Based on definition (1.28), $U_i(\boldsymbol{\beta}^*; \overline{\boldsymbol{W}}; \boldsymbol{\delta}(\boldsymbol{\beta}^*; \overline{\boldsymbol{W}}))$ follows i.i.d distribution for $i = 1, \ldots, n$. Note that as $n \to \infty$, by central limit theorem:

$$\frac{U(\boldsymbol{\beta}^*; \overline{\boldsymbol{W}}; \boldsymbol{\delta}(\boldsymbol{\beta}^*; \overline{\boldsymbol{W}}))}{\sqrt{n}} = \sqrt{n}[\frac{\sum_{i=1}^n U_i(\boldsymbol{\beta}^*; \overline{\boldsymbol{W}}; \boldsymbol{\delta}(\boldsymbol{\beta}^*; \overline{\boldsymbol{W}}))}{n} - 0] \to \mathbb{N}(\boldsymbol{0}, \boldsymbol{V}_1),$$

where

$$\begin{aligned} \boldsymbol{V}_1 = \lim_{n\to\infty} \mathbb{E}[\frac{1}{n} \sum_{i=1}^n & \boldsymbol{L}^T \overline{\boldsymbol{W}}_T [\boldsymbol{D}(\boldsymbol{\beta}^*) + \phi \overline{\boldsymbol{W}}_T (\boldsymbol{I} - \boldsymbol{D}(\boldsymbol{\beta}^*))]^{-1} \nabla \bar{l}_i((\boldsymbol{L}\boldsymbol{\beta}^*), \boldsymbol{y}) \\ & \cdot \nabla \bar{l}_i((\boldsymbol{L}\boldsymbol{\beta}^*), \boldsymbol{y})^T [\boldsymbol{D}(\boldsymbol{\beta}^*) + \phi \overline{\boldsymbol{W}}_T (\boldsymbol{I} - \boldsymbol{D}(\boldsymbol{\beta}^*))]^{-1} \overline{\boldsymbol{W}}_T \boldsymbol{L}] \end{aligned}$$

$$\frac{\mathrm{d}U(\boldsymbol{\beta}^*; \overline{\boldsymbol{W}}; \boldsymbol{\delta}(\boldsymbol{\beta}^*; \overline{\boldsymbol{W}}))}{n\, \mathrm{d}\boldsymbol{\beta}^*} = \frac{\partial U(\boldsymbol{\beta}^*; \overline{\boldsymbol{W}}; \boldsymbol{\delta}(\boldsymbol{\beta}^*; \overline{\boldsymbol{W}}))}{n\, \partial \boldsymbol{\beta}} + \frac{\partial U(\boldsymbol{\beta}^*; \boldsymbol{\delta}(\boldsymbol{\beta}^*; \overline{\boldsymbol{W}}))}{\partial \boldsymbol{\delta}} \cdot \frac{\partial \boldsymbol{\delta}}{\partial \boldsymbol{\beta}} (\boldsymbol{\beta}^*; \boldsymbol{\delta}(\boldsymbol{\beta}^*; \overline{\boldsymbol{W}})) \cdot \frac{1}{n}. \tag{1.47}$$

Since we have

$$\begin{aligned} \frac{\partial U(\boldsymbol{\beta}^*; \overline{\boldsymbol{W}}; \boldsymbol{\delta}(\boldsymbol{\beta}^*; \overline{\boldsymbol{W}}))}{\partial \boldsymbol{\delta}} &= \boldsymbol{L}^T \overline{\boldsymbol{W}}_T [\boldsymbol{D}(\boldsymbol{\beta}^*) + \phi \boldsymbol{W} (\boldsymbol{I} - \boldsymbol{D}(\boldsymbol{\beta}^*))]^{-1}, \\ \text{and} \quad \| \frac{\partial \boldsymbol{\delta}}{\partial \boldsymbol{\beta}} (\boldsymbol{\beta}^*; \overline{\boldsymbol{W}}; \boldsymbol{\delta}(\boldsymbol{\beta}^*; \overline{\boldsymbol{W}})) \|_2 / n &\to 0, \end{aligned} \tag{1.48}$$

then by law of large numbers:

$$\frac{\partial U(\boldsymbol{\beta}^*; \overline{\boldsymbol{W}}; \boldsymbol{\delta}(\boldsymbol{\beta}^*; \overline{\boldsymbol{W}}))}{n\, \partial \boldsymbol{\beta}} = \frac{\sum_{i=1}^n \partial U_i(\boldsymbol{\beta}^*; \overline{\boldsymbol{W}}; \boldsymbol{\delta}(\boldsymbol{\beta}^*; \overline{\boldsymbol{W}}))}{n\, \partial \boldsymbol{\beta}} \to \boldsymbol{V}_2, \tag{1.49}$$

where

$$\begin{aligned} \boldsymbol{V}_2 &= \lim_{n\to\infty} [\frac{1}{n} \sum_{i=1}^n \boldsymbol{L}^T \overline{\boldsymbol{W}}_T [\boldsymbol{D}(\boldsymbol{\beta}^*) + \phi \overline{\boldsymbol{W}}_T (\boldsymbol{I} - \boldsymbol{D}(\boldsymbol{\beta}^*))]^{-1} \nabla^2 \bar{l}_i(\boldsymbol{L}\boldsymbol{\beta}^*) \boldsymbol{L}] \\ &= \lim_{n\to\infty} [\boldsymbol{L}^T \overline{\boldsymbol{W}}_T [\boldsymbol{D}(\boldsymbol{\beta}^*) + \phi \overline{\boldsymbol{W}}_T (\boldsymbol{I} - \boldsymbol{D}(\boldsymbol{\beta}^*))]^{-1} \boldsymbol{D}(\boldsymbol{\beta}^*) \boldsymbol{L}] \end{aligned} \tag{1.50}$$

To conclude, as n goes to infinity, $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \to \mathbb{N}(\boldsymbol{0}, \boldsymbol{V}_2^{-1} \boldsymbol{V}_1 (\boldsymbol{V}_2^{-1})^T)$.

CHAPTER 2

SLACK EMPIRICAL LIKELIHOOD (SEL)

2.1 Introduction

Inference is one of the core topics in statistics, and the goal is to quantify the uncertainty of population parameter through confidence intervals and hypothesis tests. Estimating equations play an important role in hypothesis test problems. Given n observations $X_1, ..., X_n$ of a random variable X distributed from $\mathbb{P}$ and a parameter of interest $\boldsymbol{\theta}$, many hypothesis test problems can be formulated into the following form with the assistance of estimating equations

$$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0, \quad \text{such that} \quad \mathrm{E}(\boldsymbol{g}(X; \boldsymbol{\theta})) = \mathbf{0}, \tag{2.1}$$

where the expectation is taken with respect to the distribution $\mathbb{P}$. However, the underlying distribution $\mathbb{P}$ could be unknown in practice, and we may not know how to adaptly choose proper distribution families for real dataset. To handle the issue, a popular way is to use the empirical distribution $\mathbb{P}_n : P(X = X_i) = 1/n$ as a plug-in estimation of the unknown distribution. In order to gain more flexibility and adaptivity on hypothesis tests to handle data imperfection, [13] proposed the empirical likelihood method by replacing the empirical distribution with $\mathbb{P}_n^{(\boldsymbol{w})} : P(X = X_i) = w_i$, $\sum_{i=1}^n w_i = 1$, $w_i \geq 0$.

Hypothesis test problem in form (2.1) is widely assumed because it is beyond the traditional likelihood and even loss function setups. However, many modern statistical problems arising in the last several years go beyond that form. For example, consider the hypothesis test for the lasso estimator, which is a solution of minimizing sum square of residuals with ℓ_1 norm regularization. It is known that the estimation equation of the lasso can be characterized in the following way:

$$\boldsymbol{\beta} = \Theta_S(\boldsymbol{\beta} - \boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}); \lambda), \quad \text{where} \quad \Theta_S(t; \lambda) = \begin{cases} 0, & \text{if } |t| < \lambda; \\ (|t| - \lambda)\mathrm{sgn}(t), & \text{if } |t| \geq \lambda \end{cases}$$

is the soft-thresholding function, $\mathrm{sgn}(t)$ returns the sign of t and $\Theta_S(\cdot; \lambda)$ makes effect on all components of a vector. Due to the *many-to-one* mapping of the soft-thresholding function near

zero, there is no direct way to transform estimating equation of the lasso into the form (2.1). Hence, traditional nonparametrization for the lasso problem through estimating equations will fail.

In this paper, we consider an even more general type of formulation of hypothesis test problems than (2.1), which includes hypothesis tests for the lasso. In particular, the estimating equation for the lasso can not be written as sample additive form because of the nonlinearity of Θ_S. To achieve the purpose, we use p additional variables to restate the optimality conditions, which are slack variables. With the help of these slack variables, the estimating equation can be written as sample additive form hence nonparametrization inference can be performed through the framework of empirical likelihood.

In this article, we propose a slack empirical likelihood framework for statistical inference. By applying our framework into problems like regressions with affine inequality constrains and high dimensional inference for regressions, we show that SEL offers a mechanism of nonparametrization and robustification for the parametric assumptions. In addition, our SEL can be applied to a large variety of penalized and constrained form of optimization problem, which could reflect elementwise sparsity, row-wise sparsity, and low-rankness of the inference targeted matrix. Both asymptotic and nonasymptotic analysis are provided for SEL. Finally, some simulations are performed to demonstrate the advantages of our approach.

The organization of this article is as follows: in section 2.2, we have a brief review the empirical likelihood methods and introduce our general framework for slack empirical likelihood; we show how to construct SEL from general optimization problem in section 2.3; in section 2.4, we apply our framework for slack empirical likelihood into some specific examples, also to show the details how our methods can be applied for traditional asymptotic inference; some explorations on how to apply SEL for high dimensional problems is offered in section 2.5; numerical evidence are provided in section 2.6.

2.2 Slack Empirical Likelihood

2.2.1 Statistic-driven EL

We first review the some classical EL. Given n observations $\boldsymbol{z}_i \in \mathbb{R}^m$ and $\boldsymbol{\mu}^\circ$, a location of interest, [14] proposed to test $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}^\circ$ by $L(\hat{w}_1, \ldots, \hat{w}_n) - L(1/n, \ldots, 1/n)$, where $L(\hat{w}_1, \ldots, \hat{w}_n)$

is the optimal function value obtained by

$$\min_{[w_i]\in\mathbb{R}^n} -\sum \log w_i \equiv L(w_1,\ldots,w_n) \text{ s.t. } \sum w_i = 1, w_i \geq 0, \sum w_i \boldsymbol{z}_i = \boldsymbol{\mu}^\circ.$$

From the last constraint, w_i has a probability meaning in an associated multinomial distribution $\boldsymbol{z} = \boldsymbol{z}_i$ $(1 \leq i \leq n)$, thereby called the empirical likelihood method.

EL is actually statistic driven although it is widely used as a non-parametric tool for data analysis. Throughout the paper, we assume a supervised setup unless otherwise specified, which has n samples of p predictors and m responses available: $(\boldsymbol{x}_i, \boldsymbol{y}_i) \in \mathcal{S} \subset \mathbb{R}^p \times \mathbb{R}^m$. We call $\mathcal{S}$ the sample space and construct $\boldsymbol{X} = [\boldsymbol{x}_1 \ldots \boldsymbol{x}_n]^T \in \mathbb{R}^{n\times p}$ and $\boldsymbol{Y} = [\boldsymbol{y}_1 \ldots \boldsymbol{y}_n]^T \in \mathbb{R}^{n\times m}$. Let $\boldsymbol{B}$ be determined by a set of estimating equations:

$$\sum_{i=1}^n \boldsymbol{T}(\boldsymbol{B}; \boldsymbol{x}_i, \boldsymbol{y}_i) = \boldsymbol{0} \tag{2.2}$$

which is assumed to be *additive in samples*. In this paper, the sample additivity is a key in defining EL in various statistics contexts. For location EL, $\boldsymbol{B}$ is the location parameter matrix, but can be much more general. In statistical estimation, (2.2) can often be obtained by minimizing a certain objective—the optimal $\boldsymbol{B} \in \mathbb{R}^{p\times m}$ that minimizes $\min_{\boldsymbol{B}} f(\boldsymbol{B}; \boldsymbol{X}, \boldsymbol{Y}) \triangleq \sum_i l(\boldsymbol{B}; \boldsymbol{x}_i, \boldsymbol{y}_i)$ can be obtained by solving $\nabla_{\boldsymbol{B}} f = 0$ or (2.2) with $\boldsymbol{T}(\boldsymbol{B}; \boldsymbol{x}_i, \boldsymbol{y}_i) = \nabla_{\boldsymbol{B}} l(\boldsymbol{B}; \boldsymbol{x}_i, \boldsymbol{y}_i)$, assuming l is differentiable in $\mathbb{R}^{p\times m}$. We call $\boldsymbol{T}(\boldsymbol{B}; \boldsymbol{x}_i, \boldsymbol{y}_i)$ the residuals, and assume $\boldsymbol{T}_i(\boldsymbol{B}) \in \mathcal{G} \subset \mathbb{R}^{p\times m}$. The *residual space* $\mathcal{G}$ is often taken to be the full Euclidian space but not necessarily so which may impose more structural properties on $\{w_i\}$.

With the desired residuals provided by (2.2), we can define the EL statistic for any given point $\boldsymbol{B}^\circ$ in the parameter space $\Omega \subset \mathbb{R}^{p\times m}$ by solving the following optimization problem (with $\boldsymbol{w} = [w_i] \in \mathbb{R}^n$):

$$\begin{aligned} \min_{\boldsymbol{w}} -\sum_{i=1}^n \log(nw_i) \quad \text{s.t.} \quad & \sum_{i=1}^n w_i = 1, w_i \geq 0 \\ & \sum_{i=1}^n w_i \boldsymbol{T}(\boldsymbol{B}^\circ; \boldsymbol{x}_i, \boldsymbol{y}_i) = \boldsymbol{0}. \end{aligned} \tag{2.3}$$

An essential component of (2.3) is the last constraint, the so-called *structural constraint*. We will denote $\boldsymbol{T}(\boldsymbol{B}^\circ; \boldsymbol{x}_i, \boldsymbol{y}_i)$ by $\boldsymbol{T}_i^\circ$ for simplicity. A deep result is that under the null hypothesis H_0 : $\boldsymbol{B} = \boldsymbol{B}^\circ$, twice the optimal function value obtained in (2.3) follows an asymptotic χ^2_p distribution as $n \to +\infty$ in a nonparametric sense under some regularity conditions [16].

The residuals $\boldsymbol{T}_i^\circ$ are obtained on the ith observation using the $\boldsymbol{B}^\circ$, such as $\boldsymbol{x}_i(\boldsymbol{x}_i^T\boldsymbol{B}^\circ - \boldsymbol{y}_i^T)$ in regression. For example, when $\boldsymbol{X}$ is a column of ones and $l(\boldsymbol{\mu}^\circ;\boldsymbol{z}_i) = \|\boldsymbol{\mu}^\circ - \boldsymbol{z}_i\|_2^2/2$, $\boldsymbol{T}(\boldsymbol{\mu}^\circ;\boldsymbol{x}_i,\boldsymbol{z}_i) = \boldsymbol{\mu}^\circ - \boldsymbol{z}_i$, and so (2.3) degenerates to the location empirical likelihood, and if $l(\boldsymbol{\beta}^\circ;\boldsymbol{x}_i,y_i) = (\boldsymbol{x}_i^T\boldsymbol{\beta}^\circ - y_i)^2/2$, $\boldsymbol{T}(\boldsymbol{\beta}^\circ;\boldsymbol{x}_i,y_i) = (\boldsymbol{\beta}^{\circ T}\boldsymbol{x}_i - y_i)\boldsymbol{x}_i^T$, corresponding to the celebrated regression empirical likelihood [15]. In a sense, the residuals and structural constraint are rooted in a parametric model, but EL offers a mechanism of nonparametrization and robustification.

To gain more insights, we rewrite (2.3) for supervised learning with an estimation criterion of $f(\boldsymbol{B}) = \bar{l}(\boldsymbol{\Theta};\boldsymbol{Y}) = \sum_i l_0(\boldsymbol{B}^T\boldsymbol{x}_i;\boldsymbol{y}_i)$, where a loss $\ell_0 \in \mathcal{C}^1$ is placed on the *systematic component* $\boldsymbol{\Theta} = \boldsymbol{X}\boldsymbol{B}$. Clearly,

$$\nabla_{\boldsymbol{B}} f(\boldsymbol{B}) = \boldsymbol{X}^T\nabla_{\boldsymbol{\Theta}}\bar{l}(\boldsymbol{\Theta}).$$

Given $\boldsymbol{B}^\circ \in \Omega$, let $\boldsymbol{R} = \nabla_{\boldsymbol{\Theta}}\bar{l}\,|_{\boldsymbol{\Theta}=\boldsymbol{X}\boldsymbol{B}^\circ}$. We can also write the matrix $\boldsymbol{R} = [\boldsymbol{r}_1,\ldots,\boldsymbol{r}_n]^T$ with $\boldsymbol{r}_i = \nabla l_0(\mathrm{vec}\,(\boldsymbol{\Theta}[i,]);\boldsymbol{y}_i)$. Then $\boldsymbol{T}_i^\circ = \boldsymbol{x}_i\boldsymbol{r}_i^T$. An instance is given by a vector generalized linear model (GLM) with m responses. Let b be the cumulant function and adopt the canonical link $g = (b')^{-1}$. Then $l(\boldsymbol{B};\boldsymbol{x}_i,\boldsymbol{y}_i)$ or $l_0(\boldsymbol{B}^T\boldsymbol{x}_i;\boldsymbol{y}_i)$ is $-\langle\boldsymbol{B}^T\boldsymbol{x}_i,\boldsymbol{y}_i\rangle + \langle\boldsymbol{1},b(\boldsymbol{B}^T\boldsymbol{x}_i)\rangle$. Direct calculation shows

$$\boldsymbol{R} = b'(\boldsymbol{X}\boldsymbol{B}^\circ) - \boldsymbol{Y},$$

where $b'(\cdot)$ is applied element-wise. When $b'(\cdot)$ is the identity function, it gives the multivariate regression empirical likelihood.

Let $\mathcal{C} = \{w_1,\ldots,w_n : \sum_{i=1}^n w_i = 1, w_i \geq 0, 1 \leq i \leq n\}$ and $\boldsymbol{W} = \mathrm{diag}\{w_i\}$. The associated EL problem can restated in the following form

$$\min_{\boldsymbol{w}\in\mathcal{C}} -\langle\boldsymbol{1},\log(n\boldsymbol{w})\rangle \quad \text{s.t.} \quad \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{R} = \boldsymbol{0}. \tag{2.4}$$

To generalize the idea of EL, the structural constraint $\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{R} = \boldsymbol{0}$ is the key.

2.2.2 Introducing Slack Variables

The definitions and extensions in the previous section apply well when the objective function is differentiable in the full parameter space. Modern statistical applications however pose new challenges. In particular, (a) when the parameter space Ω is restricted, evaluating the gradient in $\mathbb{R}^{p\times m}$ may not deliver reasonable residuals; (b) the objective function for high dimensional sparse learning is typically nonsmooth. As we will see, introducing auxiliary slack variables helps. We demonstrate our techniques with some examples.

Example 1. (Nonnegative regression EL) As an extension of the celebrated regression EL, we consider EL in a setting where all coefficients are nonnegative. This corresponds to the nonnegative least squares problem: $\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2$ s.t. $\beta_j \geq 0$ $(1 \leq j \leq p)$, or $\boldsymbol{\beta} \in \mathbb{R}_+^p$ where $\mathbb{R}_+ = [0, \infty]$. The parameter space is restricted and *closed*. Regression EL can be applied if $\boldsymbol{\beta}^\circ$ is an interior point, but if $\boldsymbol{\beta}^\circ$ lies on the boundary, i.e., $\beta_j^\circ = 0$ for some j, which is of practical interest in significance tests, regular EL does not apply, and the normal-equation-based residuals must be corrected.

Example 2. (EL with affine inequality constraints) Consider test with constraints $\boldsymbol{A}\boldsymbol{\beta} \preceq \boldsymbol{\alpha}$ with $\boldsymbol{A} \in \mathbb{R}^{k \times p}$ (k can be smaller than or greater than p). If the constraints contain equalities $\boldsymbol{A}_1\boldsymbol{\beta} = \boldsymbol{\alpha}_1$, they can be converted to inequality constraints $\boldsymbol{A}_1\boldsymbol{\beta} \preceq \boldsymbol{\alpha}_1$ and $-\boldsymbol{A}_1\boldsymbol{\beta} \preceq \boldsymbol{\alpha}_1$, so we can just consider a general case $\boldsymbol{A}\boldsymbol{\beta} \preceq \boldsymbol{\alpha}$. The constraints should be compatible such that the true value $\boldsymbol{\beta}^\circ$ is indeed in the feasible region of these constrains. This corresponds to the least squares problem with constrains: $\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2$ s.t. $\boldsymbol{A}\boldsymbol{\beta} \preceq \boldsymbol{\alpha}$. Still The parameter space is restricted and closed. And if $\boldsymbol{\beta}^\circ$ is an interior point, regression EL can still be applied.

Example 3. (Large-p EL) Consider a sparse learning problem $\min_{\boldsymbol{\beta}} \sum_{i=1}^n l_0(\boldsymbol{x}_i^T\boldsymbol{\beta}) + \sum_{j=1}^p P(|\beta_j|; \lambda)$, where l_0 is a loss function defined on the systematic component $\boldsymbol{x}_i^T\boldsymbol{\beta}$ and P is a penalty function to promote the sparsity in $\boldsymbol{\beta}$. Examples of P include the ℓ_1 penalty, ℓ_0-penalty and SCAD [4], which are all popular for building a high-dimensional linear model. We assume that the regularization parameter λ is *given* (either by theory—see, e.g., [2], or by tuning such as cross-validation) so that the criterion is fully specified. A new class of empirical likelihoods say L_1-EL or L_0-EL would be useful for high-dimensional inference, but the non differentiability along with possible nonconvexity makes it difficult to obtain sample-additive EEs.

Example 1 is tricky because the closed restricted parameter space $\mathbb{R}_+^p$ makes the gradient-derived residuals improper for any target point on its boundary. Example 3 shows a nondifferentiability issue caused by the objective function. The good news is that we can employ the same trick of slack variables to rewrite the optimality conditions to extend EL, which we call slack empirical likelihood (**SEL**).

Let $\bar{l}(\boldsymbol{X}\boldsymbol{\beta}; \boldsymbol{y}) = \sum_i l_0(\boldsymbol{x}_i^T\boldsymbol{\beta}; \boldsymbol{y}_i)$ and recall the sparse estimation problem

$$\min_{\boldsymbol{\beta}} f(\boldsymbol{\beta}) \triangleq \bar{l}(\boldsymbol{X}\boldsymbol{\beta}; \boldsymbol{y}) + \sum_{j=1}^p P(|\beta_j|; \lambda). \tag{2.5}$$

For simplicity, we assume $\bar{l}$ to be differentiable, but the technique applies to any objective function that is only directionally differentiable. Apart from convex penalties, nonconvex penalties have recently received much attention, due to their capability of accommodating coherent designs and reducing estimation bias. All of the sparsity-promoting penalties are associated with **thresholding** rules. To make the connection more clear, we use the framework of Θ-estimators [21] to study EL driven by an *arbitrary* thresholding function Θ. A thresholding function $\Theta(t;\lambda)$ is defined for $-\infty < t < \infty$ and $0 \le \lambda < \infty$ such that (i) $\Theta(-t;\lambda) = -\Theta(t;\lambda)$; (ii) $\Theta(t;\lambda) \le \Theta(t';\lambda)$ for $t \le t'$; (iii) $\lim_{t\to\infty} \Theta(t;\lambda) = \infty$; (iv) $0 \le \Theta(t;\lambda) \le \infty$ for $0 \le t < \infty$. Given any $u \ge 0$, $\Theta^{-1}(u;\lambda) \triangleq \sup\{t : \Theta(t;\lambda) \le u\}$, and so the threshold is $\tau(\lambda) \triangleq \Theta^{-1}(0;\lambda)$. For simplicity, we always assume that λ is the threshold parameter, i.e., $\tau(\lambda) = \lambda$. A vector version of Θ is defined component-wise if either t or λ is replaced by a vector. Given any thresholding Θ, we say P is induced by Θ if

$$P(t;\lambda) = P_\Theta(t;\lambda) + q(t;\lambda) = \int_0^{|t|} (\Theta^{-1}(u;\lambda) - u)\,\mathrm{d}u + q(t;\lambda), \tag{2.6}$$

where q is an arbitrary nonnegative function satisfying $q(t;\lambda) = 0$ if $t = \Theta(s;\lambda)$ for some $s \in \mathbb{R}$. The mapping from P to Θ is *many-to-one* if Θ has discontinuities. The Θ-P framework covers all aforementioned penalties: ridge scaling $\Theta_R(t;\eta) = t/(1+\eta)$ corresponds to the ℓ_2-penalty $P(t;\eta) = \eta t^2/2$, soft-thresholding $\Theta_S(t;\lambda) = \mathrm{sgn}(t)(|t| - \lambda)_+$ gives the ℓ_1-penalty $P(t;\lambda) = \lambda|t|$, and hard-thresholding $\Theta_H(t;\lambda) = t1_{|t|\ge\lambda}$ induces infinitely many ℓ_0-type penalties including the discrete ℓ_0 penalty and the capped ℓ_1 penalty; see [21] for other examples of elastic net, SCAD, MCP, ℓ_r $(0 \le r < 1)$, and so on. Given (2.6), a class of iterative thresholding procedures can be used to solve (2.5), the resulting estimates called Θ-estimates.

The key to defining EL lies in the characterization a 'fit' or Θ-estimate. Let $f_\Theta(\boldsymbol{\beta}) = \bar{l}(\boldsymbol{X}\boldsymbol{\beta};\boldsymbol{y}) + \sum_{j=1}^p P_\Theta(|\beta_j|;\lambda)$. From [21], any fixed point or locally optimal solution $\hat{\boldsymbol{\beta}}$ of the iterative thresholding procedure for solving (2.5) satisfies $\mathrm{D}_{\pm\boldsymbol{e}_j} f_\Theta(\boldsymbol{\beta}) \ge 0$ for all $1 \le j \le p$, where $\boldsymbol{e}_j$ is the vector with the jth component 1 and the other components 0, and $\mathrm{D}_{\boldsymbol{u}} f_\Theta(\boldsymbol{\beta})$ denotes the directional derivative of f_Θ at $\boldsymbol{\beta}$ with increment $\boldsymbol{u}$, i.e., $\mathrm{D}_{\boldsymbol{u}} f_\Theta(\boldsymbol{\beta}) = \lim_{\epsilon\to 0+}[f_\Theta(\boldsymbol{\beta}+\epsilon\boldsymbol{u}) - f_\Theta(\boldsymbol{\beta})]/\epsilon$. These conditions do lead to a set of Θ-*equations* [21, Theorem 1]:

$$\hat{\boldsymbol{\beta}} = \Theta(\hat{\boldsymbol{\beta}} - \boldsymbol{X}^T \nabla \bar{l}(\boldsymbol{X}\hat{\boldsymbol{\beta}});\lambda), \tag{2.7}$$

under the mild assumption that $\Theta(\cdot;\lambda)$ is continuous at $\hat{\boldsymbol{\beta}} - \boldsymbol{X}^T\nabla\bar{l}(\boldsymbol{X}\hat{\boldsymbol{\beta}})$. Moreover, all Θ-estimates enjoy provable statistical guarantees.

However, unless Θ is the ridge scaling, (2.7) is **not** sample additive due to the nonlinearity of Θ. To achieve the purpose, we use p additional variables (and inequalities) to restate the optimality conditions. Given $\boldsymbol{\beta}^\circ \in \mathbb{R}^p$, define $\mathcal{J} = \{j : \beta_j^\circ \neq 0\}$ and $\mathcal{J}^c = \{j : \beta_j^\circ = 0\}$. For any $j \in \mathcal{J}$, a simple derivative calculation shows that $\Theta^{-1}(|\beta_j^\circ|;\lambda)\operatorname{sgn}(\beta_j^\circ) = \beta_j^\circ - \boldsymbol{X}[,j]^T\nabla\bar{l}(\boldsymbol{X}\boldsymbol{\beta}^\circ)$, and for $j \in \mathcal{J}^c$, $\beta_j^\circ = 0$, and so $-\lambda \le \boldsymbol{X}[,j]^T\nabla\bar{l}(\boldsymbol{X}\boldsymbol{\beta}^\circ) \le \lambda$ by definition. Let

$$\gamma_j^\circ = \begin{cases} \Theta^{-1}(|\beta_j^\circ|;\lambda)\operatorname{sgn}(\beta_j^\circ) - \beta_j^\circ, & \text{if } j \in \mathcal{J} \\ 0, & \text{if } j \in \mathcal{J}^c. \end{cases} \tag{2.8}$$

(2.7) is now replaced by $\boldsymbol{X}^T\nabla\bar{l}(\boldsymbol{X}\boldsymbol{\beta}^\circ) + \boldsymbol{\gamma}^\circ + \boldsymbol{s} = \boldsymbol{0}$, which *is* sample additive, where $\boldsymbol{s} \in \mathbb{R}^p$, $\boldsymbol{s}_{\mathcal{J}} = \boldsymbol{0}$, and $|s_j| \le \lambda$, $j \in \mathcal{J}^c$. Correspondingly, given any $\Theta(\cdot;\lambda)$ and $\boldsymbol{\beta}^\circ \in \mathbb{R}^p$, we cast SEL as a **joint** optimization with respect to the weights $\boldsymbol{w}$ and the slack variables $\boldsymbol{s} = [s_j]$:

$$\begin{aligned} \min_{\boldsymbol{w}\in\mathcal{C},\boldsymbol{s}\in\mathbb{R}^p} -\langle \boldsymbol{1}, \log(\boldsymbol{w})\rangle \quad \text{s.t.} \quad & \boldsymbol{X}^T\boldsymbol{W}\nabla\bar{l}(\boldsymbol{X}\boldsymbol{\beta}^\circ) + \boldsymbol{\gamma}^\circ + \boldsymbol{s} = \boldsymbol{0} \\ & \boldsymbol{s}\circ\boldsymbol{\beta}^\circ = \boldsymbol{0},\ \|\boldsymbol{s}\|_\infty \le \lambda, \end{aligned} \tag{2.9}$$

where $\circ$ denotes the elementwise product and $\|\boldsymbol{s}\|_\infty = \max|s_j|$. If we define $\mathcal{Z} = \mathcal{J}^c$, then $\boldsymbol{s}$ satisfies $-\lambda \le s_j \le \lambda$ for $j \in \mathcal{Z}$ and $s_j = 0$ for $j \in \mathcal{J}$.

A useful version is the sparsity EL under an ℓ_0-constraint $\|\boldsymbol{\beta}\|_0 \le q$. From [19], all locally optimal solutions to $\min_{\|\boldsymbol{\beta}\|_0\le q}\bar{l}(\boldsymbol{X}\boldsymbol{\beta};\boldsymbol{y})$ satisfy $\hat{\boldsymbol{\beta}} = \Theta^{\#}(\hat{\boldsymbol{\beta}} - \boldsymbol{X}^T\nabla\bar{l}(\boldsymbol{X}\hat{\boldsymbol{\beta}});q)$, assuming no ties occur and $\|\boldsymbol{X}\|_2 \le 1/L$ with L the Lipschitz constant of $\nabla\bar{l}$. Here, the quantile thresholding $\Theta^{\#}(\boldsymbol{\alpha};q)$ for any $\boldsymbol{\alpha} \in \mathbb{R}^p$ is defined to be a vector $\boldsymbol{\gamma} \in \mathbb{R}^p$ satisfying $\gamma_{(j)} = \alpha_{(j)}$ if $1 \le j \le q$, and 0 otherwise, where $\alpha_{(1)},\ldots,\alpha_{(p)}$ are the order statistics of $\alpha_1,\ldots,\alpha_p$ satisfying $|\alpha_{(1)}| \ge \cdots \ge |\alpha_{(p)}|$. Rewriting the $\Theta^{\#}$-equation by use of slack variables, we get the SEL problem for any given $\|\boldsymbol{\beta}^\circ\|_0 \le q$:

$$\begin{aligned} \min_{\boldsymbol{w}\in\mathcal{C},\boldsymbol{s}} -\langle \boldsymbol{1}, \log(\boldsymbol{w})\rangle \quad \text{s.t.} \quad & \boldsymbol{X}^T\boldsymbol{W}\nabla\bar{l}(\boldsymbol{X}\boldsymbol{\beta}^\circ) + \boldsymbol{s} = \boldsymbol{0} \\ & \boldsymbol{s}\circ\boldsymbol{\beta}^\circ = \boldsymbol{0},\ \|\boldsymbol{s}\|_\infty < |\beta_{(q)}^\circ|, \end{aligned} \tag{2.10}$$

where $|\beta_{(q)}^\circ|$ is the qth largest element in $|\beta_j^\circ|$, $1 \le j \le p$. Note that all $\boldsymbol{s}$-constraints in SEL, either in equalities or inequalities, are **linear**.

Nonnegative regression EL can be defined similarly. Consider $\min_{\boldsymbol{\beta}\succeq\boldsymbol{0}}\sum_i l_0(\boldsymbol{\beta};\boldsymbol{x}_i,y_i) = f(\boldsymbol{\beta})$, where l is differentiable in the augmented parameter space $\mathbb{R}^p$ but not necessarily convex. Any

optimal solution $\boldsymbol{\beta}$ obeys $\mathrm{D}_{\boldsymbol{u}} f(\boldsymbol{\beta}) \geq 0$ for all feasible $\boldsymbol{u}$. Specifically, taking $\boldsymbol{u} = \pm \boldsymbol{e}_j$ for $j \in \mathcal{J}(\boldsymbol{\beta}) = \{j : \beta_j \neq 0\}$ and $\boldsymbol{u} = \boldsymbol{e}_j$ for $j \in \mathcal{J}^c(\boldsymbol{\beta})$ leads to the estimating equations with slack variables:

$$\sum_i \{\nabla l(\boldsymbol{\beta}; \boldsymbol{x}_i, y_i) + (1/n)\boldsymbol{s}\} = \mathbf{0},$$

where $\boldsymbol{s}_{\mathcal{J}^c} \succeq \mathbf{0}$ and $\boldsymbol{s}_{\mathcal{J}} = \mathbf{0}$. Given $\boldsymbol{\beta}^\circ$, adding w_i in the first term and replacing $1/n$ by $1/w_i$ gives the SEL test statistic: $\min_{\boldsymbol{w} \in \mathcal{C}, \boldsymbol{s}} -\langle \mathbf{1}, \log(n\boldsymbol{w}) \rangle$ s.t. $\boldsymbol{X}^T \boldsymbol{W} \nabla \bar{l}(\boldsymbol{X}\boldsymbol{\beta}^\circ) + \boldsymbol{s} = \mathbf{0}, \boldsymbol{s} \circ \boldsymbol{\beta}^\circ = \mathbf{0},\ \boldsymbol{s} \succeq \mathbf{0}$. When $l(\boldsymbol{\beta}; \boldsymbol{x}_i, y_i) = (\boldsymbol{x}_i^T \boldsymbol{\beta} - y_i)^2/2$, we get

$$\begin{aligned} \min_{\boldsymbol{w} \in \mathcal{C}, \boldsymbol{s}} \ & -\langle \mathbf{1}, \log(n\boldsymbol{w}) \rangle \quad \text{s.t.} \quad \boldsymbol{X}^T \boldsymbol{W} \nabla \bar{l}(\boldsymbol{X}\boldsymbol{\beta}^\circ) + \boldsymbol{s} = \mathbf{0} \\ & \qquad\qquad\qquad\qquad\qquad\quad \boldsymbol{s} \circ \boldsymbol{\beta}^\circ = \mathbf{0},\ \boldsymbol{s} \succeq \mathbf{0}. \end{aligned} \tag{2.11}$$

If $\boldsymbol{\beta}^\circ \succ \mathbf{0}$, then $\boldsymbol{s} = \mathbf{0}$ and (2.11) reduces to the ordinary regression EL.

For EL with affine inequality constraints, Consider $\min_{\boldsymbol{A}\boldsymbol{\beta} \preceq \boldsymbol{a}} \sum_i l(\boldsymbol{\beta}; \boldsymbol{x}_i, y_i) = f(\boldsymbol{\beta})$, now when we consider the optimal solution $\boldsymbol{\beta}$ for all feasible $\boldsymbol{u}$: when $\boldsymbol{A}[i,]\boldsymbol{\beta} < a_i$, $\boldsymbol{u} = \pm \boldsymbol{A}[i,]/\|\boldsymbol{A}[i,]\|_{\ell_2}$, and if $\boldsymbol{A}[i,]\boldsymbol{\beta} = a_i$, $\boldsymbol{u} = -\boldsymbol{A}[i,]/\|\boldsymbol{A}[i,]\|_{\ell_2}$. Now let $\mathcal{J} = \{j = 1, 2, ..., k \mid \boldsymbol{A}[j,]\boldsymbol{\beta} < a_i\}$ and $\mathcal{J}^c = \{j = 1, 2, ..., k \mid \boldsymbol{A}[j,]\boldsymbol{\beta} = a_i\}$. Then given $\boldsymbol{\beta}^\circ$ in the feasible region, suppose the loss is ℓ_2 type, we can obtain the SEL statistics:

$$\begin{aligned} \min_{\boldsymbol{w} \in \mathcal{C}, \boldsymbol{s}} \ & -\langle \mathbf{1}, \log(n\boldsymbol{w}) \rangle \quad \text{s.t.} \quad \boldsymbol{X}^T \boldsymbol{W} \nabla \bar{l}(\boldsymbol{X}\boldsymbol{\beta}^\circ) + \boldsymbol{s} = \mathbf{0} \\ & \qquad\qquad\qquad\qquad\qquad\quad \boldsymbol{A}[\mathcal{J},]\boldsymbol{s} = \mathbf{0},\ \boldsymbol{A}[\mathcal{J}^c]\boldsymbol{s} \preceq \mathbf{0}. \end{aligned} \tag{2.12}$$

2.3 Constructing SEL from General Optimizations

Consider the following optimization problem:

$$\min_{\boldsymbol{B}} f(\boldsymbol{B}) := \min_{\boldsymbol{B}} \bar{l}(\boldsymbol{X}\boldsymbol{B}; \boldsymbol{Y}) + P(\boldsymbol{B}; \lambda), \tag{2.13}$$

where $\bar{l}(\boldsymbol{X}\boldsymbol{B}; \boldsymbol{Y}) = \sum_i l_0(\boldsymbol{x}_i^T \boldsymbol{B}; \boldsymbol{y}_i)$ with a well-defined differentiable loss function l_0, which is assumed to be L-Lipschitz continuous. $\boldsymbol{Y} \in \mathbb{R}^{n \times m}$, $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ and $\boldsymbol{B} \in \mathbb{R}^{p \times m}$. Here $P(\boldsymbol{B}; \lambda)$ is general, which can be element-wise sparsity, group-wise sparsity and low-rankness inducing penalties. For ease of introduction, we first study SEL construction with a special family of penalties in section, then we extend our analysis to SEL with general penalties or constraints.

2.3.1 SEL for Θ Estimators

To start with, we first study SEL constructions from optimizations with a family of regularizations P_Θ, which are induced by some thresholding functions $\Theta(t;\lambda)$ defined as follows:

Definition 4. (Thresholding function). A thresholding function is a real valued function $\Theta(t;\lambda)$ defined for $-\infty < t < \infty$ and $0 \le \lambda < \infty$ such that (i) $\Theta(-t;\lambda) = -\Theta(t;\lambda)$; (ii) $\Theta(t;\lambda) \le \Theta(t';\lambda)$ for $t \le t'$; (iii) $\lim_{t\to\infty} \Theta(t;\lambda) = \infty$; (iv) $0 \le \Theta(t;\lambda) \le t$ for $0 \le t < \infty$.

We always assume λ as the threshold parameter such that $\lambda = \Theta^{-1}(0;\lambda)$. This family of penalties P_Θ can be constructed from Θ as follows:

$$P_\Theta(t;\lambda) = \int_0^{|t|} \left(\Theta^{-1}(u;\lambda) - u\right) \mathrm{d}u = \int_0^{|t|} (\sup\{s : \Theta(s;\lambda) \le u\} - u)\mathrm{d}u. \tag{2.14}$$

As shown in Theorem 1 in [21], the local minimum points $\hat{\boldsymbol{B}}$ for optimization (2.13) can be characterized by the following Θ equation when penalty P belongs to the family of P_Θ :

$$\hat{\boldsymbol{B}} = \Theta(\hat{\boldsymbol{B}} - \boldsymbol{X}^T\nabla\bar{l}(\boldsymbol{X}\hat{\boldsymbol{B}});\lambda), \tag{2.15}$$

under mild condition such that $\Theta(:;\lambda)$ is continuous at $\hat{\boldsymbol{B}} - \boldsymbol{X}^T\nabla\bar{l}(\boldsymbol{X}\hat{\boldsymbol{B}})$.

(Element-wise thresholdings) Consider the case when the penalty function $P_\Theta(;\lambda)$ for a matrix $\boldsymbol{B}$ is define as follows:

$$P_\Theta(\boldsymbol{B};\lambda) = \sum_{i=1}^{p}\sum_{j=1}^{m} P(B[i,j];\lambda), \tag{2.16}$$

then the corresponding $\Theta(\boldsymbol{B};\lambda)$ is a function on all components of $\boldsymbol{B}$. Given $\boldsymbol{B}^\circ \in \mathbb{R}^{p\times m}$, define $\mathcal{J} = \{(i,j) : B^\circ_{ij} \neq 0\}$ and $\mathcal{J}^c = \{(i,j) : B^\circ_{ij} = 0\}$. If $(i,j) \in \mathcal{J}$, by inverting the Θ function, we have $\Theta^{-1}(|B^\circ_{i,j}|;\lambda)\,\mathrm{sgn}(B^\circ_{i,j}) = B^\circ_{i,j} - [\boldsymbol{X}^T\nabla\bar{l}(\boldsymbol{X}\boldsymbol{B}^\circ)]_{i,j}$; If $(i,j) \in \mathcal{J}^c$, we have $B^\circ_{i,j} = 0$, which implies $|[\boldsymbol{X}^T\nabla\bar{l}(\boldsymbol{X}\boldsymbol{B}^\circ)]_{ij}| \le \lambda$. Define a matrix $\boldsymbol{\Gamma}^\circ$ such that

$$\Gamma^\circ_{i,j} = \begin{cases} \Theta^{-1}(|B^\circ_{i,j}|;\lambda)\,\mathrm{sgn}(B^\circ_{i,j}) - B^\circ_{i,j} & \text{if } (i,j) \in \mathcal{J}, \\ 0 & \text{if } (i,j) \in \mathcal{J}^c. \end{cases} \tag{2.17}$$

Then by introducing the slack variable $\boldsymbol{S}$ such that $\boldsymbol{X}^T\nabla\bar{l}(\boldsymbol{X}\boldsymbol{B}^\circ) + \boldsymbol{\Gamma}^\circ + \boldsymbol{S} = \boldsymbol{0}$, we have constraints $\boldsymbol{S}\circ\boldsymbol{B}^\circ = \boldsymbol{0}$ and $\|\boldsymbol{S}\|_{\max} \le \lambda$. Finally, the corresponding SEL is:

$$\begin{aligned} &\min_{w\in\mathcal{C},\boldsymbol{S}\in\mathbb{R}^{p\times m}} -\langle 1, \log(\boldsymbol{n}\boldsymbol{w})\rangle \ \text{ s.t.} \quad \boldsymbol{X}^T\boldsymbol{W}\nabla\bar{l}(\boldsymbol{X}\boldsymbol{B}^\circ) + \boldsymbol{\Gamma}^\circ + \boldsymbol{S} = \boldsymbol{0} \\ &\boldsymbol{S}\circ\boldsymbol{B}^\circ = \boldsymbol{0}, \quad \|\boldsymbol{S}\|_{\max} \le \lambda, \end{aligned} \tag{2.18}$$

where $\mathcal{C} = \{w_1, \ldots, w_n : \sum_{i=1}^n w_i = 1, w_i \geq 0, 1 \leq i \leq n\}$ and $\boldsymbol{W} = \operatorname{diag}\{w_i\}$.

(Group-wise thresholdings) Sometimes it is more interesting to consider group-wise sparsity structure on coefficient $\boldsymbol{B}$. Let $P_{2,\Theta}$ be a penalty function for a matrix that promotes sparsity in the following way

$$P_{2,\Theta}(\boldsymbol{B}; \lambda) = \sum_{i=1}^{p} P_{\Theta}(\|\boldsymbol{B}[i,]\|_2; \lambda). \tag{2.19}$$

As discussed in [20], we define multivariate threshold function as follows:

Definition 5. (Multivariate Threshold function). Given any $\Theta, \vec{\Theta}$ is defined for any vector $\boldsymbol{\alpha} \in \mathbb{R}^m$ such that $\vec{\Theta}(\boldsymbol{\alpha}; \lambda) = \boldsymbol{\alpha}\Theta(\|\boldsymbol{\alpha}\|_2; \lambda)/\|\boldsymbol{\alpha}\|_2$ for $\boldsymbol{\alpha} \neq \mathbf{0}$ and $\mathbf{0}$ otherwise. For any matrix $\boldsymbol{A} = (\boldsymbol{\alpha}_1 \ldots \boldsymbol{\alpha}_n)^{\mathrm{T}} \in \mathbb{R}^{n \times m}, \vec{\Theta}(\boldsymbol{A}; \lambda) = \left\{\vec{\Theta}(\boldsymbol{\alpha}_1; \lambda) \ldots \vec{\Theta}(\boldsymbol{\alpha}_n; \lambda)\right\}^{\mathrm{T}}$.

Then the corresponding Θ equation can be written as $\hat{\boldsymbol{B}} = \vec{\Theta}(\hat{\boldsymbol{B}} - \boldsymbol{X}^T \nabla \bar{l}(\boldsymbol{X}\hat{\boldsymbol{B}}); \lambda)$. Given $\boldsymbol{B}^\circ \in \mathbb{R}^{p \times m}$, define $\mathcal{J} = \{i : \boldsymbol{B}^\circ[i,] \neq \mathbf{0}\}$ and $\mathcal{J}^c = \{i : \boldsymbol{B}^\circ[i,] = \mathbf{0}\}$. If $i \in \mathcal{J}$, by inverting the $\vec{\Theta}$ function, we have $\boldsymbol{B}^\circ[i,]\Theta^{-1}(\|\boldsymbol{B}^\circ[i,]\|_2; \lambda)/\|\boldsymbol{B}^\circ[i,]\|_2 = \boldsymbol{B}^\circ[i,] - [\boldsymbol{X}^T \nabla \bar{l}(\boldsymbol{X}\boldsymbol{B}^\circ)][i,]$; If $i \in \mathcal{J}^c$, we have $\boldsymbol{B}^\circ[i,] = 0$, which implies $\|[\boldsymbol{X}^T \nabla \bar{l}(\boldsymbol{X}\boldsymbol{B}^\circ)][i,]\|_2 \leq \lambda$. Define a matrix $\boldsymbol{\Gamma}^\circ$ such that

$$\boldsymbol{\Gamma}^\circ[i,] = \begin{cases} \boldsymbol{B}^\circ[i,]\Theta^{-1}(\|\boldsymbol{B}^\circ[i,]\|_2; \lambda)/\|\boldsymbol{B}^\circ[i,]\|_2 - \boldsymbol{B}^\circ[i,] & \text{if } i \in \mathcal{J}, \\ \mathbf{0} & \text{if } i \in \mathcal{J}^c. \end{cases} \tag{2.20}$$

Then by introducing a slack variable $\boldsymbol{S}$ such that $\boldsymbol{X}^T \nabla \bar{l}(\boldsymbol{X}\boldsymbol{B}^\circ) + \boldsymbol{\Gamma}^\circ + \boldsymbol{S} = \mathbf{0}$, we have constraints $\boldsymbol{S}[\mathcal{J},] = \mathbf{0}$ and $\|\boldsymbol{S}\|_{2,\infty} \leq \lambda$. Finally, the corresponding SEL is:

$$\begin{aligned} &\min_{w \in \mathcal{C}, \boldsymbol{S} \in \mathbb{R}^{p \times m}} -\langle 1, \log(\boldsymbol{n}\boldsymbol{w})\rangle \text{ s.t.} \quad \boldsymbol{X}^T \boldsymbol{W} \nabla \bar{l}(\boldsymbol{X}\boldsymbol{B}^\circ) + \boldsymbol{\Gamma}^\circ + \boldsymbol{S} = \mathbf{0} \\ &\boldsymbol{S}[\mathcal{J},] = \mathbf{0}, \quad \|\boldsymbol{S}\|_{2,\infty} \leq \lambda. \end{aligned} \tag{2.21}$$

(Thresholdings on singular values) Consider the following low-rankness promoting penalty induced by P_Θ:

$$P_{\sigma,\Theta}(\boldsymbol{B}; \lambda) = \sum_{i=1}^{p \wedge m} P_{\Theta}(\sigma_i^{\boldsymbol{B}}; \lambda), \tag{2.22}$$

where $\sigma_i^{\boldsymbol{B}}, i = 1, .., p \wedge m$ are the singular values of $\boldsymbol{B}$ ordered in the descending order. As discussed in [22], the corresponding Θ function defined on a matrix is:

Definition 6. (Matrix threshold function). Given any threshold function $\Theta(\cdot; \lambda)$, its matrix version Θ^σ is defined for $\boldsymbol{B} \in \mathbb{R}^{n \times m}$ as follows

$$\Theta^\sigma(\boldsymbol{B}; \lambda) = \boldsymbol{U} \operatorname{diag}\{\Theta(\sigma_i^{\boldsymbol{B}}; \lambda)\}\boldsymbol{V}^{\mathrm{T}}$$

where $\boldsymbol{U}, \boldsymbol{V}$, and $\sigma_i^{\boldsymbol{B}}$ are obtained from the complete SVD of $\boldsymbol{B} : \boldsymbol{B} = \boldsymbol{U} \operatorname{diag}(\sigma_i^{\boldsymbol{B}})\boldsymbol{V}^{\mathrm{T}}$.

Based on the analysis in [22], the corresponding Θ-equation can be written as $\boldsymbol{B} = \Theta^{\sigma}(\boldsymbol{B} - \boldsymbol{X}^T\nabla\bar{l}(\boldsymbol{XB};\boldsymbol{Y});\lambda)$. Given $\boldsymbol{B}^{\circ} \in \mathbb{R}^{p\times m}$, let $\boldsymbol{B}^{\circ} = \boldsymbol{U}^{\circ}_{p\times p}\boldsymbol{D}^{\circ}_{p\times m}\boldsymbol{V}^{\circ T}_{m\times m}$ is the complete form SVD, where the singular values of $\boldsymbol{B}^{\circ}$ are ordered in the descending order on the diagonal part of $\boldsymbol{D}^{\circ}$. Let $\boldsymbol{U}^{\circ} = [\boldsymbol{U}^{\circ\circ}, \boldsymbol{U}^{\circ\circ}_{\perp}]$ and $\boldsymbol{V}^{\circ} = [\boldsymbol{V}^{\circ\circ}, \boldsymbol{V}^{\circ\circ}_{\perp}]$ such that $\boldsymbol{B}^{\circ} = \boldsymbol{U}^{\circ\circ}_{p\times r^{\circ}}\boldsymbol{D}[1:r^{\circ};1:r^{\circ}]^{\circ}\boldsymbol{V}^{\circ\circ T}_{m\times r^{\circ}}$ is the compact SVD, where r° be the rank of $\boldsymbol{B}^{\circ}$, then we have $\boldsymbol{D}^{\circ}[i,i] \neq \boldsymbol{0}$ if $1 \leq i \leq r^{\circ}$ while $\boldsymbol{D}^{\circ}[i,i] = \boldsymbol{0}$ if $r^{\circ}+1 \leq i \leq p \wedge m$. An important equation can be obtained by plugging $\boldsymbol{B}^{\circ}$ into the Θ-equation:

$$\begin{aligned}
\boldsymbol{D}^{\circ} &= \boldsymbol{U}^{\circ T}\Theta^{\sigma}(\boldsymbol{B}^{\circ} - \boldsymbol{X}^T\nabla\bar{l}(\boldsymbol{XB}^{\circ});\lambda)\boldsymbol{V}^{\circ} \\
&\overset{(i)}{=} \boldsymbol{U}^{\circ T}\boldsymbol{U}'\Theta^{\sigma}(\boldsymbol{D}';\lambda)\boldsymbol{V}'^{T}\boldsymbol{V}^{\circ} \\
&\overset{(ii)}{=} \boldsymbol{U}^{\circ\circ T}\boldsymbol{U}''\Theta^{\sigma}(\boldsymbol{D}';\lambda)\boldsymbol{V}''^{T}\boldsymbol{V}^{\circ\circ} \\
&\overset{(iii)}{=} \Theta^{\sigma}(\boldsymbol{D}^{\circ} - \boldsymbol{U}^{\circ\circ T}\boldsymbol{X}^T\nabla\bar{l}(\boldsymbol{XB}^{\circ})[1:r^{\circ},1:r^{\circ}]\boldsymbol{V}^{\circ\circ};\lambda),
\end{aligned} \tag{2.23}$$

where $\boldsymbol{B}^{\circ} - \boldsymbol{X}^T\nabla\bar{l}(\boldsymbol{XB}^{\circ}) = \boldsymbol{U}'\boldsymbol{D}'\boldsymbol{V}'^{T}$ is the complete SVD and $\boldsymbol{U}' = [\boldsymbol{U}'', \boldsymbol{U}''_{\perp}]$ and $\boldsymbol{V}' = [\boldsymbol{V}'', \boldsymbol{V}''_{\perp}]$ are defined similarly. (i) is based on the definition of Θ^{σ} function, (ii) holds because the lower diagonal part of $\Theta^{\sigma}(\boldsymbol{D}';\lambda)$ is zero matrix, and (iii) is based on the uniqueness of compact SVD. In fact, the corresponding columns of $\boldsymbol{U}''$ with $\boldsymbol{U}^{\circ\circ}$ and $\boldsymbol{V}''$ with $\boldsymbol{V}^{\circ\circ}$ must be the same up to any rotation/reflection applied to both sets of columns for any repeated singular values.

Based on equation (2.23), if we have $1 \leq i \leq r^{\circ}$, which implies $\boldsymbol{D}^{\circ}[i,i] \neq \boldsymbol{0}$, we can inverse the Θ^{σ} function as before to obtain $\Theta^{\sigma^{-1}}(\boldsymbol{D}^{\circ}[1:r^{\circ},1:r^{\circ}];\lambda) = \boldsymbol{D}^{\circ}[1:r^{\circ},1:r^{\circ}] - \mathrm{diag}(\sigma_i^{\boldsymbol{X}^T\nabla\bar{l}(\boldsymbol{XB}^{\circ})})[1:r^{\circ},1:r^{\circ}]$. In addition, when we have $r^{\circ}+1 \leq i \leq p\wedge m$, which implies $\boldsymbol{D}^{\circ}[i,i] = \boldsymbol{0}$, the inequality constraint can be obtained:

$$\sigma_{r^{\circ}+1}(\boldsymbol{X}^T\nabla\bar{l}(\boldsymbol{XB}^{\circ})) \leq \lambda. \tag{2.24}$$

Define a matrix $\boldsymbol{\Gamma}^{\circ}$ such that

$$\boldsymbol{\Gamma}^{\circ} = \boldsymbol{U}^{\circ\circ}(\Theta^{\sigma^{-1}}(\boldsymbol{D}^{\circ}[1:r^{\circ},1:r^{\circ}];\lambda) - \boldsymbol{D}^{\circ}[1:r^{\circ},1:r^{\circ}])\boldsymbol{V}^{\circ\circ T}. \tag{2.25}$$

We can construct a slack variable $\boldsymbol{S}$ such that $\boldsymbol{X}^T\nabla\bar{l}(\boldsymbol{XB}^{\circ}) + \boldsymbol{\Gamma}^{\circ} + \boldsymbol{S} = \boldsymbol{0}$, then the corresponding SEL is:

$$\begin{aligned}
&\min_{w\in\mathcal{C},\boldsymbol{S}\in\mathbb{R}^{p\times m}} -\langle 1, \log(\boldsymbol{nw})\rangle \ \text{ s.t.} \quad \boldsymbol{X}^T\boldsymbol{W}\nabla\bar{l}(\boldsymbol{XB}^{\circ}) + \boldsymbol{\Gamma}^{\circ} + \boldsymbol{S} = \boldsymbol{0} \\
&\boldsymbol{U}^{\circ\circ T}\boldsymbol{S} = \boldsymbol{0}, \quad \boldsymbol{S}\boldsymbol{V}^{\circ\circ} = \boldsymbol{0}, \quad \|\boldsymbol{S}\|_2 \leq \lambda.
\end{aligned} \tag{2.26}$$

Remark 2.3.1. *When the penalty P_{Θ} is ℓ_1 norm, then $\boldsymbol{\Gamma}^{\circ} = \lambda\boldsymbol{U}^{\circ\circ}\boldsymbol{V}^{\circ\circ T}$; while when the penalty is P_H, then $\Gamma^{\circ} = \boldsymbol{0}$.*

2.3.2 SEL for General Penalties or Constraints

This section we generalize the above methodology into optimization with general penalties and even constraints. First, as discussed in [21] and [22], when a thresholding function $\Theta(t;\lambda)$ is given, then the corresponding penalties can be constructed as:

$$P(t;\lambda) = P_\Theta(t;\lambda) + q(t;\lambda),$$

where q is an arbitrary function satisfying $q(t,\lambda) \geq 0$, $\forall t \in \mathbb{R}$ and $q(t;\lambda) = 0$ if $t = \Theta(s;\lambda)$ for some $s \in \mathbb{R}$. Due to the nonconvex nature of the optimization problem, it is useful to construct the following surrogate function:

$$g(\boldsymbol{B};\boldsymbol{B}^-) = \bar{l}(\boldsymbol{X}\boldsymbol{B}^-) + \langle \nabla \bar{l}(\boldsymbol{X}\boldsymbol{B}^-), \boldsymbol{B} - \boldsymbol{B}^- \rangle + \rho \|\boldsymbol{B} - \boldsymbol{B}^-\|_F^2 + P(\boldsymbol{B};\lambda). \tag{2.27}$$

In addition, as long as a step size is proper chosen $\rho > \sqrt{L}\|\boldsymbol{X}\|_2$, if we set $\boldsymbol{B}^{(t+1)} \in \operatorname{argmin}_{\boldsymbol{B}} g(\boldsymbol{B};\boldsymbol{B}^{(t)})$, the sequence of iterates satisfies

$$f(\boldsymbol{B}^{(t+1)}) \leq g(\boldsymbol{B}^{(t+1)};\boldsymbol{B}^{(t)}) \leq g(\boldsymbol{B}^{(t)};\boldsymbol{B}^{(t)}) = f(\boldsymbol{B}^{(t)}). \tag{2.28}$$

Then the optimal solutions of problem (2.50) can be characterized as:

$$\boldsymbol{B} = \Theta(\boldsymbol{B} - \boldsymbol{X}^T \nabla \bar{l}(\boldsymbol{X}\boldsymbol{B})/\rho;\lambda), \tag{2.29}$$

which correspond to the set of fixed points of the original optimization problem (2.13). Then similar techniques can be applied to construct SEL from this Θ-equation. For the constrained-type problems like

$$\min_{\boldsymbol{B}} \bar{l}(\boldsymbol{X}\boldsymbol{B};\boldsymbol{Y}) \quad \text{s.t.} \quad \|\boldsymbol{B}\|_0 \leq J, \tag{2.30}$$

the quartile thresholding $\Theta^{\#}(\cdot;q)$ is helpful. Given any $\boldsymbol{s} \in \mathbb{R}^{n\times m}$, we define $\Theta^{\#}(\boldsymbol{s};q)$ to be a vector $\boldsymbol{t}$ satisfying $t_{(j)} = s_{(j)}$ if $1 \leq j \leq q$ and 0 otherwise, where $s_{(1)},...,s_{(n)}$ are the order statistics of $s_1,...,s_n$ satisfying $|s_1| \geq ... \geq |s_n|$, and $t_{(1)},...,t_{(n)}$ are defined similarly. To avoid ambiguity, we assume no ties occur in performing $\Theta^{\#}(\boldsymbol{s};q)$. Then the optimality condition for problem (2.30) is

$$\text{vec } \boldsymbol{B} = \Theta^{\#}(\text{vec}\,(\boldsymbol{B} - \boldsymbol{X}^T \nabla \bar{l}(\boldsymbol{X}\boldsymbol{B})/\rho); J)$$

based on constructing a similar surrogate function:

$$g(\boldsymbol{B};\boldsymbol{B}^-) = \bar{l}(\boldsymbol{X}\boldsymbol{B}^-) + \langle \nabla \bar{l}(\boldsymbol{X}\boldsymbol{B}^-), \boldsymbol{B} - \boldsymbol{B}^- \rangle + \rho \|\boldsymbol{B} - \boldsymbol{B}^-\|_F^2 \text{ s.t. } \|\boldsymbol{B}\|_0 \leq J. \tag{2.31}$$

2.3.3 SEL for Sparse Reduced Rank Regression

In this section, we consider the case where the target matrix $\boldsymbol{B}$ is simultaneous sparse and low-rank. To impose low-rankness, we in addition to the penalized optimization, we add another low-rank constraint $\text{rank}(\boldsymbol{B}) \leq r$. We first consider the simplest case when the loss l is induced by the Frobenius norm. Luckily, this low-rank constraint doesn't induce the need of an above type of surrogate function when a reparametrization is applied: let $\boldsymbol{B} = \boldsymbol{S}\boldsymbol{V}^T$, where $\boldsymbol{S} \in \mathbb{R}^{p\times r}$ carries the sparsity structure of $\boldsymbol{B}$ and $\boldsymbol{V} \in \mathbb{O}^{m\times r}$ is an orthogonal matrix. Then the original optimization problem (2.13) can be reparametrized to:

$$\min_{\boldsymbol{S}\in\mathbb{R}^{p\times r},\boldsymbol{V}\in\mathbb{O}^{m\times r}} \frac{\|\boldsymbol{X}\boldsymbol{S}\boldsymbol{V}^T - \boldsymbol{Y}\|_F^2}{2} + P(\boldsymbol{S};\lambda). \tag{2.32}$$

A simple block coordinate descent algorithm for $\boldsymbol{S}$ given $\boldsymbol{V}$ and $\boldsymbol{V}$ given $\boldsymbol{S}$ can be applied to solve problem (2.32). To develop SEL for this type of parametrization, we need sample-additive estimation equations for optimizations of the two blocks. First, given the orthogonal matrix $\boldsymbol{V}$, since we have $\|\boldsymbol{X}\boldsymbol{S}\boldsymbol{V}^T - \boldsymbol{Y}\|_F^2 = \|\boldsymbol{X}\boldsymbol{S} - \boldsymbol{Y}\boldsymbol{V}\|_F^2$, the sample additive estimating equations can be derived based on the aforementioned sections with the transformed response matrix $\boldsymbol{Y}\boldsymbol{V}$. It is more tricky to derive an sample additive estimating equation for $\boldsymbol{V}$ given $\boldsymbol{S}$ with the constraint $\boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}$. We apply the gradient on Riemannian manifold of $\boldsymbol{V}$ to achieve this goal. Denote a function $F(\boldsymbol{V})$, the Euclidean gradient of $\boldsymbol{V}$ as $\nabla F(\boldsymbol{V})$ and the Stiefel manifold gradient as $\nabla F^{(s)}(\boldsymbol{V})$. Then given any matrix $\boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2 \in \mathbb{R}^{m\times r}$, the canonical metric for Stiefel manifold is defined as $g_V(\boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2) = \text{Tr}(\boldsymbol{\Delta}_1^T(\boldsymbol{I} - \frac{1}{2}\boldsymbol{V}\boldsymbol{V}^T)\boldsymbol{\Delta}_2)$. The identity induced by the metric is $\langle \nabla F(\boldsymbol{V}), \boldsymbol{\Delta}\rangle = g_V(\nabla F^{(s)}(\boldsymbol{V}), \boldsymbol{\Delta})$. Hence, we have $\nabla F(\boldsymbol{V}) = (\boldsymbol{I} - \frac{1}{2}\boldsymbol{V}\boldsymbol{V}^T)\nabla F^{(s)}$. Since $\boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}$, we have $\boldsymbol{V}^T\nabla F(\boldsymbol{V}) + \nabla F(\boldsymbol{V})^T\boldsymbol{V} = \boldsymbol{0}$. Then we can see

$$\begin{aligned}
\nabla F(\boldsymbol{V}) &= \nabla F(\boldsymbol{V}) - \frac{1}{2}\boldsymbol{V}(\boldsymbol{V}^T\nabla F(\boldsymbol{V}) + \nabla F(\boldsymbol{V})^T\boldsymbol{V}) \\
&= \nabla F(\boldsymbol{V}) - \boldsymbol{V}\boldsymbol{V}^T\nabla F(\boldsymbol{V}) + \frac{1}{2}\boldsymbol{V}\boldsymbol{V}^T\nabla F(\boldsymbol{V}) - \frac{1}{2}\boldsymbol{V}\nabla F(\boldsymbol{V})^T\boldsymbol{V} \\
&= (\boldsymbol{I} - \frac{1}{2}\boldsymbol{V}\boldsymbol{V}^T)(\nabla F(\boldsymbol{V}) - \boldsymbol{V}\nabla F(\boldsymbol{V})^T\boldsymbol{V}),
\end{aligned} \tag{2.33}$$

which implies that $\nabla F(\boldsymbol{V}) = \nabla F(\boldsymbol{V}) - \boldsymbol{V}\nabla F(\boldsymbol{V})^T\boldsymbol{V}$. Let $F_i(\boldsymbol{V}) = \|\boldsymbol{y}_i - \boldsymbol{x}_i^T\boldsymbol{S}\boldsymbol{V}^T\|_2^2/2$ such that $\|\boldsymbol{X}\boldsymbol{S}\boldsymbol{V}^T - \boldsymbol{Y}\|_F^2 = \sum_{i=1}^n F_i(\boldsymbol{V})$. Then we have $\nabla F_i(\boldsymbol{V}) = \boldsymbol{x}_i(\boldsymbol{x}_i^T\boldsymbol{S}\boldsymbol{V}^T - \boldsymbol{y}_i)$ and the sample additive

estimating equation for $\boldsymbol{V}$ given $\boldsymbol{S}$:

$$\boldsymbol{x}_i(\boldsymbol{x}_i^T\boldsymbol{S}\boldsymbol{V}^T - \boldsymbol{y}_i) - \boldsymbol{V}\boldsymbol{x}_i(\boldsymbol{x}_i^T\boldsymbol{S}\boldsymbol{V}^T - \boldsymbol{y}_i)\boldsymbol{V} = \boldsymbol{0}.$$

2.4 Asymptotic Analysis for SEL

2.4.1 SEL with Affine Inequality Constraints

Consider test with affine constraints $\boldsymbol{A}\boldsymbol{\beta} \preceq \boldsymbol{\alpha}$ with $\boldsymbol{A} \in \mathbb{R}^{k\times p}$ with rank r (k can be smaller than or greater than p). The constraints should be compatible such that the true value $\boldsymbol{\beta}^\circ$ is indeed in the feasible region of these constraints in equation (2.12). By introducing the slack variable, the null hypothesis is $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}^\circ$, such that: there is a vector $\boldsymbol{s} \in \mathbb{R}^p$, so that:

$$\begin{aligned} &\mathrm{E}[\boldsymbol{X}^T\nabla\bar{l}(\boldsymbol{X}\boldsymbol{\beta}) + \boldsymbol{s}] = \boldsymbol{0}, \\ &\boldsymbol{A}[\mathcal{J},]\boldsymbol{s} = \boldsymbol{0},\ \boldsymbol{A}[\mathcal{J}^c,]\boldsymbol{s} \preceq \boldsymbol{0}. \end{aligned} \tag{2.34}$$

However, the constraints are not identifiable so we need to do some operations to revise the constrained matrix $\boldsymbol{A}[\mathcal{J},]$ and $\boldsymbol{A}[\mathcal{J}^c,]$:

1. If $\boldsymbol{A}[\mathcal{J}^c,]\boldsymbol{s} \preceq \boldsymbol{0}$ implies some equality constraints from the solution region of $\boldsymbol{s}$ defined by only inequalities, then we put them into $\boldsymbol{A}[\mathcal{J},]\boldsymbol{s} = \boldsymbol{0}$;

2. Pick up the linear independent constrains in both equality and inequality ones, such that $\boldsymbol{A}[\mathcal{J},] \in \mathbb{R}^{r_1\times p}$ and $\mathrm{rank}(\boldsymbol{A}[\mathcal{J},]) = r_1 \le r \le \min(k,p)$; $\boldsymbol{A}[\mathcal{J}^c,] \in \mathbb{R}^{k_2\times p}$ and $\mathrm{rank}(\boldsymbol{A}[\mathcal{J}^c,]) = r_2 \le r \le \min(k,p)$. $r_1 + r_2 = r$ and $k_2 \le 2r_2$.

If $\boldsymbol{A} = -\boldsymbol{I}$ and $\boldsymbol{\alpha} = \boldsymbol{0}$, then (2.12) reduces to the non-negative regression EL; if $\mathcal{J} = \{1, 2, ..., k\}$ and $\mathcal{J}^c = \emptyset$, (2.12) reduces to the traditional regression EL. The following conditions are required for us to obtain an asymptotic analysis of large n SEL:

1. Let $\boldsymbol{\epsilon} = -\nabla\bar{l}(\boldsymbol{X}\boldsymbol{\beta}^\circ)$. Suppose $\boldsymbol{z}_i = -\boldsymbol{x}_i^T\epsilon_i$, $i = 1, 2, ..., n$ follows a distribution with mean $\boldsymbol{0}$ and covariance matrix $\boldsymbol{V}$;

2. We assume that the interior of convex hull of $\boldsymbol{z}_1, ..., \boldsymbol{z}_n$ contains zero with probability 1 as $n \to \infty$;

3. The minimum and maximum eigenvalues of the matrix $\boldsymbol{V}$ is bounded below and above by constants;

Then we have the following theorem to characterize the asymptotic distribution of large n SEL under affine inequality constraints:

Theorem 2.4.1. *Suppose the above conditions* 1-3 *hold. Then under the inequality constraints* $\boldsymbol{A}\boldsymbol{\beta} \preceq \boldsymbol{\alpha}$ *with* $\boldsymbol{A} \in \mathbb{R}^{k\times p}$ *and rank*$(\boldsymbol{A}) = r$*, the SEL statistics for test* (2.34) *follows a chi-bar-square distribution as follows:*

$$\begin{aligned} &P(c_n(\boldsymbol{\beta}^\circ) \geq t) = \sum_{j=|\mathcal{J}|}^{p} P(\chi_j^2 \geq t)\delta_j, \\ &\delta_j = \sum_{|\mathcal{I}|=j} P(\boldsymbol{\nu}_{\boldsymbol{A}_{\mathcal{I}-\mathcal{J}}} \preceq \mathbf{0}) P(\boldsymbol{\nu}_{\boldsymbol{A}_{\mathcal{I}^c}} \,|\, \boldsymbol{\nu}_{\boldsymbol{A}_{\mathcal{I}}} \preceq \mathbf{0}), \text{ with } p > |\mathcal{I}|,\ \mathcal{J} \subsetneq \mathcal{I}, \\ &\delta_{|\mathcal{J}|} = 1 - \sum_{j=|\mathcal{J}|+1}^{p} \delta_j, \end{aligned} \tag{2.35}$$

where we denote $\sqrt{n}\sum_{i=1}^{n} \frac{1}{n}\boldsymbol{z}_{\mathcal{H}i} \xrightarrow{d} \boldsymbol{\nu}_{A_{\mathcal{H}}}$ *for any index set* $\mathcal{H}$ *and* $\boldsymbol{z}_{\mathcal{H}i}$ *is the subvector of* $\boldsymbol{z}_i$ *indexed by* $\mathcal{H}$*.*

Since $|\mathcal{J}| = r_1$ which is the rank of all the independent equality constraints, for $\boldsymbol{A} = -\boldsymbol{I}$ and $\mathcal{J} \subset \{1, 2, ..., p\}$ and $\mathcal{J}^c \neq \emptyset$, then $c_n(\boldsymbol{\beta}^\circ)$ reduces to the nonnegative regression EL; while if $\mathcal{J} = \{1, 2, ..., p\}$ and $\mathcal{J}^c = \emptyset$, then $c_n(\boldsymbol{\beta}^\circ)$ reduces to the traditional regression EL with χ_p^2 distribution. We provide some numeric evidence that the distribution of SEL for large n cases with affine inequality constraints is indeed in a chi-bar-square form. In particular, we show that if only a single chi-square distribution is applied to generate p-values SEL, then the corresponding tests will be either too optimistic or conservative.

Two specific examples are considered: nonnegative regressions and regressions with variable sections. Suppose the data generating model is $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^\circ + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ follow a normal distribution with covariance $\sigma^2\boldsymbol{I}$. Let $n = 100$, $p = 2$, $\sigma = 1$, $\boldsymbol{\beta}^\circ = (5, 0)^T$, $X_{ij} \sim \mathcal{N}(0, 1)$ for all $1 \leq i \leq n$ and $1 \leq j \leq p$. The simulation settings of the two cases are as follows:

1. Nonnegative regression: The p-value of SEL test is generated from a mixture of chi-square distribution with degrees of freedom one and two, where the weights for χ_2^2 and χ_1^2 are both 0.5;

2. Regression with variable selection: Let $\lambda = \sigma\sqrt{\log p/n}$. The p-value of SEL test is generated from a mixture of chi-square distribution with degrees of freedom one and two, where the weight for χ_2^2 is the probability that the maximum of 100 standard Gaussian random variables is greater than λ, while the weight for χ_1^2 can be obtained by one minus the above weight for χ_2^2;

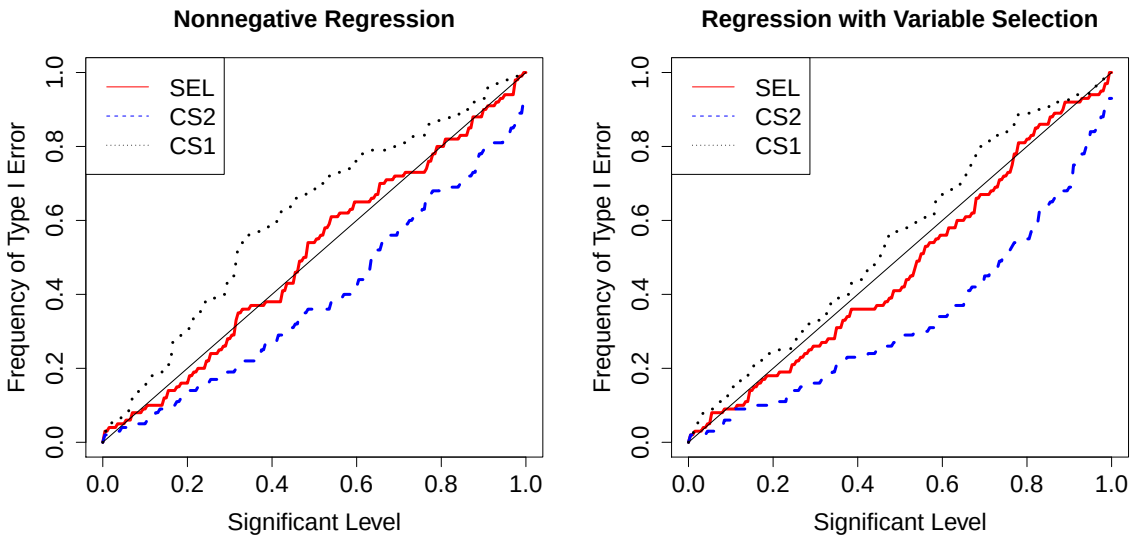

Figure 2.1: Percentage of the type I error made by 100 hypothesis tests on for $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}^\circ$ v.s. the respective significant level by large n SEL.

The simulations are repeated for 100 times where we test the hypothesis $\boldsymbol{\beta} = \boldsymbol{\beta}^\circ$ in both cases and record all the p-values of SEL tests. For comparison, we also record all the p-values when we 'assume' the distribution of SEL test is a single χ^2_2 or χ^2_1. Finally, by ranging the significant level α from zero to one, we show how the frequencies of the type I error change with respect to the significant level. The simulation results are shown in the following figure 2.1, where 'CS1' and 'CS2' stand for the testing results when the $p-$ values are generated from single χ^2_1 and χ^2_2 distributions. From the figure, we can see the validity of our proposed theorem for large n SEL.

2.5 Analysis of Large p SEL

Recall that our defined large p SEL contains a bias term $\boldsymbol{\gamma}^\circ$, which can be obtained directly from the null hypothesis $\boldsymbol{\beta}^\circ$. Therefore, the null hypothesis can be refined and the bias term can be eliminated. Finally the making inference for large p SEL corresponds the following estimating equation:

$$\boldsymbol{\beta} = \boldsymbol{\Theta}_H(\boldsymbol{\beta} - \boldsymbol{X}^T \nabla \bar{l}(\boldsymbol{X}\boldsymbol{\beta}^\circ)/n; \lambda), \tag{2.36}$$

where $\boldsymbol{\Theta}_H$ is the hard-thresholding operator. Unfortunately, when $p > n$, it is impossible to directly learn the marginal limiting distribution of the empirical likelihood ratio as in the lower dimensional

case. However, we can still make practical inference conditional on the selection event of the given data $\boldsymbol{X}, \boldsymbol{y}$. Some notations are worth to mention: let $\boldsymbol{\alpha} \succeq \boldsymbol{\beta}$ mean all components of $\boldsymbol{a}$ is greater than the respective component of $\boldsymbol{\beta}$. For a matrix $\boldsymbol{A}$ and an index set $\mathcal{J}$, $\boldsymbol{A}_{\mathcal{J}}$ means the the submatrix of $\boldsymbol{A}$ with columns indexed by $\mathcal{J}$ and $\boldsymbol{A}_{\mathcal{J},\mathcal{I}}$ is denoted as the submatrix of $\boldsymbol{A}$ with column indexed by $\mathcal{I}$ and row indexed by $\mathcal{J}$.

2.5.1 A Limiting Result for Large p SEL for Regression

In this section, we consider the example of sparse regressions to illustrate our analysis of SEL. For high dimensional regression with P_H penalty, (2.43) corresponds to the null hypothesis $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}^\circ$, and there is a vector $\boldsymbol{s} \in \mathbb{R}^p$, so that:

$$\begin{aligned} &\mathrm{E}[\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{\beta}^\circ - \boldsymbol{y}) + \boldsymbol{s}] = \boldsymbol{0}, \\ &\boldsymbol{s}_{\mathcal{J}} = \boldsymbol{0},\ \|\boldsymbol{s}\|_\infty \le \lambda. \end{aligned} \tag{2.37}$$

The corresponding regression large p SEL is:

$$\begin{aligned} \min_{\boldsymbol{w}\in\mathcal{C},\boldsymbol{s}} -\langle \boldsymbol{1}, \log(n\boldsymbol{w})\rangle \quad \text{s.t.} \quad & \boldsymbol{X}^T\boldsymbol{W}\nabla\bar{l}(\boldsymbol{X}\boldsymbol{\beta}^\circ) + \boldsymbol{s} = \boldsymbol{0} \\ & \boldsymbol{s} \circ \boldsymbol{\beta}^\circ = \boldsymbol{0},\ \|\boldsymbol{s}\|_\infty \le \lambda. \end{aligned} \tag{2.38}$$

We consider the case when the selection is consistent, where we suppose given $\boldsymbol{X}, \boldsymbol{y}$ and λ, there exists a $\hat{\boldsymbol{\beta}}$ satisfying estimating equation (2.36), with nonzero index set $\mathcal{J} = \{j : \hat{\beta}_j \neq 0\}$ the same with the truth $\boldsymbol{\beta}^\circ$. We first provide some equivalent characterizations of the selection event, where we can obtain a sparse solution $\hat{\boldsymbol{\beta}}$ of estimating equation (2.36) given $\boldsymbol{X}, \boldsymbol{y}$ and λ with nonzero index set $\mathcal{J}$. Clearly, by directly inverting the estimating equation, we have $\hat{\boldsymbol{\beta}}_{\mathcal{J}} = (\boldsymbol{X}_{\mathcal{J}}^T\boldsymbol{X}_{\mathcal{J}})^{-1}\boldsymbol{X}_{\mathcal{J}}^T\boldsymbol{y}$ while $\hat{\boldsymbol{\beta}}_{\mathcal{J}^c} = \boldsymbol{0}$. In addition, the following inequality constraints are required:

$$\begin{cases} |\hat{\boldsymbol{\beta}}_{\mathcal{J}}| \succeq \lambda\boldsymbol{1} \\ \|\boldsymbol{X}_{\mathcal{J}^c}^T(\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{y})\|_\infty < n\lambda, \end{cases} \iff \begin{cases} |(\boldsymbol{X}_{\mathcal{J}}^T\boldsymbol{X}_{\mathcal{J}})^{-1}\boldsymbol{X}_{\mathcal{J}}^T\boldsymbol{y}| \succeq \lambda\boldsymbol{1} \\ \|\boldsymbol{X}_{\mathcal{J}^c}^T(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}_{\mathcal{J}}})\boldsymbol{y}\|_\infty < n\lambda, \end{cases} \tag{2.39}$$

where $\boldsymbol{P}_{\boldsymbol{X}_{\mathcal{J}}} = \boldsymbol{X}_{\mathcal{J}}(\boldsymbol{X}_{\mathcal{J}}^T\boldsymbol{X}_{\mathcal{J}})^{-1}\boldsymbol{X}_{\mathcal{J}}^T$ is a projection matrix induced by $\boldsymbol{X}_{\mathcal{J}}$ and $\boldsymbol{1}$ is a vector with all ones. Furthermore, under the knowledge of null hypothesis, we'd like to represent the above constraints into constraints of $\boldsymbol{\epsilon}$, which is

$$\begin{cases} |\boldsymbol{\beta}_{\mathcal{J}}^\circ + (\boldsymbol{X}_{\mathcal{J}}^T\boldsymbol{X}_{\mathcal{J}})^{-1}\boldsymbol{X}_{\mathcal{J}}^T\boldsymbol{\epsilon}| \succeq \lambda\boldsymbol{1} \\ \|\boldsymbol{X}_{\mathcal{J}^c}^T(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}_{\mathcal{J}}})\boldsymbol{\epsilon}\|_\infty < n\lambda. \end{cases} \tag{2.40}$$

Equation (2.40) is helpful to to show the equivalence between prime and dual problem of SEL.

Based on the above analysis, we propose the following assumptions for out main theorem:

1. Suppose $\boldsymbol{\beta}^\circ$ is sparse with nonzero index $\mathcal{J} = \{j : \beta_j^\circ \neq 0\}$, $\min_{j\in\mathcal{J}} |\boldsymbol{\beta}_j^\circ| > (1+\kappa)\lambda$ for a constant $\kappa > 0$, and $|\mathcal{J}| < n$. Assume $\boldsymbol{\epsilon} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^\circ$. Let $\boldsymbol{z}_i = -\boldsymbol{x}_i^T\epsilon_i$, and $\boldsymbol{z}_1, ...\boldsymbol{z}_n$ follows a distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{V}$. In addition, the minimum and maximum eigenvalues of the matrix $\boldsymbol{V}$ is bounded below and above by constants;

2. Given $\boldsymbol{X}, \boldsymbol{y}, \lambda$, there exists vector $\hat{\boldsymbol{\beta}}$, such that it satisfies the estimating equation (2.36), and the nonzero index set is also $\mathcal{J}$.

3. Let $\boldsymbol{z}_{Ji}$ be the subvector of $\boldsymbol{z}_i$ indexed by $\mathcal{J}$. We assume that the interior of the convex hull of $\boldsymbol{z}_{\mathcal{J}1}, ..., \boldsymbol{z}_{\mathcal{J}n}$ contains zero with probability 1 as $n \to \infty$;

4. The minimum and maximum eigenvalues of the matrix $\boldsymbol{X}_{\mathcal{J}}^T\boldsymbol{X}_{\mathcal{J}}/n$ is bounded below and above by constants;

5. There exists $\delta > 0$, such that $\|\boldsymbol{X}_{\mathcal{J}^c}^T\boldsymbol{X}_{\mathcal{J}}(\boldsymbol{X}_{\mathcal{J}}^T\boldsymbol{X}_{\mathcal{J}})^{-1}\|_\infty \leq \delta$, and $\delta n^{1/2} = o(\lambda)$.

To make practical inference for high dimensional regression SEL, we propose the following theorem:

Theorem 2.5.1. *Under the above assumptions 1-5, denote the variable* $\boldsymbol{z}_\lambda^{\boldsymbol{X},\boldsymbol{\beta}^\circ}$ *as*

$$(\sum_{i=1}^n \boldsymbol{z}_{\mathcal{J}i})^T(\sum_{i=1}^n \boldsymbol{z}_{\mathcal{J}i}\boldsymbol{z}_{\mathcal{J}i}^T)^{-1}(\sum_{i=1}^n \boldsymbol{z}_{\mathcal{J}i}). \tag{2.41}$$

Then if as n *and* $p \to \infty$, $\lambda \to 0$, *there exist constants* c_1, c_2 *such that with probability converging to* 1, *the optimal SEL ratio* $c(\boldsymbol{\beta}^\circ) = -2\langle \mathbf{1}, \log(n\hat{\boldsymbol{w}})\rangle$ *satisfies*

$$P(c(\boldsymbol{\beta}^\circ) \geq t) = \quad P(\boldsymbol{z}_\lambda^{\boldsymbol{X},\boldsymbol{\beta}^\circ} \geq t) + o_p(1), \tag{2.42}$$

given any positive value t, *where* $\hat{\boldsymbol{w}}$ *is the optimizer of weights in problem* (2.43).

Our theory renders a limiting behavior of the distribution of $c_n(\boldsymbol{\beta}^\circ)$ when section consistency can be obtained. However, signal strength assumptions is unnecessary since the information is already incorporated into the null hypothesis. We demonstrate the validity of theorem 2.5.1, the limiting theory of large p SEL. We set λ to be large enough such that the assumptions of the main theorem can be satisfied. The details of simulation settings are as follows:

1. Case 1: let $n = 100$, $p = 200$, $\sigma = 1$, $\boldsymbol{\beta}^\circ = (5, 0, ..., 0)^T$, $X_{ij} \sim \mathcal{N}(0, 1)$ for all $1 \leq i \leq n$ and $1 \leq j \leq p$, $\lambda = \sigma\sqrt{2\log p}$;

2. Case 2: let $n = 100$, $p = 200$, $\sigma = 2.5$, $\boldsymbol{\beta}^\circ = (5, 0, ..., 0)^T$, $X_{ij} \sim \mathcal{N}(0, 1)$ for all $1 \leq i \leq n$ and $1 \leq j \leq p$, $\lambda = \sigma\sqrt{2\log p}$;

3. Case 3: let $n = 200$, $p = 500$, $\sigma = 1$, $\boldsymbol{\beta}^\circ = (5, 0, ..., 0)^T$, $X_{ij} \sim \mathcal{N}(0, 1)$ for all $1 \leq i \leq n$ and $1 \leq j \leq p$, $\lambda = \sigma\sqrt{2 \log p}$;

4. Case 4: let $n = 100$, $p = 200$, $\sigma = 1$, $\boldsymbol{\beta}^\circ = (5, 5, 0, ..., 0)^T$, $X_{ij} \sim \mathcal{N}(0, 1)$ for all $1 \leq i \leq n$ and $1 \leq j \leq p$, $\lambda = \sigma\sqrt{2 \log p}$.

The simulation is repeated for 100 times where we still test the hypothesis $\boldsymbol{\beta} = \boldsymbol{\beta}^\circ$ and record all the P-values. Finally, by ranging the significant level α from zero to one, we show how the frequencies of the type I error change with respect to the significant level in the following Figure 2.2. From the figure, we can see that the type I error can be successfully controlled by the respective significant levels when the selection consistency can happen, which implies that our main theorem is effective in a large range of settings in high dimensional inference problem.

To make practical inference in real applications, we consider the case of SEL where the λ is fixed in the following section.

2.5.2 Nonasymptotic Analysis of Large p SEL for a Fixed λ

In this section, we consider large p SEL with a general loss function and a fixed λ:

$$\begin{aligned} \min_{\boldsymbol{w} \in \mathcal{C}, \boldsymbol{s}} -\langle \mathbf{1}, \log(n\boldsymbol{w}) \rangle \quad \text{s.t.} \quad & \boldsymbol{X}^T \boldsymbol{W} \nabla \bar{l}(\boldsymbol{X}\boldsymbol{\beta}^\circ) + \boldsymbol{s} = \mathbf{0} \\ & \boldsymbol{s} \circ \boldsymbol{\beta}^\circ = \mathbf{0}, \ \|\boldsymbol{s}\|_\infty \leq \lambda. \end{aligned} \tag{2.43}$$

We show that the optimal slack empirical likelihood ratio follows a specific distribution given the data when λ is fixed. For notation ease, denote $\boldsymbol{z}_i = [\boldsymbol{z}_{\mathcal{J}i}^T, \boldsymbol{z}_{\mathcal{J}^c i}^T]^T = -\boldsymbol{x}^T \epsilon_i$ where $\boldsymbol{z}_{\mathcal{J}i} = -\boldsymbol{x}_{\mathcal{J}}^T \epsilon_i$ and $\boldsymbol{z}_{\mathcal{J}^c i} = -\boldsymbol{x}_{\mathcal{J}^c}^T \epsilon_i$. We assume that the following condition hold:

1. The optimization (2.43) is feasible: there exists a $\boldsymbol{w}^0 = \{w_1^0, ..., w_n^0\}$ such that $\sum_{i=1}^n w_i^0 \boldsymbol{z}_{\mathcal{J}i} = \mathbf{0}$, $-\lambda \mathbf{1} \preceq \sum_{i=1}^n w_i^0 \boldsymbol{z}_{\mathcal{J}^c i} \preceq \lambda \mathbf{1}$, and $0 < w_i' < 1$ for $i = 1, 2, ..., n$.

The above assumption guarantees that the Slater's conditions hold, therefore, dual optimization can render an equivalent characterization of the prime problem.

To obtain the Lagrange dual function, first, by introducing Lagrange multipliers $\phi \in \mathbb{R}$, $\boldsymbol{u}_0 \in \mathbb{R}^{|\mathcal{J}|}$, $\boldsymbol{h} \in \mathbb{R}^p$ and $\boldsymbol{v}^-, \boldsymbol{v}^+ \in \mathbb{R}^{|\mathcal{J}^c|}$, the optimization problem can be formulated to the following Lagrangian form:

$$\begin{aligned} T(\boldsymbol{w}, \boldsymbol{s}, \phi, \boldsymbol{h}, \boldsymbol{u}_0, \boldsymbol{v}^-, \boldsymbol{v}^+) = -\langle \mathbf{1}, \log(n\boldsymbol{w}) \rangle + \phi(\sum_{i=1}^n w_i - 1) + \boldsymbol{h}^T(-\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{\epsilon} + \boldsymbol{s}) \\ + \boldsymbol{u}_0^T \boldsymbol{s}_{\mathcal{J}} + \boldsymbol{v}^{-T}(-\boldsymbol{s}_{\mathcal{J}^c} - \lambda \mathbf{1}) + \boldsymbol{v}^{+T}(\boldsymbol{s}_{\mathcal{J}^c} - \lambda \mathbf{1}), \end{aligned} \tag{2.44}$$

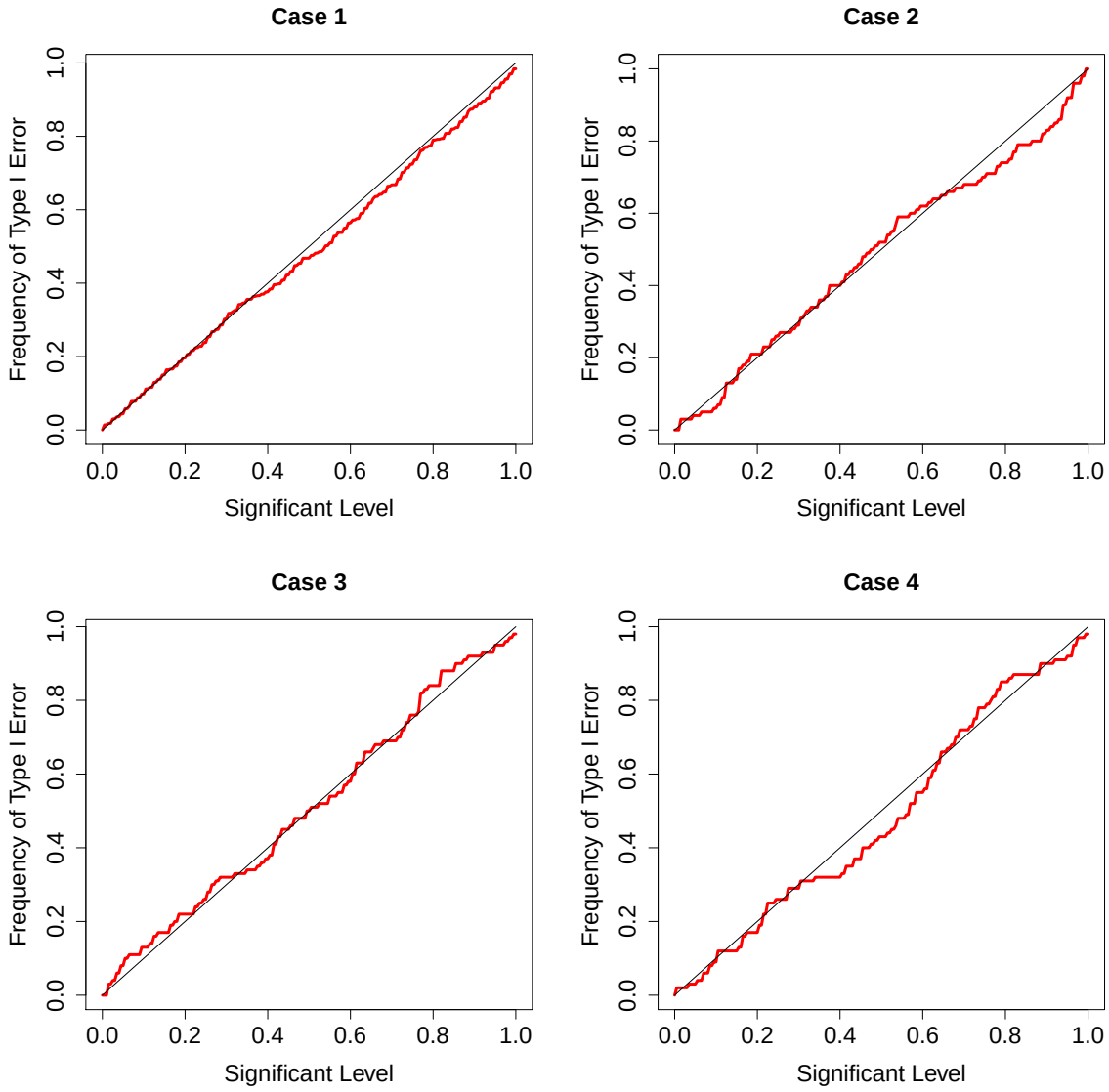

Figure 2.2: Percentage of the type I error made by 100 hypothesis tests on for $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}^\circ$ v.s. the respective significant level by conditional inference of SEL.

and the Lagrange dual function is:

$$d(\phi, \boldsymbol{h}, \boldsymbol{u}_0, \boldsymbol{v}^-, \boldsymbol{v}^+) = \min_{\boldsymbol{w}, \boldsymbol{s}} T(\boldsymbol{w}, \boldsymbol{s}, \phi, \boldsymbol{h}, \boldsymbol{u}_0, \boldsymbol{v}^-, \boldsymbol{v}^+), \tag{2.45}$$

which is expected to be maximized. Taking derivative of the Lagrangian T with respect to $\boldsymbol{w}$ and $\boldsymbol{s}$, we have

$$\begin{cases} -\dfrac{1}{\hat{w}_i} + \boldsymbol{h}^T \boldsymbol{z}_i + \phi = 0, \\ \boldsymbol{h}_{\mathcal{J}} = -\boldsymbol{u}_0, \\ \boldsymbol{h}_{\mathcal{J}^c} = \boldsymbol{v}^- - \boldsymbol{v}^+. \end{cases} \tag{2.46}$$

Plugging equation (2.72) into the Lagrangian to get the Lagrange dual function, we have:

$$\begin{aligned} d(\phi, \boldsymbol{h}, \boldsymbol{u}_0, \boldsymbol{v}^-, \boldsymbol{v}^+) &= -\sum_{i=1}^{n} \log(\frac{n}{\phi + \boldsymbol{h}^T \boldsymbol{z}_i}) + \phi(\sum_{i=1}^{n} \frac{1}{\phi + \boldsymbol{h}^T \boldsymbol{z}_i} - 1) \\ &\quad + \sum_{i=1}^{n} \frac{\boldsymbol{h}^T \boldsymbol{z}_i}{\phi + \boldsymbol{h}^T \boldsymbol{z}_i} - \boldsymbol{v}^{+T} \lambda \mathbf{1} - \boldsymbol{v}^{-T} \lambda \mathbf{1}, \\ &= -\sum_{i=1}^{n} \log(\frac{n}{\phi + \boldsymbol{h}^T \boldsymbol{z}_i}) + n - \phi - \boldsymbol{v}^{+T} \lambda \mathbf{1} - \boldsymbol{v}^{-T} \lambda \mathbf{1}. \end{aligned} \tag{2.47}$$

Note that this dual problem implicitly require that the constraints $\phi + \boldsymbol{h}^T \boldsymbol{z}_i > 0$ for $i = 1, ..., n$, this can be guarantee by the prespecified condition and the first equation of (2.72). To maximize equation (2.73), taking derivatives with respect to $\boldsymbol{u}_0$ and ϕ, we have:

$$\begin{cases} \nabla_{\boldsymbol{u}_0} d(\phi, \boldsymbol{h}, \boldsymbol{u}_0, \boldsymbol{v}^-, \boldsymbol{v}^+) &= \sum_{i=1}^{n} \dfrac{-\boldsymbol{z}_{\mathcal{J}i}}{\phi + \boldsymbol{h}^T \boldsymbol{z}_i} = \mathbf{0}, \\ \nabla_{\phi} d(\phi, \boldsymbol{h}, \boldsymbol{u}_0, \boldsymbol{v}^-, \boldsymbol{v}^+) &= \sum_{i=1}^{n} \dfrac{1}{\phi + \boldsymbol{h}^T \boldsymbol{z}_i} - 1 = 0. \end{cases} \tag{2.48}$$

In addition, taking derivatives with respect to $\boldsymbol{v}^-$ and $\boldsymbol{v}^+$, we have:

$$\begin{cases} \nabla_{\boldsymbol{v}^-} d(\phi, \boldsymbol{h}, \boldsymbol{u}_0, \boldsymbol{v}^-, \boldsymbol{v}^+) &= \sum_{i=1}^{n} \dfrac{\boldsymbol{z}_{\mathcal{J}^c i}}{\phi + \boldsymbol{h}^T \boldsymbol{z}_i} - \lambda \mathbf{1}, \\ \nabla_{\boldsymbol{v}^+} d(\phi, \boldsymbol{h}, \boldsymbol{u}_0, \boldsymbol{v}^-, \boldsymbol{v}^+) &= \sum_{i=1}^{n} \dfrac{-\boldsymbol{z}_{\mathcal{J}^c i}}{\phi + \boldsymbol{h}^T \boldsymbol{z}_i} - \lambda \mathbf{1}. \end{cases} \tag{2.49}$$

Not all components of equation (2.49) can be set to zeros since the dual optimization is limited to $\boldsymbol{v}^- \succeq \mathbf{0}$ and $\boldsymbol{v}^+ \succeq \mathbf{0}$. However, we always have

$$\begin{cases} \boldsymbol{v}^{-T} (\sum_{i=1}^{n} \dfrac{\boldsymbol{z}_{\mathcal{J}^c i}}{\phi + \boldsymbol{h}^T \boldsymbol{z}_i} - \lambda \mathbf{1}) = 0, \\ \boldsymbol{v}^{+T} \sum_{i=1}^{n} (\dfrac{-\boldsymbol{z}_{\mathcal{J}^c i}}{\phi + \boldsymbol{h}^T \boldsymbol{z}_i} - \lambda \mathbf{1}) = 0, \end{cases} \tag{2.50}$$

since either the components of $\boldsymbol{v}^-$ and $\boldsymbol{v}^+$ or the corresponding components of gradients $\nabla_{\boldsymbol{v}^-} d$ and $\nabla_{\boldsymbol{v}^+} d$ are zeros when maximizing the dual function. Then by combining equation (2.72), (2.48) and (2.50), we have:

$$\phi + \boldsymbol{v}^{+T}\lambda\mathbf{1} + \boldsymbol{v}^{-T}\lambda\mathbf{1} - n = 0, \tag{2.51}$$

which gives $\phi = n - \boldsymbol{v}^{+T}\lambda\mathbf{1} - \boldsymbol{v}^{-T}\lambda\mathbf{1}$ and implies that $\hat{w}_i = 1/(n - \boldsymbol{u}_0^T\boldsymbol{z}_{\mathcal{J}i} + \boldsymbol{v}^{+T}(-\boldsymbol{z}_{\mathcal{J}^c i} - \lambda\mathbf{1}) + \boldsymbol{v}^{-T}(\boldsymbol{z}_{\mathcal{J}^c i} - \lambda\mathbf{1}))$.

Next, based on the form of $\hat{\boldsymbol{w}}$, we show that the obtained optimal $\hat{\boldsymbol{w}}$ by solving SEL can indeed define a unique distribution. Note that the optimization (2.43) can be written as the following equivalent form:

$$\min_{\boldsymbol{w}\in\mathcal{C}} -\langle\mathbf{1}, \log(n\boldsymbol{w})\rangle \quad \text{s.t.} \quad \sum_{i=1}^{n} w_i\boldsymbol{z}_{\mathcal{J}i} = \mathbf{0}, \quad -\lambda\mathbf{1} \le \sum_{i=1}^{n} w_i\boldsymbol{z}_{\mathcal{J}^c i} \le \lambda\mathbf{1}. \tag{2.52}$$

Then the following theorem indicates that through dual optimization, the induced distribution of $\hat{\boldsymbol{w}}$ is unique.

Theorem 2.5.2. *Denote $\hat{\boldsymbol{w}} = \{\hat{w}_1, ..., \hat{w}_n\}$ such that*

$$\hat{w}_i = \frac{1}{n - \boldsymbol{u}_0^T\boldsymbol{z}_{\mathcal{J}i} + \boldsymbol{v}^{-T}(\boldsymbol{z}_{\mathcal{J}^c i} - \lambda) + \boldsymbol{v}^{+T}(-\boldsymbol{z}_{\mathcal{J}^c i} - \lambda)}$$

for some $\boldsymbol{u}_0 \in \mathbb{R}^{|\mathcal{J}|}$, and $\boldsymbol{v}^-, \boldsymbol{v}^+ \in \mathbb{R}^{|\mathcal{J}^c|}$ with $\boldsymbol{v}^- \succeq \mathbf{0}$ and $\boldsymbol{v}^+ \succeq \mathbf{0}$, then for any other $\boldsymbol{w}' = \{w_1', ..., w_n'\}$ that is different from $\hat{\boldsymbol{w}}$, such that $\sum_{i=1}^n w_i'\boldsymbol{z}_{\mathcal{J}i} = \mathbf{0}$, $-\lambda\mathbf{1} \preceq \sum_{i=1}^n w_i'\boldsymbol{z}_{\mathcal{J}^c i} \preceq \lambda\mathbf{1}$, and $0 < w_i' < 1$ for $i = 1, 2, ..., n$, we have

$$-\langle\mathbf{1}, \log(n\boldsymbol{w}')\rangle < -\langle\mathbf{1}, \log(n\hat{\boldsymbol{w}})\rangle.$$

Since the optimizer $\hat{\boldsymbol{w}}$ is defined as a unique distribution according to the data, the optimal value $c_0(\boldsymbol{\beta}^\circ) = -2\langle\mathbf{1}, \log(n\boldsymbol{w})\rangle$ follows a specific distribution, which can be applied to make practical nonparametric inference. In order to calculate p-value, we can apply bootstrap to generate psedo-samples of $\boldsymbol{x}$ and $\boldsymbol{y}$, and then apply SEL optimization to calculate the corresponding $c_0(\boldsymbol{\beta}^\circ)$ for the samples. Therefore, the distribution of $c_0(\boldsymbol{\beta}^\circ)$ can be obtained and the nonparametric inference can be applied. In the following section, we show some numeric evidence to support our main theorems and demonstrate that our SEL approach can indeed provide robustification for some statistical inference problems.

2.6 Experiments

2.6.1 Simulation Studies

In this part, we compare our nonparametric inference method with other approaches that rely on the correct model and distribution specification. Throughout the simulations, all SEL optimizations are solved by *CVXR* package in R. In particular, two circumstances are considered:

(Estimating sparse multivariate mean) Let the data generating model be $\boldsymbol{y} = \boldsymbol{\beta}^\circ + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon}$ follows a specific distribution. Let $\boldsymbol{y}_1, ..., \boldsymbol{y}_n \in \mathbb{R}^p$ are i.i.d observations of responses with respect to $\boldsymbol{\epsilon}_1, ..., \boldsymbol{\epsilon}_n$. Suppose that $\boldsymbol{\beta}^\circ$ is sparse with nonzero index set $\mathcal{J}$. Based on the estimating equation, given any λ, we first calculate the hard thresholding estimator with:

$$\begin{cases} \hat{\beta}_i = \frac{1}{n}\sum_{k=1}^n y_{ki}, & \text{if } |\frac{1}{n}\sum_{k=1}^n y_{ki}| \geq \lambda, \\ \hat{\beta}_i = 0, & \text{if } |\frac{1}{n}\sum_{k=1}^n y_{ki}| < \lambda, \end{cases} \tag{2.53}$$

where y_{ki} is the ith component of $\boldsymbol{y}_k$. Suppose this hard thresholding estimator has nonzero index set $\mathcal{I}$ with $\mathcal{J} \subset \mathcal{I}$ and $|\mathcal{I}| < n$, then clearly, the selection event induced by $\hat{\boldsymbol{\beta}}$ is

$$\min_{i\in\mathcal{I}} |\beta_i^\circ + \frac{1}{n}\sum_{k=1}^n \epsilon_{ki}| \geq \lambda, \quad \text{and } \max_{i\in\mathcal{I}^c} |\frac{1}{n}\sum_{k=1}^n \epsilon_{ki}| < \lambda. \tag{2.54}$$

When the distribution of the noise $\boldsymbol{\epsilon}$ is assumed to be normal with mean zero and variance $\sigma^2\boldsymbol{I}$, the exact inference for components $\mathcal{I}-\mathcal{J}$ can be made by drawing samples from Gaussian distributions $\mathcal{N}(0,\sigma^2)$ lower truncated at λ with 0.5 probability and $\mathcal{N}(0,\sigma^2)$ upper truncated at $-\lambda$ with 0.5 probability. Therefore, we can calculate the converge probability of the truth by constructing confidence intervals of components indexed by $\mathcal{I}-\mathcal{J}$ through the above exact distribution.

On the other hand, the respective SEL to test $\boldsymbol{\beta}^\circ$ is defined as follows:

$$\begin{aligned} \min_{\boldsymbol{w}\in\mathcal{C},\boldsymbol{s}} -\langle \mathbf{1}, \log(n\boldsymbol{w})\rangle \quad \text{s.t.} \quad & \sum_{i=1}^n w_i(\boldsymbol{\beta}^\circ - \boldsymbol{y}_i) + \boldsymbol{s} = \mathbf{0} \\ & \boldsymbol{s}\circ\boldsymbol{\beta}^\circ = \mathbf{0},\ \|\boldsymbol{s}\|_\infty \leq \lambda. \end{aligned} \tag{2.55}$$

In addition, the P-value of the optimal SEL ratio $c(\boldsymbol{\beta}^\circ) = -2\langle \mathbf{1}, \log(n\hat{\boldsymbol{w}})\rangle$ can be constructed from bootstrap by resampling $\boldsymbol{\epsilon} = \boldsymbol{y} - \boldsymbol{\beta}^\circ$. We consider the cases where $n = 100$, $p = 200$ and $n = 200$, $p = 500$, the truth $\boldsymbol{\beta}^\circ = [5, 0, ..., 0]^T$, $\lambda = \sqrt{\log p/n}$ and three types of distribution of $\boldsymbol{\epsilon}$:

1. Case 1: $\boldsymbol{\epsilon} \sim \mathcal{N}(0, 3)$;
2. Case 2: $\boldsymbol{\epsilon} \sim \mathcal{N}(0, 3)$ with probability 0.9 while $\boldsymbol{\epsilon} \sim T$ distribution with $df = 3$ (which also has a variance of 3) with probability 0.1;

Table 2.1: Comparisons between SEL and exact Gaussian inference

	$n = 100$, $p = 200$			$n = 200$, $p = 500$			
Unit %	Case 1	Case 2	Case 3	Case 1	Case 2	Case 3	
$C = 0.90$	89.60	89.31	89.55	85.77	88.28	86.97	SEL
	89.96	80.50	68.25	87.78	87.37	54.51	GEI
$C = 0.95$	95.30	96.33	94.18	93.99	93.79	92.99	SEL
	94.98	87.42	80.50	93.79	92.38	61.52	GEI
$C = 0.99$	99.33	98.11	99.47	98.60	98.20	97.59	SEL
	99.19	95.60	68.25	98.60	96.59	54.51	GEI

3. Case 3: $\boldsymbol{\epsilon} \sim T$ distribution with $df = 3$.

Therefore, for Gaussian exact inference, the noise specifications vary from correct to contaminated and finally totally misspecified. We show the coverage probability of the constructed confidence intervals at different confidence levels $C = 0.90, 0.95$ and 0.99 in Table 2.1 for both our SEL method and the Gaussian exact inference (GEI) approach. From the tables, we can see that the exact inference by assuming Gaussian noise will easily fall even if there is only 10% contamination of the noises, however, our SEL approach can keep the coverage level even if the distribution is totally misspecified. The simulation results demonstrates the robustness of our SEL method with repect to the distribution assumptions of the effective noise.

(High dimensional regression) Consider the regression model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^\circ + \boldsymbol{\epsilon}$ where β° is sparse with nonzero index $\mathcal{J}$ and the target is to make inference on the coefficient $\boldsymbol{\beta}^\circ$. Post-selection approach [8] is a well-known exact inference method for the above type of sparse regressions which relies on the regression model and Gaussian assumption of the noises. First, an estimator $\hat{\boldsymbol{\beta}}$ with nonzero index $\mathcal{I}$ is obtained by the formulation of the lasso problem given λ, where $\mathcal{J} \subset \mathcal{I}$ is necessary. Then, via the KKT condition of the lasso problem, we can obtain some valuable information of the data $\boldsymbol{X}, \boldsymbol{y}$, which is in the form of affine constraints of $\boldsymbol{y}$. Finally, based on the Gaussian assumption, the exact distribution of the target statistics, like the OLS estimator on $\mathcal{I}$, conditional on the above obtained affine constraints can be calculated, which is often in the truncated Gaussian form. Therefore, the statistical inference on the target statistics can be carried out. We compare our large p SEL (2.43) with the above truncated Gaussian inference for the lasso (TGL). To obtained the P-value of the optimal SEL ratio, we generate bootstrap sample $\boldsymbol{X}', \boldsymbol{y}'$. Then in order to avoid duplicate effective noise defined by the estimating equations, we still use

Table 2.2: Comparisons between SEL and truncated Gaussian inference of the lasso

	$n = 100$, $p = 200$			$n = 200$, $p = 500$			
Unit %	Case 1	Case 2	Case 3	Case 1	Case 2	Case 3	
$C = 0.90$	88.78	90.20	86.50	88.84	89.41	84.25	SEL
	88.16	84.08	70.14	90.26	87.06	63.19	TGL
$C = 0.95$	93.67	95.39	92.02	94.30	96.47	90.80	SEL
	93.88	90.00	75.46	95.49	92.94	68.71	TGL
$C = 0.99$	98.98	99.59	96.73	98.81	97.65	97.75	SEL
	98.78	94.08	80.37	99.29	98.82	73.01	TGL

original X to be the predictor, while the effective noise is generate according to $\boldsymbol{X}^T(\boldsymbol{X}'^T)^+\boldsymbol{X}'^T(\boldsymbol{y}' - \boldsymbol{X}'\boldsymbol{\beta}^\circ)$, where $(\boldsymbol{X}'^T)^+$ is the Moore-Penrose inverse of $\boldsymbol{X}'^T$. We consider the cases where $n = 100$, $p = 200$ and $n = 200$, $p = 500$, the truth $\boldsymbol{\beta}^\circ = [5, 0, ..., 0]^T$, $\lambda = 1.5 * \sqrt{\log p/n}$ and the same three types of distribution of $\boldsymbol{\epsilon}$ as above case of estimating sparse multivariate mean. The simulation results are shown in Table 2.2. From the simulation results, we can see that when the distribution assumption is correct, our method can perform almost as good as the exact inference approach. However, when the distribution is totally misspecified, the exact approach will fall completely, while our large p SEL can still some reasonable results. The simulations results support our opinion that our SEL approach can provide nonparametrization and robustification for statistical inference.

2.6.2 Real Data

Riboflavin data denotes how the log-transformed riboflavin production rate varies according to variables measuring the logarithm of the expression level of 4088 genes of Bacillus subtilis. Many research has been done to discover the association between response and predictors, like [1]. Two tasks are of particular interest: one is to select particular meaningful genes that could contribute to the response, the other is to quantify the uncertainty of the selected genes. Our proposed methodology focus on the second task. The specific data set can be founded in *hdi* R package, which contains $n = 71$ observations and $p = 4088$ predictors. To validate our methodology, we wish to test whether the gene 'YXLD at' is significant as it is shown in [1], this gene is significant at 0.05 level through multi sample-splitting inference method but not significant when tested via projected-based method [32];[26]. We apply the following procedure to construct a valid hypothesis test based on our SEL methodology:

1. Choosing proper λ to test the null hypothesis $\boldsymbol{\beta}^\circ = \mathbf{0}$ and obtaining the first p-value p_1;

2. Denoting j as the index of 'YXLD at' and Calculating the OLS estimated value at index j: $(\boldsymbol{x}_j^T\boldsymbol{x}_j)^{-1}\boldsymbol{x}_j^T\boldsymbol{y} = -0.55$;

3. Choosing another proper λ such that the SEL is feasible if $\boldsymbol{\beta}^\circ$ is changed on the component of 'YXLD at' around -0.55, then computing the 95% confidence interval of 'YXLD at'.

After performing the above steps, we can get $p_1 = 0.01$ and the confidence interval is $[-0.75, -0.33]$, which implies that at 5% significant level, we should reject the null hypothesis that the coefficients are zeros while we don't have enough evidence to reject the statement that the gene 'YXLD at' is significant. The result delivers similar information as in [1], which supports the efficacy of our proposed methodology.

2.7 Summary

In this section, we extend traditional empirical likelihood approach by incorporating a slack variable to characterize some inequality constraints of the inference target. Both asymptotic and nonasymptotic theories are provided to support the inference. In addition, some simulations show that our inference approach offers a mechanism of nonparametrization and robustification for the parametric assumptions. In the future, it is important to explore composite slack empirical likelihood where the inference target is solved through a joint optimization instead of prespecified. This composite approach provide an efficient way to learn the hypothesis, which is extremely useful in high dimensional inference problem.

2.8 Outlines of Proofs

2.8.1 Proof of Theorem 2.4.1

To obtain the Lagrange dual function, first, by introducing Lagrange multipliers $\phi \in \mathbb{R}$, $\boldsymbol{u}_0 \in \mathbb{R}^{|\mathcal{J}|}$, $\boldsymbol{h} \in \mathbb{R}^p$ and $\boldsymbol{v} \in \mathbb{R}^{|\mathcal{J}^c|}$, the optimization problem can be formulated to the following Lagrangian form:

$$\begin{aligned} T(\boldsymbol{w}, \boldsymbol{s}, \phi, \boldsymbol{h}, \boldsymbol{u}_0, \boldsymbol{v}) = -\langle \mathbf{1}, \log(n\boldsymbol{w})\rangle + \phi(\sum_{i=1}^{n} w_i - 1) + \boldsymbol{h}^T(-\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{\epsilon} + \boldsymbol{s}) \\ + \boldsymbol{u}_0^T\boldsymbol{A}[\mathcal{J},]\boldsymbol{s} + \boldsymbol{v}^T(\boldsymbol{A}[\mathcal{J}^c,]\boldsymbol{s}), \end{aligned} \tag{2.56}$$

and the Lagrange dual function is:

$$d(\phi, \boldsymbol{h}, \boldsymbol{u}_0, \boldsymbol{v}) = \min_{\boldsymbol{w},\boldsymbol{s}} T(\boldsymbol{w}, \boldsymbol{s}, \phi, \boldsymbol{h}, \boldsymbol{u}_0, \boldsymbol{v}), \tag{2.57}$$

which is expected to be maximized. To see that the Slater's condition is hold, note that the equality constraints for the primal problem are affine and we will show the conditions on inequality constraints later. Taking derivative of the Lagrangian T with respect to $\boldsymbol{w}$ and $\boldsymbol{s}$, we have

$$\begin{cases} -\frac{1}{\hat{w}_i} + \boldsymbol{h}^T \boldsymbol{z}_i + \phi = 0, \\ \boldsymbol{h}_{\mathcal{J}} = -\boldsymbol{A}[\mathcal{J},]^T \boldsymbol{u}_0, \\ \boldsymbol{h}_{\mathcal{J}^c} = -\boldsymbol{A}[\mathcal{J}^c,]^T \boldsymbol{v}. \end{cases} \tag{2.58}$$

Plugging equation (2.58) into the Lagrangian to get the Lagrange dual function, we have:

$$\begin{aligned} d(\phi, \boldsymbol{h}, \boldsymbol{u}_0, \boldsymbol{v}) &= -\sum_{i=1}^{n} \log(\frac{n}{\phi + \boldsymbol{h}^T \boldsymbol{z}_i}) + \phi(\sum_{i=1}^{n} \frac{1}{\phi + \boldsymbol{h}^T \boldsymbol{z}_i} - 1) + \sum_{i=1}^{n} \frac{\boldsymbol{h}^T \boldsymbol{z}_i}{\phi + \boldsymbol{h}^T \boldsymbol{z}_i}, \\ &= -\sum_{i=1}^{n} \log(\frac{n}{\phi + \boldsymbol{h}^T \boldsymbol{z}_i}) + n - \phi. \end{aligned} \tag{2.59}$$

Note that this dual problem implicitly require that the constraints $\phi + \boldsymbol{h}^T \boldsymbol{z}_i > 0$ for $i = 1, ..., n$, this can be guarantee by the prespecified condition and the first equation of (2.58). To maximize equation (2.59), taking derivatives with respect to $\boldsymbol{u}_0$ and ϕ, we have:

$$\begin{cases} \nabla_{\boldsymbol{u}_0} d(\phi, \boldsymbol{h}, \boldsymbol{u}_0, \boldsymbol{v}) &= \sum_{i=1}^{n} \frac{-\boldsymbol{z}_{\mathcal{J}i}}{\phi + \boldsymbol{h}^T \boldsymbol{z}_i} = \boldsymbol{0}, \\ \nabla_{\phi} d(\phi, \boldsymbol{h}, \boldsymbol{u}_0, \boldsymbol{v}) &= \sum_{i=1}^{n} \frac{1}{\phi + \boldsymbol{h}^T \boldsymbol{z}_i} - 1 = 0, \end{cases} \tag{2.60}$$

where $\boldsymbol{z}_{\mathcal{J}i}$ is the subvector of $\boldsymbol{z}_i$ indexed by $\mathcal{J}$ and $\boldsymbol{z}_{\mathcal{J}_i^c}$ can be defined similarly. In addition, taking derivatives with respect to $\boldsymbol{v}$, we have:

$$\nabla_{\boldsymbol{v}} d(\phi, \boldsymbol{h}, \boldsymbol{u}_0, \boldsymbol{v}) = \sum_{i=1}^{n} \frac{\boldsymbol{z}_{\mathcal{J}^c i}}{\phi + \boldsymbol{h}^T \boldsymbol{z}_i}. \tag{2.61}$$

Not all components of equation (2.61) can be set to zeros since the dual optimization is limited to $\boldsymbol{v} \succeq \boldsymbol{0}$. However, we always have

$$\boldsymbol{v}^T (\sum_{i=1}^{n} \frac{\boldsymbol{z}_{\mathcal{J}^c i}}{\phi + \boldsymbol{h}^T \boldsymbol{z}_i}) = 0, \tag{2.62}$$

since the components of $\boldsymbol{v}$ or the corresponding components of gradients $\nabla_{\boldsymbol{v}} d$ are zeros when maximizing the dual function. Then by combining equation (2.58), (2.60) and (2.62), we have $\hat{\phi} = n$ which gives that $\hat{w}_i = 1/(n - \boldsymbol{u}_0^T \boldsymbol{z}_{\mathcal{J}i} - \boldsymbol{v}^T \boldsymbol{z}_{\mathcal{J}^c i})$. In addition, based on the condition 2, we have $0 < \hat{w}_i < 1$, so the requirement of Slater's condition for the inequality of $\boldsymbol{w}$ is satisfied.

Denote $\boldsymbol{u} = -[\boldsymbol{u}_0^T, \boldsymbol{v}^T]^T/n$. To better characterize the dual variable $\boldsymbol{u}$, let's assume for one realization, $i \in \mathcal{I}$, $\boldsymbol{A}[\mathcal{I},]\boldsymbol{s} = 0$ or $i \in \mathcal{I}$ and for $i \in \mathcal{I}^c$, $\boldsymbol{A}[i,]s < 0$, where $\mathcal{I}$ is a subset for $\{1,..,r\}$ and $\mathcal{J}$ is a subset for $\mathcal{I}$, $\mathcal{I}^c$ is the complement of I with respect to $\{1,..,k\}$, which implies $\boldsymbol{u}_{\mathcal{I}^c} = \boldsymbol{0}$ and $\boldsymbol{u}_{\mathcal{I}-\mathcal{J}} \preceq \boldsymbol{0}$. The difference between $\mathcal{I}$ and $\mathcal{J}$ is that for $\boldsymbol{A}[\mathcal{J},]\boldsymbol{s} = 0$, the constraints are active because $\boldsymbol{\beta}^\circ$; while for $\boldsymbol{A}[\mathcal{I}-\mathcal{J},]\boldsymbol{s} = 0$ is just because the realization of the random error. Note that the rank for $\boldsymbol{A}[\mathcal{I},]$ is always less than or equal to the rank of the matrix $\boldsymbol{A}$, which is $r \leq \min(k,p)$. Thus $\mathcal{I}$ is random and we still obtain the final distribution of the likelihood ratio statistics conditioning on each component corresponding to one case for $\mathcal{A}$.

By plugging $\hat{\boldsymbol{w}}$ to the empirical likelihood, we can get: $\hat{\boldsymbol{u}} = \text{argmax}_{\boldsymbol{u}}(\boldsymbol{u}) = \sum_{i=1}^n \log(1+\boldsymbol{u}^T\boldsymbol{z}_{\mathcal{I}i})$, under the constraints $\boldsymbol{u}_{\mathcal{I}-\mathcal{J}} \preceq 0$. Therefore, we want to obtain the distribution of $c_n(\boldsymbol{\beta}^\circ) = 2\sum_{i=1}^n \log(1+\hat{\boldsymbol{u}}^T\boldsymbol{z}_{\mathcal{I}i})$ under the cases of all possible $\mathcal{I}$. If there are no constraints on $\boldsymbol{u}$, we have $\sqrt{n}\sum_{i=1}^n \frac{1}{n}\boldsymbol{z}_{\mathcal{I}i} \xrightarrow{d} \boldsymbol{A}[\mathcal{I},]\boldsymbol{\nu}$ and $\sum_{i=1}^n \frac{1}{n}\boldsymbol{z}_{\mathcal{I}i}\boldsymbol{z}_{\mathcal{I}i}^T \xrightarrow{d} \boldsymbol{A}[\mathcal{I},]\boldsymbol{V}\boldsymbol{A}[\mathcal{I},]^T$ where $\boldsymbol{\nu} \sim \mathcal{N}(0,\boldsymbol{V})$. Let $\boldsymbol{\nu}_{\boldsymbol{A}_\mathcal{I}} = \boldsymbol{A}[\mathcal{I},]\boldsymbol{\nu}$ and $\boldsymbol{V}_{\boldsymbol{A}_\mathcal{I}} = \boldsymbol{\nu}^T\boldsymbol{A}[\mathcal{I},]^T(\boldsymbol{A}[\mathcal{I},]\boldsymbol{V}\boldsymbol{A}[\mathcal{I},]^T)^{-1}\boldsymbol{A}[\mathcal{I},]\boldsymbol{\nu}$. Then based on Lemma A.0.1 we have $\sqrt{n}\boldsymbol{u} \xrightarrow{d} \boldsymbol{V}_{\boldsymbol{A}_\mathcal{I}}^{-1}\boldsymbol{\nu}_{\boldsymbol{A}_\mathcal{I}}$ and $c_n(\boldsymbol{\beta}^\circ) \xrightarrow{d} \boldsymbol{\nu}_{\boldsymbol{A}_\mathcal{I}}^T\boldsymbol{V}_{\boldsymbol{A}_\mathcal{I}}^{-1}\boldsymbol{\nu}_{\boldsymbol{A}_\mathcal{I}}$ under this realization.

For the inequality constraints, apply Taylor expansion:

$$\begin{aligned}
\boldsymbol{A}[\mathcal{I}^c,]\boldsymbol{s} = \boldsymbol{A}[\mathcal{I}^c,]\boldsymbol{X}^T\boldsymbol{W}(\boldsymbol{X}\boldsymbol{\beta}^\circ - \boldsymbol{y}) = \frac{1}{n}\sum_{i=1}^n \frac{\boldsymbol{A}[\mathcal{I}^c,]\boldsymbol{X}[i,]^T(\boldsymbol{x}_i^T\boldsymbol{\beta}^\circ - \boldsymbol{y}_i)}{1+\boldsymbol{u}^T\boldsymbol{z}_{\mathcal{I}i}} \\
= \frac{1}{n}\sum_{i=1}^n \boldsymbol{A}[\mathcal{I}^c,]\boldsymbol{X}[i,]^T(\boldsymbol{x}_i^T\boldsymbol{\beta}^\circ - \boldsymbol{y}_i) - \frac{1}{n}\sum_{i=1}^n \boldsymbol{A}[\mathcal{I}^c,]\boldsymbol{X}[i,]^T(\boldsymbol{x}_i^T\boldsymbol{\beta}^\circ - \boldsymbol{y}_i)\boldsymbol{z}_{\mathcal{I}i}^T\boldsymbol{u} \\
+o(\frac{1}{n}\sum_{i=1}^n \boldsymbol{A}[\mathcal{I}^c,]\boldsymbol{X}[i,]^T(\boldsymbol{x}_i^T\boldsymbol{\beta}^\circ - \boldsymbol{y}_i)\boldsymbol{z}_{\mathcal{I}i}^T\boldsymbol{u}),
\end{aligned} \tag{2.63}$$

we notice that

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \boldsymbol{A}[\mathcal{I}^c,]\boldsymbol{X}[i,]^T\boldsymbol{z}_{\mathcal{I}i}^T\boldsymbol{u}(\boldsymbol{x}_i^T\boldsymbol{\beta}^\circ - \boldsymbol{y}_i) \\
&= \frac{1}{n}\sum_{i=1}^n \boldsymbol{A}[\mathcal{I}^c,]\boldsymbol{X}[i,]^T(\boldsymbol{x}_i^T\boldsymbol{\beta}^\circ - \boldsymbol{y}_i)^2\boldsymbol{X}[i,]\boldsymbol{A}[\mathcal{I},]\boldsymbol{u} \\
&= \hat{\boldsymbol{V}}_{\mathcal{I}^c\mathcal{I}}(\hat{\boldsymbol{V}}_{\mathcal{I}\mathcal{I}}^{-1})\frac{1}{n}\sum_{i=1}^n \boldsymbol{z}_{\mathcal{I}i},
\end{aligned} \tag{2.64}$$

where $\hat{\boldsymbol{V}}_{\mathcal{I}^c\mathcal{I}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{A}[\mathcal{I}^c,]\boldsymbol{X}[i,]^T(\boldsymbol{x}_i^T\boldsymbol{\beta}^\circ - \boldsymbol{y}_i)^2\boldsymbol{X}[i,]\boldsymbol{A}[\mathcal{I}^c,]^T$ and $\hat{\boldsymbol{V}}_{\mathcal{I}\mathcal{I}}^{-1} = (\frac{1}{n}\sum_{i=1}^n \boldsymbol{z}_{\mathcal{I}i}\boldsymbol{z}_{\mathcal{I}i}^T)^{-1}$. Let $\boldsymbol{\nu}_{\boldsymbol{A}_{\mathcal{I}^c}} = \boldsymbol{A}[\mathcal{I}^c,]\boldsymbol{\nu}$, by law of the large numbers, the above decomposition implies

$$\sqrt{n}\boldsymbol{A}[\mathcal{I}^c,]\boldsymbol{s} \to \boldsymbol{\nu}_{\boldsymbol{A}_{\mathcal{I}^c}} \,|\, \boldsymbol{\nu}_{\boldsymbol{A}_\mathcal{I}}, \tag{2.65}$$

which is independent of $\boldsymbol{v}_{\mathcal{I}}$. For exactly one case of $\mathcal{I}$, under the following realizations, the Slater's condition is satisfied asymptotically:

$$\begin{aligned} \boldsymbol{\nu}_{\boldsymbol{A}_{\mathcal{I}^c}} \,|\, \boldsymbol{\nu}_{\boldsymbol{A}_{\mathcal{I}}} &\preceq \mathbf{0}, \\ \sqrt{n}\boldsymbol{u}_{\mathcal{I}-\mathcal{J}} &\preceq \mathbf{0}, \end{aligned} \tag{2.66}$$

since they imply $\boldsymbol{A}[\mathcal{I}^c,]\boldsymbol{s} \preceq \mathbf{0}$ for large enough n, which satisfies the Slater's condition for the inequality constraints of $\boldsymbol{s}$. Based on Corollary A.0.4, we have:

$$P(\boldsymbol{\nu}_{\boldsymbol{A}_{\mathcal{I}}}^T \boldsymbol{V}_{\boldsymbol{A}_{\mathcal{I}}}^{-1} \boldsymbol{\nu}_{\boldsymbol{A}_{\mathcal{I}}} \geq t, -\boldsymbol{\nu}_{\boldsymbol{A}_{\mathcal{I}}} \succeq \mathbf{0}) = P(\boldsymbol{\nu}_{\boldsymbol{A}_{\mathcal{I}}}^T \boldsymbol{V}_{\boldsymbol{A}_{\mathcal{I}}}^{-1} \boldsymbol{\nu}_{\boldsymbol{A}_{\mathcal{I}}} \geq t) P(-\boldsymbol{\nu}_{\boldsymbol{A}_{\mathcal{I}}} \succeq \mathbf{0}) \tag{2.67}$$

Consider all possible $\mathcal{I}$, finally we can obtain:

$$\begin{aligned} P(c_n(\boldsymbol{\beta}^\circ) \geq t) &= \sum_{A} P(c_n(\boldsymbol{\beta}^\circ) \geq t, \sqrt{n}\boldsymbol{u}_{\mathcal{I}-\mathcal{J}} \preceq \mathbf{0}, \sqrt{n}\boldsymbol{A}[\mathcal{I}^c,]\boldsymbol{s} \preceq \mathbf{0}), \\ &\xrightarrow{d} \sum_{\mathcal{I}} P(\boldsymbol{\nu}_{\boldsymbol{A}_{\mathcal{I}}}^T \boldsymbol{V}_{\boldsymbol{A}_{\mathcal{I}}}^{-1} \boldsymbol{\nu}_{\boldsymbol{A}_{\mathcal{I}}} \geq t, \boldsymbol{\nu}_{\boldsymbol{A}_{\mathcal{I}-\mathcal{J}}} \preceq \mathbf{0}, \boldsymbol{\nu}_{\boldsymbol{A}_{\mathcal{I}^c}} \,|\, \boldsymbol{\nu}_{\boldsymbol{A}_{\mathcal{I}}} \preceq \mathbf{0}) \\ &= \sum_{\mathcal{I}} P(\boldsymbol{\nu}_{\boldsymbol{A}_{\mathcal{I}}}^T \boldsymbol{V}_{\boldsymbol{A}_{\mathcal{I}}}^{-1} \boldsymbol{\nu}_{\boldsymbol{A}_{\mathcal{I}}} \geq t) P(\boldsymbol{\nu}_{\boldsymbol{A}_{\mathcal{I}-\mathcal{J}}} \preceq \mathbf{0}) P(\boldsymbol{\nu}_{\boldsymbol{A}_{\mathcal{I}^c}} \,|\, \boldsymbol{\nu}_{\boldsymbol{A}_{\mathcal{I}}} \preceq \mathbf{0}), \end{aligned} \tag{2.68}$$

the probability of $\mathcal{I} = \mathcal{J}$ is 1 minus the sum of all other probabilities. Thus $c_n(\boldsymbol{\beta}^\circ)$ follows a chi-bar-square distribution as follows:

$$\begin{aligned} &P(c_n(\boldsymbol{\beta}^\circ) \geq t) = \sum_{j=|\mathcal{J}|}^{p} P(\chi_j^2 \geq t)\delta_j, \\ &\delta_j = \sum_{|\mathcal{I}|=j} P(\boldsymbol{\nu}_{\boldsymbol{A}_{\mathcal{I}-\mathcal{J}}} \preceq \mathbf{0}) P(\boldsymbol{\nu}_{\boldsymbol{A}_{\mathcal{I}^c}} \,|\, \boldsymbol{\nu}_{\boldsymbol{A}_{\mathcal{I}}} \preceq \mathbf{0}), \text{ with } p > |\mathcal{I}|,\ \mathcal{J} \subsetneq \mathcal{I}, \\ &\delta_{|\mathcal{J}|} = 1 - \sum_{j=|\mathcal{J}|+1}^{p} \delta_j. \end{aligned} \tag{2.69}$$

2.8.2 Proof of Theorem 2.5.1

In this section, we prove the above main theorem. By introducing Lagrange multipliers $\phi \in \mathbb{R}$, $\boldsymbol{u}_0 \in \mathbb{R}^{|\mathcal{J}|}$ and $\boldsymbol{h}_0 \in \mathbb{R}^p$, the optimization problem can be formulated to the following Lagrangian form:

$$T(\boldsymbol{w}, \boldsymbol{s}, \phi, \boldsymbol{h}, \boldsymbol{u}_0) = -\langle \mathbf{1}, \log(n\boldsymbol{w}) \rangle + \phi(\sum_{i=1}^{n} w_i - 1) + \boldsymbol{h}^T(-\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{\epsilon} + \boldsymbol{s}) + \boldsymbol{u}_0^T \boldsymbol{s}_{\mathcal{J}}, \tag{2.70}$$

and the Lagrange dual function is:

$$d(\phi, \boldsymbol{h}, \boldsymbol{u}_0) = \min_{\boldsymbol{w}, \boldsymbol{s}} T(\boldsymbol{w}, \boldsymbol{s}, \phi, \boldsymbol{h}, \boldsymbol{u}_0), \tag{2.71}$$

which is expected to be maximized. Note that the equality constraints for the primal problem are affine and we will show that the Slater's conditions for the inequality constrains hold later. Taking derivative of the Lagrangian $T(\boldsymbol{w}, \phi, \boldsymbol{h}, \boldsymbol{s}, \boldsymbol{u}_0)$ with respect to $\boldsymbol{w}$ and $\boldsymbol{s}$, we have

$$\begin{cases} -\frac{1}{\hat{w}_i} + \boldsymbol{h}^T \boldsymbol{x}_i \epsilon_i + \phi = 0, \\ \boldsymbol{h}_{\mathcal{J}} = \boldsymbol{u}_0, \\ \boldsymbol{h}_{\mathcal{J}^c} = \boldsymbol{0}. \end{cases} \tag{2.72}$$

Plugging equation (2.72) into the Lagrangian to get the Lagrange dual function, we have:

$$\begin{aligned} d(\phi, \boldsymbol{u}_0) &= -\sum_{i=1}^{n} \log(\frac{n}{\phi + \boldsymbol{u}_0^T \boldsymbol{z}_i}) + \phi(\sum_{i=1}^{n} \frac{1}{\phi + \boldsymbol{z}_i} - 1) + \sum_{i=1}^{n} \frac{\boldsymbol{u}_0^T \boldsymbol{z}_i}{\phi + \boldsymbol{u}_0^T \boldsymbol{z}_i} \\ &= -\sum_{i=1}^{n} \log(\frac{n}{\phi + \boldsymbol{u}_0^T \boldsymbol{z}_i}) + n - \phi. \end{aligned} \tag{2.73}$$

Note that this dual problem implicitly require that the constraints $\phi + \boldsymbol{u}_0^T \boldsymbol{z}_i > 0$ for $i = 1, ..., n$, but we will show our prespecified conditions could guarantee them later. To maximize equation (2.73), taking derivatives with respect to $\boldsymbol{u}_0$ and ϕ, we have:

$$\begin{cases} \nabla_{\boldsymbol{u}_0} d(\phi, \boldsymbol{u}_0) & = \sum_{i=1}^{n} \frac{\boldsymbol{z}_i}{\phi + \boldsymbol{u}_0^T \boldsymbol{z}_i} = 0, \\ \nabla_{\phi} d(\phi, \boldsymbol{h}) & = \sum_{i=1}^{n} \frac{1}{\phi + \boldsymbol{u}_0^T \boldsymbol{z}_i} - 1 = 0, \end{cases} \tag{2.74}$$

which gives $\phi = n$ and implies that $\hat{w}_i = 1/(n + \boldsymbol{u}_0^T \boldsymbol{z}_i)$. Let $\boldsymbol{u} = n\boldsymbol{u}_0$, the dual optimization becomes

$$\max_{\boldsymbol{u}} \sum_{i=1}^{n} \log(1 + \boldsymbol{u}^T \boldsymbol{z}_i) \quad \text{s.t.} \quad 0 \leq \hat{\boldsymbol{w}}_i(\boldsymbol{u}) \leq 1, i = 1, ..., n; \quad \|\hat{\boldsymbol{s}}(\boldsymbol{u})\|_\infty \leq \lambda. \tag{2.75}$$

Now we denote the unconstrained maximizer of $\max_{\boldsymbol{u}} \sum_{i=1}^{n} \log(1 + \boldsymbol{u}^T \boldsymbol{z}_i)$ as $\tilde{\boldsymbol{u}}$, where $\tilde{\boldsymbol{u}}$ is a traditional M-estimator for the large n case. Then based on the condition 3, the interior of the convex hull of $\boldsymbol{z}_1, ..., \boldsymbol{z}_n$ contains zero, which implies that $0 < \hat{w}_i(\tilde{\boldsymbol{u}}) < 1, i = 1, ..., n$. In addition, denote $\boldsymbol{r}(\tilde{\boldsymbol{u}}) = \hat{\boldsymbol{s}}(\tilde{\boldsymbol{u}})_{\mathcal{J}^c} - \boldsymbol{X}_{\mathcal{J}^c}^T \boldsymbol{\epsilon}/n$, then we have

$$\begin{aligned} \hat{\boldsymbol{s}}(\tilde{\boldsymbol{u}})_{\mathcal{J}^c} = \boldsymbol{X}_{\mathcal{J}^c}^T \hat{\boldsymbol{W}}(\tilde{\boldsymbol{u}}) \boldsymbol{\epsilon} = \frac{1}{n} \sum_{i=1}^{n} \frac{\boldsymbol{X}_{i,\mathcal{J}^c}^T \epsilon_i}{1 + \tilde{\boldsymbol{u}}^T \boldsymbol{z}_i} \\ = \frac{1}{n} \boldsymbol{X}_{\mathcal{J}^c}^T (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}_{\mathcal{J}}}) \boldsymbol{\epsilon} + \frac{1}{n} \boldsymbol{X}_{\mathcal{J}^c}^T \boldsymbol{P}_{\boldsymbol{X}_{\mathcal{J}}} \boldsymbol{\epsilon} + \boldsymbol{r}(\tilde{\boldsymbol{u}}). \end{aligned} \tag{2.76}$$

Based on the asymptotic property of $\tilde{\boldsymbol{u}}$ (Lemma A.0.1) such that $\max_i(\tilde{\boldsymbol{u}}^T \boldsymbol{z}_i) = o_p(1)$ we have $\|\boldsymbol{r}(\tilde{\boldsymbol{u}})\|_\infty = o_p(\|\boldsymbol{X}_{\mathcal{J}^c}^T \boldsymbol{\epsilon}/n\|_\infty)$. Moreover, based on the selection event induced condition 2, we have

$$\|\frac{1}{n}\boldsymbol{X}_{\mathcal{J}^c}^T(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}_{\mathcal{J}}})\boldsymbol{\epsilon}\|_\infty < \lambda. \tag{2.77}$$

For the last term $\frac{1}{n}\boldsymbol{X}_{\mathcal{J}^c}^T\boldsymbol{P}_{\boldsymbol{X}_{\mathcal{J}}}\boldsymbol{\epsilon}$, since $\|\frac{1}{n}\boldsymbol{X}_{\mathcal{J}^c}^T\boldsymbol{P}_{\boldsymbol{X}_{\mathcal{J}}}\boldsymbol{\epsilon}\|_\infty \leq \|\boldsymbol{X}_{\mathcal{J}^c}^T\boldsymbol{X}_{\mathcal{J}}(\boldsymbol{X}_{\mathcal{J}}^T\boldsymbol{X}_{\mathcal{J}})^{-1}\|_\infty\|\boldsymbol{X}_{\mathcal{J}}^T\boldsymbol{\epsilon}\|_\infty$, based on Lemma A.0.2, since $\boldsymbol{X}_{\mathcal{J}}\boldsymbol{\epsilon}$ is a $|\mathcal{J}|$-dimensional vector with finite variance, with probability converging to 1, there is a constant C such that

$$\|\frac{1}{n}\boldsymbol{X}_{\mathcal{J}^c}^T\boldsymbol{P}_{\boldsymbol{X}_{\mathcal{J}}}\boldsymbol{\epsilon}\|_\infty \leq C\delta/n^{1/2} = o(\lambda). \tag{2.78}$$

Therefore with probability one, we have,

$$\|\hat{\boldsymbol{s}}(\tilde{\boldsymbol{u}})_{\mathcal{J}^c}\|_\infty < \lambda + o(\lambda) \leq \lambda. \tag{2.79}$$

Since we also have $\hat{\boldsymbol{s}}_{\mathcal{J}} = \boldsymbol{0}$, the above analysis implies that the inequality constraints $\|\hat{\boldsymbol{s}}(\tilde{\boldsymbol{u}})\|_\infty \leq \lambda$ holds automatically for the unconstrained optimizer $\tilde{\boldsymbol{u}}$ when n is sufficiently large. Note that both $0 \leq w_i \leq 1$ for $i = 1, 2, ..., n$ and $\|\boldsymbol{s}\|_\infty \leq \lambda$ are all affine inequality constraints, so the existence of $\tilde{\boldsymbol{u}}$ such that $0 \leq \hat{w}_i(\tilde{\boldsymbol{u}}) \leq 1, i = 1, ..., n$ and $\|\hat{\boldsymbol{s}}(\tilde{\boldsymbol{u}})\|_\infty \leq \lambda$ hold guarantees the requirement for inequality constraints of the Slater's condition. Hence, strict feasibility holds and the duality gap is zero, which means the transformation from prime to dual is equivalent. Moreover, when $\min_{j\in\mathcal{J}}|\boldsymbol{\beta}_j^\circ| > (1+\kappa)\lambda$, the event $\{|\boldsymbol{\beta}_{\mathcal{J}}^\circ + (\boldsymbol{X}_{\mathcal{J}}^T\boldsymbol{X}_{\mathcal{J}})^{-1}\boldsymbol{X}_{\mathcal{J}}^T\boldsymbol{\epsilon}| \succeq \lambda\boldsymbol{1}\}$ holds with probability approaching one since $\|\boldsymbol{X}_{\mathcal{J}}^T\boldsymbol{\epsilon}/n\|_\infty = O(1/n) = o(\lambda)$ and $(\boldsymbol{X}_{\mathcal{J}}^T\boldsymbol{X}_{\mathcal{J}})^{-1}/n$ is positive definite and finite. Finally, given any $t > 0$, based on asymptotic analysis of traditional EL, we have $c_n(\boldsymbol{\beta}^\circ) = \boldsymbol{\epsilon}^T\boldsymbol{X}_{\mathcal{J}}(\boldsymbol{X}_{\mathcal{J}}\boldsymbol{D}(\boldsymbol{\epsilon})^2\boldsymbol{X}_{\mathcal{J}})^{-1}\boldsymbol{X}_{\mathcal{J}}^T\boldsymbol{\epsilon} + o_p(1)$, so the obtained conditional critical probability can be obtained

$$\begin{aligned}&P(c_n(\boldsymbol{\beta}^\circ) \geq t \mid \{|\boldsymbol{\beta}_{\mathcal{J}}^\circ + (\boldsymbol{X}_{\mathcal{J}}^T\boldsymbol{X}_{\mathcal{J}})^{-1}\boldsymbol{X}_{\mathcal{J}}^T\boldsymbol{\epsilon}| \succeq \lambda\boldsymbol{1}, \text{ and } \|\boldsymbol{X}_{\mathcal{J}^c}^T(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}_{\mathcal{J}}})\boldsymbol{\epsilon}\|_\infty < n\lambda\}) \\ &= P(\boldsymbol{z}_\lambda^{\boldsymbol{X},\boldsymbol{\beta}^\circ} \geq t + o_p(1)),\end{aligned} \tag{2.80}$$

and the finally conclusion can be obtained if the distribution of $\boldsymbol{z}_\lambda^{\boldsymbol{X},\boldsymbol{\beta}^\circ}$ is continuous at t.

2.8.3 Proof of Theorem 2.5.2

$$
\begin{aligned}
-\langle \mathbf{1}, \log(n\boldsymbol{w}')\rangle &= -\langle \mathbf{1}, \log(n\hat{\boldsymbol{w}})\rangle - \sum_{i=1}^{n} \log(\frac{w_i'}{\hat{w}_i}) \\
&< -\langle \mathbf{1}, \log(n\hat{\boldsymbol{w}})\rangle - \sum_{i=1}^{n} (\frac{w_i'}{\hat{w}_i} - 1) \\
&= -\langle \mathbf{1}, \log(n\hat{\boldsymbol{w}})\rangle - \sum_{i=1}^{n} (w_i'(n - \boldsymbol{u}_0^T \boldsymbol{z}_{\mathcal{J}i} \\
&+ \boldsymbol{v}^{-T}(\boldsymbol{z}_{\mathcal{J}^c i} - \lambda) + \boldsymbol{v}^{-T}(-\boldsymbol{z}_{\mathcal{J}^c i} - \lambda)) - 1) \\
&= -\langle \mathbf{1}, \log(n\hat{\boldsymbol{w}})\rangle - \sum_{i=1}^{n} (nw_i' - 1) - \boldsymbol{u}_0^T (\sum_{i=1}^{n} w_i' \boldsymbol{z}_{\mathcal{J}i}) \\
&- \boldsymbol{v}^{-T} (\sum_{i=1}^{n} w_i' (\boldsymbol{z}_{\mathcal{J}^c i} - \lambda) - \boldsymbol{v}^{-T} (\sum_{i=1}^{n} w_i' (-\boldsymbol{z}_{\mathcal{J}^c i} - \lambda)) \\
&\leq -\langle \mathbf{1}, \log(n\hat{\boldsymbol{w}})\rangle,
\end{aligned}
$$

where the first inequality follows by the fact that $\hat{\boldsymbol{w}}$ and $\boldsymbol{w}'$ are different and the last inequality holds based on $\boldsymbol{v}^- \succeq \mathbf{0}$ and $\boldsymbol{v}^+ \succeq \mathbf{0}$.

APPENDIX A

TECHNICAL LEMMAS

Lemma A.0.1. *Suppose $\boldsymbol{z}_i \in \mathbb{R}^p$, $i = 1, ..., n$, $p < n$ are independent and mean 0 and finite and positive definite variance $\boldsymbol{V}$, denote*

$$\hat{\boldsymbol{h}} = \underset{\boldsymbol{h}}{\operatorname{argmax}} \sum_{i=1}^{n} \log(1 + \boldsymbol{h}^T \boldsymbol{z}_i) \tag{A.1}$$

$$c_n = 2 \sum_{i=1}^{n} \log(1 + \hat{\boldsymbol{h}}^T \boldsymbol{z}_i).$$

Then we have $\hat{\boldsymbol{h}} \to (\sum_{i=1}^n \boldsymbol{z}_i \boldsymbol{z}_i^T)^{-1} \sum_{i=1}^n \boldsymbol{z}_i$ and $c_n \to (\sum_{i=1}^n \boldsymbol{z}_i^T)(\sum_{i=1}^n \boldsymbol{z}_i \boldsymbol{z}_i^T)^{-1}(\sum_{i=1}^n \boldsymbol{z}_i)$.

Proof. To solve optimization (A.1), we take the derivative and find that $\hat{\boldsymbol{h}}$ should satisfy

$$\sum_{i=1}^{n} \frac{\boldsymbol{z}_i}{1 + \boldsymbol{h}^T \boldsymbol{z}_i} = \boldsymbol{0}. \tag{A.2}$$

Then we need to show $\hat{\boldsymbol{h}}$ is root-n consistency with respect to zero $\|\hat{\boldsymbol{h}}\|_2 = O_p(n^{-1/2})$. Let $\boldsymbol{g}_i(\boldsymbol{h}) = \boldsymbol{z}_i/(1 + \boldsymbol{h}^T \boldsymbol{z}_i)$, then $\hat{\boldsymbol{h}}$ satisfies $\sum_{i=1}^n \boldsymbol{g}_i(\hat{\boldsymbol{h}}) = \boldsymbol{0}$.

We consider to directly show the root-n consistency of $\hat{\boldsymbol{h}}$. Denote $d(\boldsymbol{h}) = \sum_{i=1}^n \log(1 + \boldsymbol{h}^T \boldsymbol{z}_i)$. Then, we want to show that for any given $\epsilon > 0$, there exists a large constant C such that

$$P\{\sup_{\|\boldsymbol{u}\|_2 = C} d(n^{-\frac{1}{2}}\boldsymbol{u}) - d(\boldsymbol{0}) < 0\} \geq 1 - \epsilon. \tag{A.3}$$

This implies that with probability at least $1-\epsilon$, there exists a local maximum in the ball $\{n^{-1/2}\boldsymbol{u} : \|\boldsymbol{u}\|_2 \leq C\}$. Hence, there is a local maximizer of $d(\boldsymbol{h})$ such that $\|\hat{\boldsymbol{h}}\|_2 = O_p(n^{-1/2})$. By Taylor's expansion (the first Taylor expansion), there is $0 < t < 1$ such that

$$\begin{aligned} d(n^{-\frac{1}{2}}\boldsymbol{u}) - d(\boldsymbol{0}) &= \frac{1}{n^{1/2}} \sum_{i=1}^{n} \boldsymbol{z}_i^T \boldsymbol{u} + (-\frac{1}{2n} \sum_{i=1}^{n} \boldsymbol{u}^T \boldsymbol{z}_i \boldsymbol{z}_i^T \boldsymbol{u} + \frac{1}{2} \boldsymbol{u}^T \boldsymbol{V} \boldsymbol{u}) \\ &\quad + (-\frac{1}{2} \boldsymbol{u}^T \boldsymbol{V} \boldsymbol{u}) + (-\frac{1}{2n} \sum_{i=1}^{n} \boldsymbol{u}^T \nabla \boldsymbol{g}_i(t n^{-\frac{1}{2}} \boldsymbol{u})^T \boldsymbol{u} + \frac{1}{2n} \sum_{i=1}^{n} \boldsymbol{u}^T \boldsymbol{z}_i \boldsymbol{z}_i^T \boldsymbol{u}) \\ &\equiv I_1 + I_2 + I_3 + I_4. \end{aligned} \tag{A.4}$$

By central limit theorem, $\sum_{i=1}^n \|\boldsymbol{z}_i\|_2/\sqrt{n} = O_p(1)$, so the term I_1 is upper bounded by CC', where $C' > 0$ is a constant chosen based on ϵ; while the second term I_2 can be smaller than C with probability tending to 1 by WLLN. I_3 is upper bounded by $-C^2\lambda_{\min}(\boldsymbol{V})$, which is a negative constant based on the positive definite of $\boldsymbol{V}$. It remains to show $I_4 = o_p(1)$.

$$
\begin{aligned}
|I_4| =& |\frac{1}{2n}\sum_{i=1}^n \boldsymbol{u}^T\nabla \boldsymbol{g}_i(tn^{-1/2}\boldsymbol{u})^T\boldsymbol{u} - \frac{1}{2n}\sum_{i=1}^n \boldsymbol{u}^T\boldsymbol{z}_i\boldsymbol{z}_i^T\boldsymbol{u}| \\
=& |\frac{1}{2n}\sum_{i=1}^n -\frac{\boldsymbol{u}^T\boldsymbol{z}_i\boldsymbol{z}_i^T\boldsymbol{u}}{(1+tn^{-1/2}\boldsymbol{u}^T\boldsymbol{z}_i)^2} - \frac{1}{2n}\sum_{i=1}^n \boldsymbol{u}^T\boldsymbol{z}_i\boldsymbol{z}_i^T\boldsymbol{u}| \\
=& |\frac{1}{2n}\sum_{i=1}^n \frac{tn^{-1/2}\boldsymbol{u}^T\boldsymbol{z}_i\boldsymbol{u}^T\boldsymbol{z}_i\boldsymbol{z}_i^T\boldsymbol{u}}{(1+tn^{-1/2}\boldsymbol{u}^T\boldsymbol{z}_i)^2}\| + \|\frac{1}{2n}\sum_{i=1}^n \frac{tn^{-1/2}\boldsymbol{u}^T\boldsymbol{z}_i\boldsymbol{u}^T\boldsymbol{z}_i\boldsymbol{z}_i^T\boldsymbol{u}}{1+tn^{-1/2}\boldsymbol{u}^T\boldsymbol{z}_i}| \\
\leq& C^3\frac{1}{2n}\sum_{i=1}^n \frac{n^{-1/2}\|\boldsymbol{z}_i\|_2^3}{(1+tn^{-1/2}\boldsymbol{u}^T\boldsymbol{z}_i)^2} + C^3\frac{1}{2n}\sum_{i=1}^n \frac{n^{-1/2}\|\boldsymbol{z}_i\|_2^3}{|1+tn^{-1/2}\boldsymbol{u}^T\boldsymbol{z}_i|}.
\end{aligned}
\tag{A.5}
$$

Based on the finite second moment and i.i.d. of $\boldsymbol{z}_i, i = 1, 2, ..., n$, we have $\frac{1}{n}\sum_{i=1}^n \|\boldsymbol{z}_i\|_2^3 = o_p(n^{1/2})$ and $\max_{1\leq i\leq n}\|\boldsymbol{z}_i\|_2 = o_p(n^{1/2})$ according to Lemma A.0.2. Denote $z_{\max} = \max_{1\leq i\leq n}\|\boldsymbol{z}_i\|_2$. Then for sufficient large n, the right hand side of (A.6) can be bounded by

$$
C^3\frac{1}{2n}\sum_{i=1}^n \frac{n^{-1/2}\|\boldsymbol{z}_i\|_2^3}{1-n^{-1/2}Cz_{\max}} + C^3\frac{1}{2n}\sum_{i=1}^n \frac{n^{-1/2}\|\boldsymbol{z}_i\|_2^3}{(1-n^{-1/2}Cz_{\max})^2} = o_p(1).
$$

Hence the term I_4 can be less than C with probability tending to 1. To summarize, we have $I_1 + I_3 + I_3 + I_4 \leq C(C'+2) - C^2\lambda_{\min}(\boldsymbol{V}) < 0$ by choosing a large enough constant C such that $C > (C'+2)/\lambda_{\min}(\boldsymbol{V})$. Hence, $d(n^{-\frac{1}{2}}\boldsymbol{u}) - d(\boldsymbol{0}) < 0$ with probability at least $1-\epsilon$. This implies that there is a local maximum with in the ball $\{n^{-1/2}\boldsymbol{u} : \|\boldsymbol{u}\|_2 \leq C\}$ with probability $1-\epsilon$. Therefore, by picking up the root closed to zero as $\hat{\boldsymbol{h}}$, we have $\|\hat{\boldsymbol{h}}\| = O_p(n^{-1/2})$ and $\boldsymbol{z}_i^T\hat{\boldsymbol{h}} \leq z_{\max}\|\hat{\boldsymbol{h}}\| = o_p(1)$ for $i = 1, 2, ..., n$.

After obtaining root-n consistency of $\hat{\boldsymbol{h}}$, we can expand the estimating equation (A.2) for $\boldsymbol{z}_i^T\boldsymbol{h}$ (the second Taylor expansion), which means that there exist constants $0 \leq t_i \leq 1$ for $i = 1, ..., n$, such that

$$
\sum_{i=1}^n \boldsymbol{z}_i - \sum_{i=1}^n \boldsymbol{z}_i\hat{\boldsymbol{h}}^T\boldsymbol{z}_i + \sum_{i=1}^n \boldsymbol{z}_i(t_i\hat{\boldsymbol{h}}^T\boldsymbol{z}_i)^2 = \boldsymbol{0}. \tag{A.6}
$$

Then by defining

$$
\boldsymbol{r} = (\sum_{i=1}^n \boldsymbol{z}_i\boldsymbol{z}_i^T)^{-1}\sum_{i=1}^n \boldsymbol{z}_i(t_i\hat{\boldsymbol{h}}^T\boldsymbol{z}_i)^2, \tag{A.7}
$$

we have

$$\hat{\boldsymbol{h}} = (\sum_{i=1}^{n} \boldsymbol{z}_i \boldsymbol{z}_i^T)^{-1} \sum_{i=1}^{n} \boldsymbol{z}_i + \boldsymbol{r}, \tag{A.8}$$

with $\|\boldsymbol{r}\|_2 = o(n^{-1/2})$. We define $\boldsymbol{v}$ as a normal random variable such that we have $\boldsymbol{v} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{V})$. Then by plugging the form of $\hat{\boldsymbol{h}}$ to the EL statistics, there exist $0 \leq t_i' \leq 1, i = 1, 2..., n$, such that (the third Taylor expansion):

$$\begin{aligned}
-2\langle \boldsymbol{1}, \log(n\hat{\boldsymbol{w}}) \rangle &= 2 \sum_{i=1}^{n} \log(1 + \hat{\boldsymbol{h}}^T \boldsymbol{z}_i) \\
&= 2 \sum_{i=1}^{n} \{ \hat{\boldsymbol{h}}^T \boldsymbol{z}_i - \frac{1}{2} \hat{\boldsymbol{h}}^T \boldsymbol{z}_i \boldsymbol{z}_i^T \hat{\boldsymbol{h}} + \frac{1}{3} (t_i' \hat{\boldsymbol{h}}^T \boldsymbol{z}_i)^3 \} \\
&= (\sum_{i=1}^{n} \boldsymbol{z}_i)^T (\sum_{i=1}^{n} \boldsymbol{z}_i \boldsymbol{z}_i^T)^{-1} (\sum_{i=1}^{n} \boldsymbol{z}_i) + 2\boldsymbol{r}^T \sum_{i=1}^{n} \boldsymbol{z}_i \\
&- 2\boldsymbol{r}^T \sum_{i=1}^{n} \boldsymbol{z}_i - \boldsymbol{r}^T (\sum_{i=1}^{n} \boldsymbol{z}_i \boldsymbol{z}_i^T) \boldsymbol{r} + \frac{2}{3} \sum_{i=1}^{n} (t_i' \hat{\boldsymbol{h}}^T \boldsymbol{z}_i)^3 \\
&\overset{(i)}{=} (\sqrt{n} \sum_{i=1}^{n} \frac{1}{n} \boldsymbol{z}_i)^T (\sum_{i=1}^{n} \frac{1}{n} \boldsymbol{z}_i \boldsymbol{z}_i^T)^{-1} (\sqrt{n} \sum_{i=1}^{n} \frac{1}{n} \boldsymbol{z}_i) + o_p(1)
\end{aligned} \tag{A.9}$$

where (i) is because $\boldsymbol{r}^T (\sum_{i=1}^{n} \boldsymbol{z}_i \boldsymbol{z}_i^T) \boldsymbol{r} = o_p(1)$ based on law of large numbers and the finiteness of $\boldsymbol{V}$ and $\| \sum_{i=1}^{n} (t_i \hat{\boldsymbol{h}}^T \boldsymbol{z}_i)^3 \|_2 \leq z_{\max} \|\hat{\boldsymbol{h}}\|_2 \| \sum_{i=1}^{n} \hat{\boldsymbol{h}}^T \boldsymbol{z}_i \boldsymbol{z}_i^T \hat{\boldsymbol{h}} \|_2 = o_p(1)$. □

Lemma A.0.2. *(Lemma 3, [14]) Let $\xi_i \geq 0$ be i.i.d. random variables and define $\xi_{\max} = \max_{1 \leq i \leq n} \xi_i$. If $E(\xi_1^2) < \infty$, then*

$$\xi_{\max} = o(n^{\frac{1}{2}}) \quad \textit{and} \quad \sum_{i=1}^{n} \frac{\xi_i^3}{n} = o(n^{\frac{1}{2}}). \tag{A.10}$$

Lemma A.0.3. *([18]) Suppose that $\boldsymbol{y}$ is $N(0, \boldsymbol{I})$, $\boldsymbol{Q}$ is an idempotent symmetric matrix of rank r and $\mathcal{A}$ is a cone defined by*

$$\mathcal{A} = \{ \boldsymbol{v} \in \mathbb{R}^q : \boldsymbol{d}_i^T \boldsymbol{v} \leq 0, i = 1, ..., k \}, \tag{A.11}$$

where $d_i \, (i = 1, ..., k)$ are vectors such that either $\boldsymbol{Q}\boldsymbol{d}_i = \boldsymbol{0}$ or $\boldsymbol{Q}\boldsymbol{d}_i = \boldsymbol{d}_i$. Then the conditional distribution of $\boldsymbol{y}^T \boldsymbol{Q} \boldsymbol{y}$, given $\boldsymbol{y} \in \mathcal{A}$, is that of a chi-squared random variable with r degrees of freedom. □

Based on this Lemma, we can easily get the following Corollary:

Corollary A.0.4. *Suppose* $\boldsymbol{v} \in \mathbb{R}^p$, $\boldsymbol{v} \sim N(\mathbf{0}, \boldsymbol{V})$ *and* $\boldsymbol{V}$ *is positive definite, then for any* t*, we have:*

$$P(\boldsymbol{v}^T \boldsymbol{V}^{-1} \boldsymbol{v} \geq t, \boldsymbol{V}^{-1} \boldsymbol{v} \succeq 0) = P(\boldsymbol{v}^T \boldsymbol{V}^{-1} \boldsymbol{v} \geq t) P(\boldsymbol{V}^{-1} \boldsymbol{v} \succeq 0)). \tag{A.12}$$

Proof. Let $\boldsymbol{y} = \boldsymbol{V}^{-\frac{1}{2}} \boldsymbol{v}$, and $\boldsymbol{Q} = \boldsymbol{I}_p$, then directly apply Lemma A.0.3. □

BIBLIOGRAPHY

[1] Peter Bühlmann, Markus Kalisch, and Lukas Meier. High-dimensional statistics with a view toward applications in biology. 2014.

[2] T. T. Cai, G. Xu, and J. Zhang. On recovery of sparse signals via ℓ_1 minimization. *IEEE Transactions on Information Theory*, 55(7):3388–3397, July 2009.

[3] Ruben Dezeure, Peter Bühlmann, and Cun-Hui Zhang. High-dimensional simultaneous inference with the bootstrap. *Test*, 26(4):685–719, 2017.

[4] Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

[5] Frank R. Hampel, Elvezio M. Ronchetti, Peter J. Rousseeuw, and Werner A. Stahel. *Robust statistics*. John Wiley & Sons, New York, 2005.

[6] Holger Höfling and Robert Tibshirani. Estimation of sparse binary pairwise markov networks using pseudo-likelihoods. *Journal of Machine Learning Research*, 10(Apr):883–906, 2009.

[7] Dimitris Karlis. An EM algorithm for multivariate Poisson distribution and related models. *Journal of Applied Statistics*, 30(1):63–77, 2003.

[8] Jason D Lee, Dennis L Sun, Yuekai Sun, Jonathan E Taylor, et al. Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927, 2016.

[9] Kung-Yee Liang and Scott L Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986.

[10] Po-Ling Loh and Martin J Wainwright. Structure estimation for discrete graphical models: Generalized covariance matrices and their inverses. *THE ANNALS of STATISTICS*, pages 3022–3049, 2013.

[11] Willard G Manning, Anirban Basu, and John Mullahy. Generalized modeling approaches to risk adjustment of skewed outcomes data. *Journal of health economics*, 24(3):465–488, 2005.

[12] Hamed Masnadi-shirazi and Nuno Vasconcelos. On the design of loss functions for classification: theory, robustness to outliers, and savageboost. In *Advances in Neural Information Processing Systems 21*, pages 1049–1056. 2009.

[13] Art B Owen. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75(2):237–249, 1988.

[14] Art B Owen. Empirical likelihood ratio confidence regions. *The Annals of Statistics*, 18(1):90–120, 1990.

[15] Art B Owen. Empirical likelihood for linear models. *The Annals of Statistics*, 19(4):1725–1747, 1991.

[16] Jin Qin and Jerry Lawless. Empirical likelihood and general estimating equations. *The Annals of Statistics*, 22(1):300–325, 1994.

[17] Pradeep Ravikumar, Martin J. Wainwright, and John D. Lafferty. High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *Ann. Statist.*, 38(3):1287–1319, 06 2010.

[18] Alexander Shapiro. Asymptotic distribution of test statistics in the analysis of moment structures under inequality constraints. *Biometrika*, 72(1):133–144, 1985.

[19] Y. She and K. Chen. Robust reduced-rank regression. *Biometrika*, 104(3):633–647, 2017.

[20] Yiyuan She. An iterative algorithm for fitting nonconvex penalized generalized linear models with grouped predictors. *Computational Statistics & Data Analysis*, 56(10):2976–2990, October 2012.

[21] Yiyuan She. On the finite-sample analysis of Θ-estimators. *Electronic Journal of Statistics*, 10(2):1874–1895, 2016.

[22] Yiyuan She and Kun Chen. Robust reduced rank regression. *Biometrika*, 104(3):633–647, 2017.

[23] Yiyuan She, Shao Tang, and Qiaoya Zhang. Indirect gaussian graph learning beyond gaussianity. *IEEE Transactions on Network Science and Engineering*, 2019.

[24] Auke Tellegen, David Watson, and Lee Anna Clark. On the dimensional and hierarchical structure of affect. *Psychological Science*, 10(4):297–303, 1999.

[25] Konstantinos Trohidis, Grigorios Tsoumakas, George Kalliris, and Ioannis P Vlahavas. Multi-label classification of music into emotions. In *ISMIR*, volume 8, pages 325–330, 2008.

[26] Sara Van de Geer, Peter Bühlmann, Ya'acov Ritov, Ruben Dezeure, et al. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014.

[27] Lan Wang, Jianhui Zhou, and Annie Qu. Penalized generalized estimating equations for high-dimensional longitudinal data analysis. *Biometrics*, 68(2):353–360, 2012.

[28] Lingzhou Xue, Hui Zou, and Tianxi Cai. Nonconcave penalized composite conditional likelihood estimation of sparse Ising models. *Ann. Statist.*, 40(3):1403–1429, 06 2012.

[29] Eunho Yang, Genevera Allen, Zhandong Liu, and Pradeep K Ravikumar. Graphical models via generalized linear models. In *Advances in Neural Information Processing Systems*, pages 1358–1366, 2012.

[30] Eunho Yang, Pradeep Ravikumar, Genevera I. Allen, and Zhandong Liu. Graphical models via univariate exponential family distributions. *Journal of Machine Learning Research*, 16(1):3813–3847, January 2015.

[31] Scott L Zeger, Kung-Yee Liang, and Paul S Albert. Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, pages 1049–1060, 1988.

[32] Cun-Hui Zhang and Stephanie S Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014.

BIOGRAPHICAL SKETCH

Peng Zhao was born in Henan, China. He received his B.S. in Statistics from Beijing Institute of Technology in 2015. In 2017, he attained a M.S. in Statistics from Florida State University and then pursue a doctorate in Statistics.