HIGH-DIMENSIONAL DATA INTEGRATION WITH MULTIPLE HETEROGENEOUS AND OUTLIER CONTAMINATED TASKS

YUAN ZHONG

A DISSERTATION SUBMITTED TO THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN MATHEMATICS AND STATISTICS
YORK UNIVERSITY
TORONTO, ONTARIO

FEBRUARY 2023

Abstract

Data integration is the process of extracting information from multiple sources and analyzing different related data sets simultaneously. The aggregated information can reduce the sample biases caused by low-quality data, boost the statistical power for joint inference, and enhance the model prediction. Therefore, this dissertation focuses on the development and implementation of statistical methods for data integration.

In clinical research, the study outcomes usually consist of various patients' information corresponding to the treatment. Since the joint inference across related data sets can provide more efficient estimates compared with marginal approaches, analyzing multiple clinical endpoints simultaneously can better understand treatment effects. Meanwhile, the data from different research are usually heterogeneous with continuous and discrete endpoints. To alleviate computational difficulties, we apply the pairwise composite likelihood method to analyze the data. We can show that the estimators are consistent and asymptotically normally distributed based on the Godambe information.

Under high dimensionality, the joint model needs to select the important features to analyze the intrinsic relatedness among all data sets. The multi-task feature learning is widely used to recover this union support through the penalized M-estimation framework. However, the heterogeneity among different data sets may cause difficulties in formulating the joint model. Thus, we propose the mixed $\ell_{2,1}$ regularized composite quasi-likelihood function to perform multi-task feature learning. In our framework, we relax the distributional assumption

of responses, and our result establishes the sign recovery consistency and estimation error bounds of the penalized estimates.

When data from multiple sources are contaminated by large outliers, the multi-task learning methods suffer efficiency loss. Next, we propose robust multi-task feature learning by combining the adaptive Huber regression tasks with mixed regularization. The robustification parameters can be chosen to adapt to the sample size, model dimension, and error moments while striking a balance between unbiasedness and robustness. We consider heavy-tailed distributions for multiple data sets that have bounded $(1+\omega)$th moment for any $\omega > 0$. Our method is shown to achieve estimation consistency and sign recovery consistency. In addition, the robust information criterion can conduct joint inference on related tasks for consistent model selection.

Keywords: Data Integration, Composite Likelihood, Penalized M-estimation, Robust M-estimation, Mixed $\ell_{2,1}$ Regularization, Adaptive Huber Regression, Outlier Contamination.

Acknowledgements

First, I would like to express my sincere gratitude to my supervisors, Professor Xin Gao and Professor Wei Xu. Without their persistent guidance and patient supervision, this thesis would not have been possibly completed. Professor Gao has taught me more than I could ever credit her, and I have been influenced by her excitement, commitment, and expertise in academics and research work. Professor Wei Xu is a life mentor and supportive of my academic and career goals. I have been greatly inspired by his brilliant insights and professional suggestions.

I want to extend my appreciation to Professor Cindy Fu, Professor Hyejin Ku, Professor Grace Yi, and Professor Zijiang Yang as my supervisory committee members and dissertation examiners for providing academic guidance. Professor Cindy Fu also offered me numerous help and encouragement during my graduate study life at York University. I also want to thank the faculty and staff in the Department of Mathematics and Statistics at York University.

My special thanks go to Dr. Yaguang Li and Dr. Hao Bai. They helped me solve many statistical problems and offered me technical support in programming for my research work. I also want to thank all friends I have met at York University, as it has been an unforgettable and valuable journey in my life.

Finally, I would like to express my deepest appreciation to my parents for all their never-ending support and encouragement.

All thoughts are stars.

Table of Contents

List of Tables

List of Figures

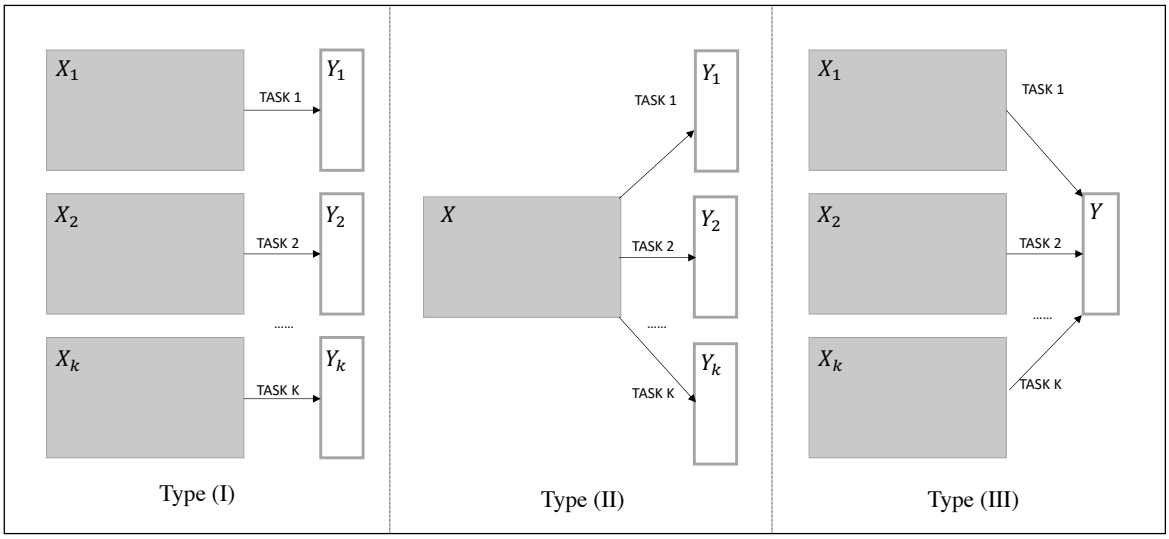
Chapter 1

Introduction

With the advancement of cyberinfrastructure technologies, increasing amounts and types of databases or data repositories are available for research in science fields. For example, the Gene Expression Omnibus (GEO) is a public archive for high-throughput microarray, and next-generation sequence functional genomics data sets [Edgar et al., 2002, Barrett et al., 2010, 2012]. The study of targeted gene expression profiles can get information based on the GEO database submitted by different institutions. Collecting related data sets and aggregating information for statistical learning is defined as a process of data integration [Council et al., 2010, Gomez-Cabrero et al., 2014]. Some examples of data integration are illustrated in Figure 1.1. In Type (I) and (III) scenarios, different learning tasks can be combined due to the similarity between data sets, which can provide the fundamental sparsity patterns of predictors. The study based on Type (II) data integration is usually used to reveal the intrinsic relationship between response variables, which can boost statistical power on the inference of parameters. By applying data integration, the synergy among different learning tasks can enhance the overall prediction accuracy compared with marginal approaches applied to the individual data.

Biomedical research usually has experimental outcomes consisting of various information

Figure 1.1: Examples of data integration. Type (I): Different learning tasks combined with different data sets; Type (II): Multiple outcomes modeled with shared predictors; Type (III): Multi-source predictors with the dependent variable (outcome of interest).



about participants corresponding to the treatment. For example, the study of omics data sets needs to model transcriptome and proteome profiles with related biological processes together in microbial biology [Zhang et al., 2010b, Meng et al., 2014, Zhang et al., 2022]. The diversity of data types indicates that the proposed models have to deal with multiple heterogeneous data sets [Gomez-Cabrero et al., 2014]. It is a theoretical challenge to formulate the joint probability density of data sets with different distributions, and the unknown relationship within and between data sets can increase the model complexity and number of parameters. Consequently, the estimating algorithm needs intensive computation in the programming. Therefore, building statistical models that can effectively alleviate the computational difficulty and provide consistent estimates is of significant interest.

In addition, the ultra-high dimensionality of multiple data sets brings tremendous information to train the learning process, and it also makes the model complicated and overly fitted for data validation. For instance, the research of gene expression profiles often contains a large number of biomarkers but has a sample scarcity problem due to the limitation of budget and participants in the experiment. In this case, data integration can aggregate information from multiple similar experiments to enhance the selection power [Gao and Carroll, 2017, Dai et al., 2020]. Different statistical methods, such as multi-task learning and fusion learning, have been developed for the joint feature selection with a wide implementation. For example, Liu et al. [2010] examined small interfering RNAs (siRNAs) efficacy across 14 different platforms for gene functional study. The molecular association between different phenotypic responses was analyzed by Zhang et al. [2010a], and their results showed that the joint feature selection identified some new biomarkers relevant to the responses. In literature, most existing procedures deal with the same type of regression or classification problems and selecting features across tasks of different natures on heterogeneous data sets has not been fully explored.

Since the data sets are collected from various sources, the data quality can be difficult to control in retrospective studies. Data integration can reduce the impact of low-quality data by introducing more reliable data. However, if the integrated data have heavy-tailed distribution or large outliers contamination, the model performance could suffer efficiency loss. For example, when modeling microarray datasets, Wang et al. [2015a] observed that the gene expression levels presented heavy tails even after normalization based on the values of the marginal kurtosises. It also happens to the data obtained from the study of functional magnetic resonance imaging (fMRI). Eklund et al. [2016] identified that the major cause of invalid fMRI inferences is that the spatial data fail to follow the Gaussian distribution. To accommodate learning tasks with outlier contamination, robust regularization methods are necessary to ensure that estimation results can be more reliable and robust for data

integration.

In this dissertation, we focus on the development and implementation of statistical methods for data integration. In Chapter 2, the joint inference across multiple related data sets is provided to show the asymptotic properties of the estimates. In addition, we propose two new methods for high-dimensional data integration. In Chapter 3, the heterogeneous multi-task feature learning is established to combine different types of learning tasks through the regularized composite quasi-likelihood. For instance, linear regression, Poisson regression, logistic regression, and multinomial regression can be jointly analyzed for the feature selection through this model. In Chapter 4, we propose the robust multi-task feature learning to model multiple related data sets that have heavy-tailed distributions or outlier contamination. In this model, the adaptive Huber regressions are combined through mixed regularization to select important features based on the robust Bayesian information criterion.

Notation

The following general notations will be used in subsequent chapters.

- For any vector $v = (v_1, v_2, \cdots, v_d)^T \in \mathbb{R}^d$, the ℓ_q norm of v is defined as $\|v\|_q = (\sum_{i=1}^d |v_i|^q)^{\frac{1}{q}}$ for some $q \geq 1$, and with $q = \infty$, $\|v\|_\infty = \sup_i\{|v_i| : i = 1, 2, \cdots, d\}$.

- The mixed $\ell_{q,r}$ norm of the vector is used for the vector that is evenly partitioned into groups. For example, for any doubly indexed vector $v = (v_{11}, v_{12}, \cdots, v_{ij}, \cdots, v_{Kd})^T$, the $\ell_{q,r}$ norm is defined as

$$\|v\|_{q,r} = \big(\sum_{j=1}^{d} \big(\sum_{i=1}^{K} v_{ij}^q\big)^{\frac{r}{q}}\big)^{\frac{1}{r}}, \tag{1.1}$$

  for $i = 1, 2, \cdots, K$ and $j = 1, 2, \cdots, d$. One special case is when $r = \infty$, $\|v\|_{q,\infty} =$

$\max_j\{(\sum_{i=1}^K v_{ij}^q)^{\frac{1}{q}}\}$. The jth block-wise subvector can be defined as follows,

$$v^{(j)} = (v_{1j}, v_{2j}, \cdots, v_{Kj})^T \tag{1.2}$$

for $j = 1, 2, \cdots, d$. In this thesis, the subset of the vector is constructed as $v_{\mathcal{E}} = \{v^{(j)} : j \in \mathcal{E}\}$ for $\mathcal{E} \subseteq \{1, 2, \cdots, d\}$, and the cardinality of the subset is denoted by $|\mathcal{E}|$. The set $\mathbb{B}_r(v^*)$ with center point v^* contains all vectors v satisfying $\|v - v^*\|_2 \leq r$.

- For any matrix $A \in \mathbb{R}^{d\times d}$, the element of the matrix is defined as $A_{[ij]}$ with $i, j = 1, 2, \cdots, d$, and the submatrix $A_{\mathcal{E}\mathcal{E}'}$ consists of all element $A_{[ij]}$ such that $i \in \mathcal{E}$ and $j \in \mathcal{E}'$. The eigenvalues of the matrix A is denoted as $\Lambda(A) = \{\lambda_1, \lambda_2, \cdots, \lambda_d\}$. The norm of matrix A is defined as follows:
  - The spectral norm: $|||A|||_2 = \Lambda_{\max}(A)$, which is the maximum eigenvalue;
  - The Frobenius norm: $|||A|||_F = (\sum_{i=1}^d \sum_{j=1}^d a_{ij}^2)^{1/2}$;
  - The ℓ_1 norm: $|||A|||_1 = \max_j \sum_{i=1}^d |A_{[ij]}|$;
  - The ℓ_∞ norm: $|||A|||_\infty = \max_i \sum_{j=1}^d |A_{[ij]}|$.
- A random variable X follows a sub-exponential distribution with parameters (ν, α), such that $E(\exp\{tX\}) \leq \exp\{t^2\nu^2/2\}$ for all $|t| < 1/\alpha$, and the ψ_1 norm

$$\|X\|_{\psi_1} = \sup_{m\geq 1} m^{-1}(E|X|^m)^{1/m} < \infty. \tag{1.3}$$

- A random variable X follows a sub-Gaussian distribution with a parameter σ, such that $E(\exp\{tX\}) \leq \exp\{t^2\sigma^2/2\}$ for all $t \in \Re$, and the ψ_2 norm is defined as

$$\|X\|_{\psi_2} = \sup_{m\geq 1} m^{-1/2}(E|X|^m)^{1/m} < \infty. \tag{1.4}$$

- Let $f(n) \gtrsim g(n)$ indicate $f(n) \geq c_1 g(n)$ for $c_1 \in (0, \infty)$; let $f(n) \lesssim g(n)$ indicate $f(n) \leq c_2 g(n)$ for $c_2 \in (0, \infty)$; and $f(n) \asymp g(n)$ if $f(n) \gtrsim g(n)$ and $f(n) \lesssim g(n)$ hold simultaneously.

Chapter 2

Joint Inference with Pairwise Composite Likelihood Method

This work has been conducted in collaboration with Dr. Bai, Dr. Gao, and Dr. Xu and published in "Multivariate Mixed Response Model with Pairwise Composite-Likelihood Method" [Bai et al., 2020]. In this chapter, we provide the implementation of the joint inference to the parameters of interest based on the pairwise composite likelihood.

2.1 Introduction

Clinical research, such as toxicity studies and laboratory examinations, can provide relevant information for measuring the effect of various treatments or experiments on patients. In practice, this type of research needs to jointly analyze multiple experimental data sets, but the research outcomes collected during the treatment can be correlated and heterogeneous with different distributions. For example, we are simultaneously studying the efficacy of treatments along with the toxicity and adverse drug reactions. The severity level could be measured as discrete or ordinal data, while the clinical examination results, such as the blood

test measures, are continuous. The experimental outcomes can be analyzed by different linear models to estimate the effect of the treatment with the relevant clinical and demographic information, but the relatedness between the data is not considered for the analysis, which may lose efficiency for the inference. Thus, it is desirable to jointly model multiple clinical data sets and analyze the treatment with other clinical information.

In the recent literature, there are various methods to model multiple data sets simultaneously. When one continuous response variable and one discrete response variable are jointly analyzed, the conditional Gaussian distribution model (CGDM) can decompose the joint distribution into a combination of the conditional distribution and the marginal distribution. In particular, Cox [1972] provided the logistic conditional distribution for binary variables, and Cox and Wermuth [1992] extended the model with a probit-type function and showed the potential connection to the latent variable model. Another conditional Gaussian distribution model, referred to as the general location model (GLOM), was proposed by Olkin and Tate [1961]. They adopted the opposite factorization, which consists of a conditional normal distribution given the categorical variables and marginal multinomial distribution. Teixeira-Pinto and Normand [2009] compared this approach with the models proposed by Sammel et al. [1997, 1999] in a comprehensive review. Yang et al. [2007] extended the model to mixed Poisson and continuous response variables through a likelihood-based approach.

In addition, the grouped continuous model (GCM) treated the categorical variables as partitioned continuous latent variables with different non-overlapping intervals, which allows the latent variables to follow a multivariate Gaussian distribution [Anderson and Pemberton, 1985, Skrondal and Rabe-Hesketh, 2007]. Poon and Lee [1987b] proposed the conditional grouped continuous model (CGCM), which can jointly model continuous response variables with categorical ones through the transformation. Catalano and Ryan [1992], Catalano [1997], and Najita et al. [2009] applied the conditional grouped continuous model to the studies of fetal toxicity for longitudinal data. Gueorguieva and Agresti [2001] proposed that the estimation

of correlated mixed response variables can be obtained by the expectation–maximization (EM) algorithm. Zhang et al. [2018] provided the parameter-expanded EM algorithm to conduct joint estimation under the full-likelihood approach.

In practice, modeling multiple heterogeneous data sets with different distributions can be computationally challenging to estimate the joint distribution. The composite likelihood method offers an alternative solution to the estimation problem based on compounded lower-dimensional distributions [Lindsay, 1988, Cox and Reid, 2004, Varin, 2008, Xu and Reid, 2011]. Faes et al. [2008] applied the composite likelihood to model multiple longitudinal data. However, their correlation structure is induced by the random effect, which does not have a closed-form expression. De Leon [2005] and De Leon and Carrièrgre [2007] developed a general mixed-data model to jointly analyze correlated nominal, ordinal, and continuous data together through the pairwise likelihood. In addition, Ekvall and Molstad [2022] used the approximate maximum likelihood estimation for the mixed-type multivariate response regression. The multivariate mixed response model proposed by Bai et al. [2020] can alleviate the computational difficulties by using three types of bivariate models, which conduct joint inference based on the pairwise composite likelihood.

Therefore, we can analyze multiple experimental data sets together and jointly estimate the effect of treatment on each outcome of interest by applying the multivariate mixed response model [Bai et al., 2020]. The model can estimate the parameters of the mean structure and the correlation among different outcomes simultaneously. We derive the asymptotic properties of the composite-likelihood estimates and derive three composite-likelihood test statistics for joint inference. The hypothesis tests can be applied to a group of parameters related to all data sets, while the model also includes some nuisance parameters. Simulation studies were conducted to examine the empirical performance of the method in comparison with the conventional approaches. In addition, we apply this method to the clinical data from a colorectal cancer study. We analyze the effect of the treatment and other clinical factors' on

multiple correlated responses of the patients.

2.2 Methodology

2.2.1 Model Setup

Suppose there are n observations $y_1, y_2, \ldots, y_i, \ldots, y_n$ in a clinical dataset, and each observation contains K multiple outcomes $y_i = (y_{1i}, y_{2i}, \ldots, y_{ki}, \ldots, y_{Ki})^T$, which are correlated and heterogeneous with continuous and binary variables. Suppose we wish to model the effects of a collection of covariates, and the generalized linear model can be constructed for each outcome as

$$g_k(E(y_{ki}|x_{ki})) = x_{ki}^T \beta_k,$$

in which the covariate x_{ki} with $k = 1, 2, \ldots, K$, and $i = 1, 2, \ldots, n$ for different responses can be the same or different, and g_k denotes as the link function used for the kth response. In particular, if the response variable is continuous, we can use the regression model to fit the data such that

$$y_{ki} = x_{ki}^T \beta_k + \varepsilon_k.$$

When modeling binary outcomes, we can use the latent variable transformation based on Dunson [2000], such that with Normal CDF $\Phi(\cdot)$,

$$\Phi^{-1}(P(y_{ki} = 1|x_{ki})) = x_{ki}^T \beta_k.$$

To analyze all K outcomes simultaneously, the joint likelihood function can be given by

$$L(\theta) = \prod_{i=1}^{n} f(y_{1i}, y_{2i}, \ldots, y_{ki}, \ldots, y_{Ki}; \theta),$$

where θ includes all parameters associated with the joint density function $f(\cdot)$. However, this joint density function $f(\cdot)$ is difficult to formulate for the mixture of continuous and discrete variables, and it is computationally intensive to estimate the parameters of interest through this multivariate model. Alternatively, we can set up the pairwise likelihood function for responses y_{ki} and y_{li} as

$$L_{kl}(\theta_{kl}) = \prod_{i=1}^{n} f_{kl}(y_{ki}, y_{li}; \theta_{kl}). \tag{2.1}$$

The set pf parameters θ_{kl} contains the coefficients of each linear model (β_k, β_l), the standard deviation of the errors (σ_k, σ_l), and pairwise correlation ρ_{kl} associated with y_{ki} and y_{li}. The joint density function $f_{kl}(\cdot)$ only needs to model three different bivariate structures, such that outcomes y_{ki}, and y_{li} are both continuous variables, outcomes y_{ki}, and y_{li} are both binary variables, and outcomes y_{ki}, and y_{li} contain one continuous and one binary variable.

The log-likelihood function is given by $\ell_{kl}(\theta_{kl}) = \log L_{kl}(\theta_{kl})$, and the score function is given by

$$U_{kl}(\theta_{kl}) = \sum_{i=1}^{n} f_{kl}(y_{ki}, y_{li}; \theta_{kl})^{-1} \frac{\partial}{\partial \theta_{kl}} f_{kl}(y_{ki}, y_{li}; \theta_{kl}). \tag{2.2}$$

The pairwise composite likelihood function of these K response variables is the product of $\binom{K}{2}$ paired likelihood functions

$$\mathrm{CL}(\theta) = \prod_{k=1}^{K-1} \prod_{l=k+1}^{K} L_{kl}(\theta_{kl}),$$

and the composite score function is constructed by differentiating the composite log-likelihood function

$$U(\theta) = \frac{\partial}{\partial \theta} \log \mathrm{CL}(\theta) = \sum_{k=1}^{K-1} \sum_{l=k+1}^{K} U_{kl}(\theta_{kl}). \tag{2.3}$$

By solve $U(\hat{\theta}) = 0$ based on (2.3), we can obtained the maximum composite likelihood estimator $\hat{\theta}$.

2.2.2 Theoretical Results

As Cox and Hinkley [1974] and Kent [1982] presented the hypothesis testing of the full likelihood function and its extensions, the composite likelihood function can be treated as the misspecified likelihood function. Its asymptotic properties were reviewed and discussed by Varin [2008], Xu and Reid [2011], and Gao and Song [2010]. Following this framework, the pairwise composite likelihood function implemented in our analysis produces the estimators, which are consistent and asymptotically normally distributed.

The Godambe information [Godambe, 1960] G of the parameters θ for the log composite likelihood function involves the sensitivity matrix H and the variability matrix J,

$$G(\theta) = H(\theta) J^{-1}(\theta) H(\theta),$$

where the sensitivity matrix and variability matrix are defined as

$$H(\theta) = E_\theta\{-\nabla U(\theta; y_i)\} \text{ and } J(\theta) = Var_\theta\{U(\theta; y_i)\}.$$

where $U(\theta; y_i)$ denotes the score function of the ith observation and the total score function $U(\theta) = \sum_{i=1}^{n} U(\theta; y_i)$.

Assumption 2.1. *The parameter space $\theta \in \Theta \subset \mathbb{R}^d$ is a closed set with fixed d. For any $k, l = 1, 2, \cdots, K$, each log-likelihood function $\ell_{kl}(\theta_{kl})$ is measurable function of (y_{ki}, y_{li}) for any θ_{kl}, and is distinct for different values of θ_{kl}. Let the true parameter be denoted by $\theta^* \in \Theta$. It is assumed that $E_{\theta^*}\{U_{kl}(\theta_{kl})\} = 0$.*

Assumption 2.2. *The sensitivity matrix $H(\theta)$ and variability matrix $H(\theta)$ are assumed with restricted eigenvalues, such that $0 < \min\{\Lambda(H(\theta^*))\} < \max\{\Lambda(H(\theta^*))\} < \infty$ and $0 < \min\{\Lambda(J(\theta^*))\} < \max\{\Lambda(J(\theta^*))\} < \infty$.*

Assumption 2.3. *The pairwise composite likelihood admits third derivatives for almost all y_i and for all $\theta \in \mathbb{B}_r(\theta^*)$. The third derivatives denoted as $\nabla^2 U(\theta)$ is assumed to satisfy*

$$\left\| \left\| \nabla^2 U(\theta) u \right\| \right\|_2 \leq \mathcal{W}$$

for some constant $\mathcal{W} > 0$ with unit vector u.

Theorem 2.1. *Under Assumptions (2.1) - (2.3), as $n \to \infty$, the maximum composite likelihood estimator $\hat{\theta} \in \mathbb{B}_r(\theta^*)$ is asymptotically normally distributed as*

$$\sqrt{n}(\hat{\theta} - \theta^*) \xrightarrow{d} N_d(0, G^{-1}).$$

Proof. The asymptotic normality of the maximum composite likelihood estimator has been established in previous works, and we just provide some important steps of the proof.

Since we have $U(\hat{\theta}) = 0$, we can show the second-order Taylor expansion,

$$\begin{aligned}
0 = U(\hat{\theta}) = U(\theta^*) + \nabla U(\theta^*)(\hat{\theta} - \theta^*) + \frac{1}{2}(\hat{\theta} - \theta^*)\nabla^2 U(\tilde{\theta})(\hat{\theta} - \theta^*) \\
(\hat{\theta} - \theta^*) = \big(- \nabla U(\theta^*) - \frac{1}{2}\nabla^2 U(\tilde{\theta})(\hat{\theta} - \theta^*)\big)^{-1} U(\theta^*) \\
\sqrt{n}(\hat{\theta} - \theta^*) = \big(- \frac{1}{n}\nabla U(\theta^*) - \frac{1}{2n}\nabla^2 U(\tilde{\theta})(\hat{\theta} - \theta^*)\big)^{-1} \frac{1}{\sqrt{n}} U(\theta^*),
\end{aligned}$$

where the point $\tilde{\theta} = \alpha\theta^* + (1-\alpha)\hat{\theta}$ for some $\alpha \in (0,1)$. We can show that $\tilde{\theta} \in \mathbb{B}_r(\theta^*)$ for some r.

Based on the central limit theory, the score function $\frac{1}{\sqrt{n}}U_{kl}(\theta^*_{kl})$ is asymptotically normally distributed. From (2.3), the composite score function can satisfy

$$\frac{1}{\sqrt{n}}U(\theta^*) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} U(\theta^*; y_i) \xrightarrow{d} N_d(0, J(\theta^*)).$$

In addition, by applying the law of large numbers,

$$-\frac{1}{n}\nabla U(\theta^*) = -\frac{1}{n}\sum_{i=1}^{n} \nabla U(\theta^*; y_i) \xrightarrow{P} H(\theta^*)$$

Based on Assumption (2.3), we can show that

$$\left|\left|\left|-\frac{1}{n}\nabla^2 U(\tilde{\theta})(\hat{\theta}-\theta^*)\right|\right|\right|_2 \leq \frac{1}{n}\left|\left|\left|\nabla^2 U(\tilde{\theta})u\right|\right|\right|_2 \|\hat{\theta}-\theta^*\|_2 \leq \frac{1}{n}\mathcal{W}^* r \lesssim \frac{1}{n}.$$

This implies that any element in the matrix $-\frac{1}{n}\nabla^2 U(\tilde{\theta})(\hat{\theta}-\theta^*)$ has order less than $1/n$.

Therefore, we can show that as $n \to \infty$,

$$\begin{aligned}\sqrt{n}(\hat{\theta}-\theta^*) &= \big(-\frac{1}{n}\nabla U(\theta^*) - \frac{1}{2n}\nabla^2 U(\tilde{\theta})(\hat{\theta}-\theta^*)\big)^{-1}\frac{1}{\sqrt{n}}U(\theta^*) \\ &\approx H^{-1}(\theta^*)\frac{1}{\sqrt{n}}\sum_{i=1}^{n} U(\theta^*; y_i) \xrightarrow{d} N_d(0, G^{-1}).\end{aligned}$$

□

The sensitivity matrix $H(\theta)$ and the variability matrix $J(\theta)$ can be evaluated by the empirical estimates under the maximum composite likelihood estimators,

$$H(\hat{\theta}) = -\frac{1}{n}\sum_{i=1}^{n}\nabla U(\theta; y_i)\big|_{\hat{\theta}} \text{ and } J(\hat{\theta}) = \frac{1}{n}\sum_{i=1}^{n} U(\theta; y_i)U(\theta; y_i)^T.$$

Furthermore, according to Theorem 2.1, the composite Wald statistic, the composite score statistic, and the composite likelihood ratio statistic for testing null-hypothesis H_0: $\hat{\theta} = \theta^*$ are given respectively by

$$W_e = n(\hat{\theta} - \theta^*)G(\hat{\theta} - \theta^*)$$
$$W_u = n^{-1}U(\theta^*)J^{-1}U(\theta^*),$$
$$W = 2\{\log \mathrm{CL}(\hat{\theta}) - \log \mathrm{CL}(\theta^*)\}.$$

Testing with Nuisance Parameters

Suppose the parameters are partitioned as $\theta = \{\psi, \lambda\}$ with $\psi \in \mathbb{R}^q$, $\lambda \in \mathbb{R}^r$, and $d = q + r$. The parameter of interest is ψ, and λ are treated as the nuisance parameters for the hypothesis testing. In this setting, the Godambe information matrix and its inverse can be partitioned as

$$G = \begin{bmatrix} G_{\psi\psi} & G_{\psi\lambda} \\ G_{\lambda\psi} & G_{\lambda\lambda} \end{bmatrix} \text{ and } G^{-1} = \begin{bmatrix} G^{\psi\psi} & G^{\psi\lambda} \\ G^{\lambda\psi} & G^{\lambda\lambda} \end{bmatrix},$$

and the inverse of the submatrix pertaining to ψ is given by $G_{\psi\psi,\lambda} = (G^{\psi\psi})^{-1} = G_{\psi\psi} - G_{\psi\lambda}G_{\lambda\lambda}^{-1}G_{\lambda\psi}$. According to the asymptotic theorem, the composite Wald statistics under the null hypothesis H_0: $\psi = \psi^*$ using $\hat{\lambda}(\psi^*)$ is given by

$$W_e(\psi^*) = n(\hat{\psi} - \psi^*)G_{\psi\psi,\lambda}(\hat{\psi} - \psi^*),$$

which has an asymptotic χ^2_q distribution. Similarly, we define the composite score statistics

$$W_u(\psi^*) = n^{-1}U(\psi^*, \hat{\lambda}(\psi^*))H^{\psi\psi}G_{\psi\psi,\lambda}H^{\psi\psi}U(\psi^*, \hat{\lambda}(\psi^*)),$$

where the matrix $H^{\psi\psi}$ can be obtained by partitioning the inverse sensitivity matrix H. Also, the composite likelihood ratio statistic can be obtained by

$$W(\psi^*) = 2\{\log \mathrm{CL}(\hat{\psi}, \hat{\lambda}) - \log \mathrm{CL}(\psi^*, \hat{\lambda}(\psi^*))\},$$

with the unrestricted maximum composite likelihood estimate $\hat{\theta} = \{\hat{\psi}, \hat{\lambda}\}$. However, the asymptotic distribution of the composite likelihood ratio under H_0 is given by $\sum_{k=1}^{K} \lambda_k \chi^2_{1(k)}$, where $\chi^2_{1(k)}$ are independent χ^2_1 variates and $\lambda_1, \lambda_2 \cdots, \lambda_K$ are the eigenvalues of the matrix $H_{\psi\psi,\lambda} G^{\psi\psi}$ with $H_{\psi\psi,\lambda} = H_{\psi\psi} - H_{\psi\lambda} H^{-1}_{\lambda\lambda} H_{\lambda\psi}$. There are different adjustments to this non-standard weighted chi-square distribution [Rotnitzky and Jewell, 1990, Geys et al., 1999, Pace et al., 2011]. For example, we can apply the adjustment by introducing the scaling factor $\bar{\lambda} = \sum_{k=1}^{K} \lambda_k / K$, then the adjusted composite likelihood ratio has the same asymptotic distribution as $W_e(\psi^*)$ and $W_u(\psi^*)$,

$$\frac{W}{\bar{\lambda}} \xrightarrow{d} \chi^2_q. \tag{2.4}$$

Therefore, the composite likelihood method can simplify the modeling of correlated responses with multiple generalized linear models and allow users to conduct statistic inferences on parameters of interest from different generalized linear models. Moreover, we can select a subset of the parameters and conduct a further inferential assessment in the presence of nuisance parameters.

2.3 Simulation

Different simulation studies were implemented to show the validity of the pairwise composite likelihood method. The estimation results from the proposed model are compared with the full-likelihood and marginal approaches, respectively.

2.3.1 Comparison with Maximum Full-Likelihood Estimation

In the multivariate regression with correlated continuous outcomes, the full-likelihood estimation can be conducted without numerical integration. Thus, we can compare the maximum composite-likelihood estimates with the full-likelihood approach through the simulation study. The simulated samples contain four continuous response variables y_{1i}, y_{2i}, y_{3i}, and y_{4i}, which are generated from Equation (2.5):

$$\begin{aligned} y_{1i} &= \alpha_{c_1} + \beta_{c_1} x_{11i} + \gamma_{c_1} x_{12i} + \varepsilon_{1i}, \\ y_{2i} &= \alpha_{c_2} + \beta_{c_2} x_{21i} + \gamma_{c_2} x_{22i} + \varepsilon_{2i}, \\ y_{3i} &= \alpha_{c_3} + \beta_{c_3} x_{31i} + \gamma_{c_3} x_{32i} + \varepsilon_{3i}, \\ y_{4i} &= \alpha_{c_4} + \beta_{c_4} x_{41i} + \gamma_{c_4} x_{42i} + \varepsilon_{4i}. \end{aligned} \tag{2.5}$$

The sets of covariates are independently simulated, such that $\{x_{11i}, x_{21i}, x_{31i}, x_{41i}\} \sim N(0,1)$ and $\{x_{12i}, x_{22i}, x_{32i}, x_{42i}\} \sim N(0, 0.5)$. The errors are correlated and generated from a multivariate normal distribution $N_4(0, \Sigma)$, and the variance-covariance matrix Σ is given by

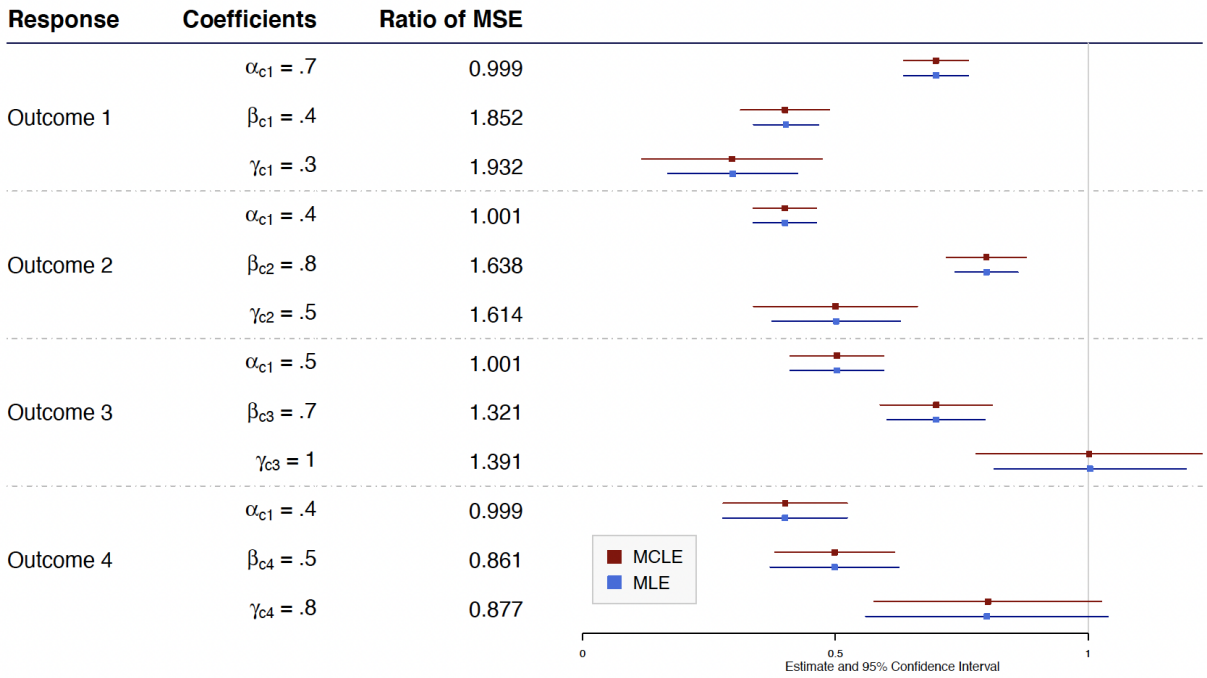
$$\begin{bmatrix} \sigma^2_{c_1} & \sigma_{c_1}\sigma_{c_2}\rho_{c_1c_2} & \sigma_{c_1}\sigma_{c_3}\rho_{c_1c_3} & \sigma_{c_1}\sigma_{c_4}\rho_{b_1c_4} \\ \sigma_{c_1}\sigma_{c_2}\rho_{c_1c_2} & \sigma^2_{c_2} & \sigma_{c_2}\sigma_{c_3}\rho_{c_2c_3} & \sigma_{c_2}\sigma_{c_4}\rho_{c_2c_4} \\ \sigma_{c_1}\sigma_{c_3}\rho_{c_1c_3} & \sigma_{c_2}\sigma_{c_3}\rho_{c_2c_3} & \sigma^2_{c_3} & \sigma_{c_3}\sigma_{c_4}\rho_{c_3c_4} \\ \sigma_{c_1}\sigma_{c_4}\rho_{c_1c_4} & \sigma_{c_2}\sigma_{c_4}\rho_{c_2c_4} & \sigma_{c_3}\sigma_{c_4}\rho_{c_3c_4} & \sigma^2_{c_4} \end{bmatrix}.$$

In the simulation, the variances are designed as $\sigma^2_{c_1} = 1$, $\sigma^2_{c_2} = 1$, $\sigma^2_{c_3} = 2.25$, and $\sigma^2_{c_4} = 4$, and an identical correlation $\rho = 0.3$ is applied between the errors in the data generating process.

The simulation results (Figure 2.1) were obtained through 1000 independent replications. The maximum composite-likelihood estimators demonstrate similar performance when com-

pared with the full-likelihood approach. The simulated results also show that the estimates are close to each other, and the maximum likelihood estimators have slightly higher relative efficiency.

Figure 2.1: The comparison between the maximum full-likelihood estimation and maximum composite likelihood estimation for the regression coefficients on the multivariate continuous outcomes. The ratio of the mean squared error (MSE) was computed using the MSE of the maximum composite-likelihood estimate (MCLE) over the MSE of the maximum likelihood estimate (MLE).



2.3.2 Comparison with Marginal Approach

The full-likelihood approach is computationally challenging for the mixed outcome regression, and marginal regression is often resorted to in order to conduct the analysis. We implemented simulation studies to evaluate the performance of our proposed method in comparison with

marginal regression. We first tested the overall performance of the point estimates when the outcomes had different levels of dependency and covariates. Next, we focused on the test of composite statistics. The joint inference across related heterogeneous responses can provide statistical inference with nuisance parameters and attains a higher statistical power in terms of dealing with joint inference.

Simulation Settings

We generate the sample data consisting of two binary responses y_{1i} and y_{2i} and two continuous responses y_{3i} and y_{4i}. The binary variables are obtained based on the corresponding latent normal variables y_{1i}^* and y_{2i}^* through the probit link function,

$$\begin{aligned} \text{probit}(\mu_{1i}) &= \mu_{1i}^*, \\ \text{probit}(\mu_{2i}) &= \mu_{2i}^*. \end{aligned}$$

The simulation studies of the responses are based on equation 2.6 associated with covariates respectively,

$$\begin{aligned} y_{1i}^* &= \alpha_{b_1} + \beta_{b_1} x_{11i} + \gamma_{b_1} x_{12i} + \varepsilon_{1i}, \\ y_{2i}^* &= \alpha_{b_2} + \beta_{b_2} x_{21i} + \gamma_{b_2} x_{22i} + \varepsilon_{2i}, \\ y_{3i} &= \alpha_{c_3} + \beta_{c_3} x_{31i} + \gamma_{c_3} x_{32i} + \varepsilon_{3i}, \\ y_{4i} &= \alpha_{c_4} + \beta_{c_4} x_{41i} + \gamma_{c_4} x_{42i} + \varepsilon_{4i}. \end{aligned} \tag{2.6}$$

We provide different simulation scenarios of the values of the covariates and three levels of correlation to analyze the response variables with the proposed model. The regression parameters are arbitrarily chosen and set to be fixed values in each simulation study. The errors in equation 2.6 follow a multivariate normal $N_4(0, \Sigma)$, and the variance-covariance

matrix Σ is given by

$$\begin{bmatrix} \sigma^2_{b_1} & \sigma_{b_1}\sigma_{b_2}\rho_{b_1b_2} & \sigma_{b_1}\sigma_{c_1}\rho_{b_1c_1} & \sigma_{b_1}\sigma_{c_2}\rho_{b_1c_2} \\ \sigma_{b_1}\sigma_{b_2}\rho_{b_1b_2} & \sigma^2_{b_2} & \sigma_{b_2}\sigma_{c_1}\rho_{b_2c_1} & \sigma_{b_2}\sigma_{c_2}\rho_{b_2c_2} \\ \sigma_{b_1}\sigma_{c_1}\rho_{b_1c_1} & \sigma_{b_2}\sigma_{c_1}\rho_{b_2c_1} & \sigma^2_{c_1} & \sigma_{c_1}\sigma_{c_2}\rho_{c_1c_2} \\ \sigma_{b_1}\sigma_{c_2}\rho_{b_1c_2} & \sigma_{b_2}\sigma_{c_2}\rho_{b_2c_2} & \sigma_{c_1}\sigma_{c_2}\rho_{c_1c_2} & \sigma^2_{c_2} \end{bmatrix}.$$

In the following data-generating processes, the values of the variance-covariance parameters are set as $\sigma^2_{b_1} = 1$, $\sigma^2_{b_2} = 1$, $\sigma^2_{c_1} = 16$, and $\sigma^2_{c_2} = 25$, and the correlation is designed at the levels of low (all $\rho = 0.3$), medium (all $\rho = 0.5$), and high (all $\rho = 0.7$) respectively to assess the underlying model. Since there is no constraint on the sign of the correlation, the negative correlation can be estimated through our algorithm without further assumptions.

Point estimates

Different simulation scenarios are designed to assess the performance of the underlying model on the point estimates by 1000 independent replications. There are two different sets of simulations for the data-generating process, and within each setting, we analyze three levels of correlation respectively. As shown in table 2.1, the values of the regression parameters and the standard deviation of the continuous response variables are given across all simulation studies. In the first simulation setting, we provide 300 samples, and the response variables are associated with covariates of distinct values. The covariates are identically and independently simulated in each linear model from a normal distribution $N(0, 1)$, respectively.

In the second simulation, 1000 independent samples are generated. There is one common covariate shared across four response variables in equation 2.6, such that $x_{11i} = x_{21i} = x_{31i} = x_{41i}$ simulated from $N(0, 1)$. In addition, we generate $\{x_{12i}, x_{22i}, x_{32i}, x_{42i}\}$ from a Bernoulli (0.5), which are different for each response. This setting represents the scenario in practice

Table 2.1: The ratio of the mean squared error (MSE) of the composite likelihood method (CLM) to the marginal model (GLM). Results based on 1000 independent simulations under two different scenarios and three different levels of correlation.

	Simulation I *			Simulation II †		
	Low	Med	High	Low	Med	High
$\alpha_{b_1} = 0.2$	1.00	0.99	0.98	0.97	0.92	0.85
$\beta_{b_1} = 0.3$	0.93	0.81	0.66	1.00	0.99	0.97
$\gamma_{b_1} = 0.3$	0.93	0.81	0.66	0.95	0.85	0.71
$\alpha_{b_2} = 0.2$	1.00	0.98	0.97	0.97	0.92	0.86
$\beta_{b_2} = 0.3$	0.94	0.84	0.69	1.00	0.98	0.95
$\gamma_{b_2} = 0.5$	0.94	0.83	0.70	0.95	0.86	0.71
$\alpha_{c_1} = 0.5$	1.00	1.00	1.00	0.96	0.89	0.79
$\beta_{c_1} = 8$	0.89	0.73	0.50	1.00	1.00	1.00
$\gamma_{c_1} = 10$	0.90	0.74	0.51	0.93	0.80	0.59
$\sigma_{c_1} = 4$	1.01	1.01	1.01	1.01	1.01	1.01
$\alpha_{c_2} = 0.4$	1.00	1.00	1.00	0.97	0.90	0.79
$\beta_{c_2} = 5$	0.92	0.77	0.53	1.00	1.00	1.00
$\gamma_{c_2} = 8$	0.92	0.75	0.50	0.94	0.80	0.57
$\sigma_{c_2} = 5$	1.01	1.01	1.01	1.01	1.01	1.01

* Simulation I: N = 300, and the four responses have different covariates;
† Simulation II: N= 1000, the responses shared one common set of covariates.

when a common factor is included in all of the response models.

In table 2.1, we provide the ratio of the mean squared error (MSE) of the proposed method to the marginal approaches. This ratio represents the relative efficiency of the proposed method in comparison with the marginal method under different settings. In most of the simulation settings, the ratio rates of the MSE are well below 1. When the responses are highly correlated and have different covariate sets, our method can reduce MSE by 50%, which indicates a large efficiency gain.

Table 2.2: Type 1 error rate and statistical power under different sample sizes ($N = 500$ and $N = 1000$).

	Type I Error		**Statistic Power**	
	$N = 500$	$N = 1000$	$N = 500$	$N = 1000$
	Composite likelihood method $H_0 : \beta_{b_1} = \beta_{b_2} = \beta_{c_1} = \beta_{c_2} = 0$			
Likelihood ratio	0.054	0.043	0.804	0.988
Wald statistics	0.058	0.043	0.800	0.989
Scoring statistics	0.058	0.042	0.798	0.989
	Multiple Test			
Bonferroni test	0.051	0.040	0.569	0.902

* The likelihood ratio statistic is adjusted as equation 2.4, which approximates to a χ^2_4.

Statistical test

The test of composite likelihood statistics can jointly assess parameters of interest across different generalized linear models, while the conventional methods cannot achieve this. The simulation studies are conducted to measure the type I error rate and the statistical power in comparison with the marginal approaches.

This simulation study is conducted to perform the hypothesis test. The correlated responses are generated based on equation 2.6 with all correlation $\rho = 0.3$. The parameters of interest are the regression coefficients $\{\beta_{b_1}, \beta_{b_2}, \beta_{c_1}, \beta_{c_2}\}$ of the first covariates across four generalized linear models, and the first covariates x_i are independently simulated from $N(0, 1)$. The regression coefficients of the second covariates $\{\gamma_{b_1}, \gamma_{b_2}, \gamma_{c_1}, \gamma_{c_2}\}$ and other parameters are nuisance parameters with $\{x_{12i}, x_{22i}, x_{32i}, x_{42i}\} \sim N(0, 0.5)$. In the simulation study to assess the type 1 error rate, the regression parameters $\{\beta_{b_1}, \beta_{b_2}, \beta_{c_1}, \beta_{c_2}\}$ are equal to zero in all generalized linear models, while other parameters have the same values as the previous

simulation in table 2.1. To assess the statistical power, we fix the values of the regression parameters as $\beta_{b_1} = \beta_{b_2} = 0.1$ for the binary responses and $\beta_{c_1} = \beta_{c_2} = 0.3$ for the continuous responses.

Table 2.2 illustrated the results of over 2000 independent replications. Since the proposed model analyzes all responses simultaneously, the simulated type I error rates are valid and close to 0.05. Through the test of the joint effect of the covariate of interest on all responses, the simulated statistical power is enhanced by our proposed model in comparison with the results from the marginal approaches. As the sample size increases from 500 to 1000, the composite likelihood statistics produce increased statistical power from 0.800 to 0.989. The overall performance demonstrates that the composite statistics are more powerful than the marginal models.

2.4 Data Analysis

In this section, the composite likelihood method is applied to the clinical data from a colorectal cancer study. The data consist of clinical observations and demographic information on 743 patients, which are mixed with both categorical and continuous data. Our research interest is to evaluate the effect of treatment and other clinical factors on the toxicity outcomes. We focus on four common toxicity events that are related to colorectal cancer treatment. First, we choose nausea and diarrhea as two categoric responses. They are ordinal data measuring the severity of the toxicity from grades 1 to 4. In our model setting, we only concern with the occurrence of nausea and diarrhea for each patient. Therefore, these two responses are designed as binary variables, which are coded as 1 if they occurred and 0 if there is no record during the treatment. The continuous responses include two blood test measures, namely the counts of the hemoglobin (HGB) and white blood cell (WBC). Each patient had several times of blood examinations during the treatment, and we took the highest value for analysis.

The explanatory variables contain the treatment effect (two different treatment therapies), demographic information, tumor status, and genetic test results for each patient. In total, we need to jointly estimate 68 parameters for the coefficients of four linear models and the correlation between each outcome.

Table 2.3: The difference in treatment effect between two treatment therapies. GLM: the generalized linear model; CLM: the composite likelihood method.

Regression parameter	**Models**	
	GLM	**CLM**
y_{b_1}: occurrence of nausea		
Interceptα_{b_1}	-0.2685 ± 2.502	-0.2793 ± 2.582
(p-value)	(0.833)	(0.832)
Treatment effectβ_{b_1}	-0.2644 ± 0.190	-0.2724 ± 0.193
(p-value)	(0.006)	(0.006)
y_{b_2}: occurrence of diarrhea		
Intercept α_{b_2}	0.6631 ± 2.557	0.6741 ± 2.605
(p-value)	(0.611)	(0.612)
Treatment effect β_{b_2}	-0.6231 ± 0.192	-0.6422 ± 0.198
(p-value)	(< 0.001)	(< 0.001)
y_{c_1}: measures of hemoglobin		
Intercept α_{c_1}	160.3758 ± 32.989	160.3775 ± 32.984
(p-value)	(< 0.001)	(< 0.001)
Treatment effect β_{c_1}	-12.492 ± 2.498	-12.496 ± 2.454
(p-value)	(< 0.001)	(< 0.001)
y_{c_2}: measures of white blood cell		
Intercept α_{c_2}	12.295 ± 9.331	12.2946 ± 9.515
(p-value)	(0.010)	(0.011)
Treatment effect β_{c_2}	-0.1591 ± 0.706	-0.1597 ± 0.702
(p-value)	(0.659)	(0.656)

Table 2.3 shows the main result of the treatment effect, and the complete result is presented in Table 2.5. We can observe that the statistical inference on the treatment effect through the two models was in agreement. The second treatment therapy results in lower

Table 2.4: The estimated parameters contain second moments of each outcome

	Esimated correlation				Estimated standard deviation
	Nausea	Diarrhea	HGB	WBC	
Nausea	1.0000	0.3954	0.0736	0.0899	-
Diarrhea		1.0000	0.0351	-0.0126	-
HGB			1.0000	0.0139	16.796
WBC				1.0000	4.7507

measures of hemoglobin and indicates a negative association with the occurrence of nausea and diarrhea, whereas the effect difference on the measures of white blood cells is insignificant. We can use the composite statistics to jointly assess the overall effect of this therapy on four responses. Table 2.4 provides the standard deviation and the correlation of four clinical outcomes estimated based on the proposed model.

Using the conventional approach, we cannot make statistical inferences across different linear models. The proposed model is able to test the hypothesis $H_0 : \beta_{b_1} = \beta_{b_2} = \beta_{c_1} = \beta_{c_2} = 0$ based on the asymptotical properties of the composite likelihood function. The test statistics of the composite Wald statistics under the H_0 is approximately 138.5890, the composite score statistics is 476.975, and the adjusted composite likelihood ratio is 264.3069, which are all greater than the critical value of χ^2_4. Therefore, we can reject the null hypothesis and conclude that two different treatments have a statistically significant difference in patient toxicity response. More specifically, in our estimation results, we infer there exists a significant occurrence difference of nausea and diarrhea and a significant difference in HGB between the two treatments.

Table 2.5: The estimated regression coefficients β and the standard deviation (sd). GLM: the estimation via the generalized linear model; CLM: estimation via the composite likelihood method; the column of $*$ lists the significant covariates.

	1.Nausea		2.Diarrhea		3.HBG		4.WBC		
Parameters	GLM	MMR	GLM	MMR	GLM	MMR	GLM	MMR	$*$
Intercept	-0.2685 (1.277)	-0.2793 (1.317)	0.6631 (1.304)	0.6741 (1.329)	160.3759 (16.831)	160.3775 (16.829)	12.2949 (4.761)	12.2946 (4.855)	3,4
Treatment	-0.2644 (0.097)	-0.2724 (0.099)	-0.6231 (0.098)	-0.6422 (0.101)	-12.4921 (1.274)	-12.4957 (1.252)	-0.1591 (0.360)	-0.1597 (0.358)	1,2,3
OS	-0.0037 (0.009)	-0.0039 (0.009)	0.0219 (0.009)	0.0228 (0.009)	0.5674 (0.115)	0.5675 (0.117)	-0.1764 (0.032)	-0.1764 (0.029)	2,3,4
OS event	0.0650 (0.195)	0.0702 (0.194)	0.1173(0.200)	0.1208 (0.194)	1.5154 (2.592)	1.5159 (2.614)	0.1290 (0.733)	0.1295 (0.562)	
PFS	0.0305 (0.020)	0.0315 (0.020)	0.0066 (0.021)	0.0062 (0.022)	-0.1628 (0.259)	-0.1627 (0.258)	0.1242 (0.073)	0.1243 (0.062)	
IN	-0.0204 (0.188)	-0.0200 (0.190)	-0.8622 (0.203)	-0.8907 (0.212)	-11.1599 (2.486)	-11.1621 (2.932)	-0.0809 (0.703)	-0.0803 (0.810)	2, 3
PD	0.1263 (0.136)	0.1319 (0.139)	-0.1375 (0.136)	-0.1432 (0.142)	-7.1013 (1.782)	-7.1005 (1.667)	-0.3695 (0.504)	-0.3689 (0.528)	3
PR	-0.0213 (0.169)	-0.0207 (0.171)	0.0691 (0.176)	0.0707 (0.174)	3.6782 (2.234)	3.6781 (2.210)	0.4003 (0.632)	0.4005 (0.5277)	
Age	-0.0142 (0.008)	-0.0091 (0.005)	0.0029 (0.008)	0.0018 (0.005)	-0.1463 (0.061)	-0.1464 (0.061)	-0.0382 (0.017)	-0.0382 (0.018)	3,4
Gender	-0.0086 (0.005)	-0.0091 (0.005)	0.0016 (0.005)	0.0018 (0.005)	9.0894 (1.798)	9.0836 (1.825)	-0.1079 (0.509)	-0.1096 (0.534)	1, 3
Colon	0.2313 (0.165)	0.2378 (0.169)	0.1017 (0.170)	0.1059 (0.177)	-1.1366 (2.189)	-1.1338 (2.105)	0.5706 (0.619)	0.5713 (0.676)	
Rectum	-0.0075 (0.154)	-0.0111 (0.157)	0.1760 (0.158)	0.1844 (0.166)	-1.6947 (2.032)	-1.6940 (1.913)	0.9096 (0.575)	0.9094 (0.649)	
Height	0.0048 (0.007)	0.0050 (0.008)	-0.0050 (0.007)	-0.0051 (0.008)	-0.0820 (0.097)	-0.0820 (0.097)	0.0237 (0.028)	0.0237 (0.028)	
Weight	0.0023 (0.003)	0.0024 (0.003)	0.0012 (0.003)	0.0012 (0.003)	0.0906 (0.044)	0.0907 (0.042)	-0.0189 (0.012)	-0.0189 (0.012)	3
PERF1	0.2046 (0.107)	0.2132 (0.110)	0.0898 (0.109)	0.0933 (0.110)	-4.8490 (1.414)	-4.8469 (1.352)	0.6028 (0.400)	0.6036 (0.380)	3
PERF2	0.5041 (0.187)	0.5244 (0.187)	0.2637 (0.189)	0.2692 (0.193)	-7.1494 (3.814)	-7.1455 (2.628)	1.8158 (0.688)	1.8176 (0.685)	1, 3,4
KRAS	-0.0407 (0.292)	-0.0387 (0.295)	0.0080 (0.293)	0.0150 (0.294)	1.3456 (3.814)	1.3455 (3.285)	-0.7323 (1.079)	-0.7325 (0.936)	

In section 4, we illustrate some estimation results of the regression parameters and correlation between outcomes. The complete analysis of this data consists of 68 parameters.

Chapter 3

Heterogeneous Multi-task Feature Learning with Mixed $\ell_{2,1}$ Regularization

In this chapter, the mixed $\ell_{2,1}$ regularized composite quasi-likelihood function is proposed to perform multi-task feature learning with different types of responses, including continuous and discrete responses. The theoretical results establish the sign recovery consistency and estimation error bounds of the penalized estimates under regularity conditions.

3.1 Introduction

Data integration, as a process of analyzing multiple related data sets simultaneously, can conduct joint inference by aggregating information from different sources. Statistical models, such as multi-task learning and fusion learning, have been proposed to conduct joint learning through a structured regularization for grouped parameters [Caruana, 1997, Ando and Zhang, 2005, Gao and Carroll, 2017, Zhang and Yang, 2017, Thung and Wee, 2018]. The mixed $\ell_{2,1}$ norm has been used for the grouped regularization to combine different statistical tasks, such as multivariate regression models [Liu et al., 2009] and multiple classification problems

[Obozinski et al., 2010, Zhou et al., 2011]. Rakotomamonjy et al. [2011] generalized the grouped penalization to a larger class of mixed norm penalties, such as the mixed $\ell_{q,r}$ norm with $q \geq 1$ and $0 \leq r \leq 1$. Furthermore, a variety of algorithms have been developed for differently structured mixed regularization [Argyriou et al., 2006, Gong et al., 2013, Jalali et al., 2010].

Lounici et al. [2011] showed that the regression coefficients estimated from multi-task learning satisfy the oracle inequality, and the result can be extended to non-Gaussian errors. The union support recovery of the multi-task feature learning was established by Obozinski et al. [2011], Negahban and Wainwright [2011], and Wang et al. [2015b] for both deterministic and random designs. Obozinski et al. [2011] proposed a sparsity-overlap function measuring the shared sparsities in the regression coefficient vectors for different responses. Multi-task learning often involves high-dimensional heterogeneous data sets [Gomez-Cabrero et al., 2014]. Most existing procedures focus on the same type of regression or classification problems across different tasks, where the response variables are either all continuous or all discrete. To deal with heterogeneous data sets with different types of response variables, Gao and Carroll [2017] proposed a method of fusion learning which uses the composite likelihood to combine the marginal likelihoods of different distributions across multiple tasks.

When the joint likelihood of heterogeneous data sets is difficult to formulate, the composite likelihood is a convenient likelihood-based method to perform joint estimation, inference, and feature selection. Even though the composite likelihood is not a true likelihood, the maximum composite likelihood estimates are still consistent and asymptotically normally distributed [Godambe, 1960, Lindsay, 1988, Cox and Reid, 2004, Varin, 2008, Gao and Song, 2010, Lindsay et al., 2011, Yi, 2014]. When the response variables are correlated across different tasks, the second Bartlett identity no longer holds. Namely, the covariance matrix of the composite score vectors is not equal to the negative Hessian matrix of the composite likelihood. Both matrices need to be separately estimated when we perform joint inference on

the correlated multiple tasks. Under this framework, different types of tasks, such as linear regression, Poisson regression, logistic regression, and multinomial regression, can be jointly analyzed through the multi-task feature learning. The features are shared by multiple tasks, and the design matrices of different tasks are allowed to be different.

In Gao and Carroll [2017], non-convex penalty functions, such as the group smoothly clipped absolute deviations (SCAD) penalty, is imposed on the composite likelihood constructed from multiple tasks to perform the joint sparse estimation. In this thesis, we propose to use the mixed $\ell_{2,1}$ norms to perform the group penalization on composite likelihood. We establish the union support recovery consistency and estimation error bounds of the penalized estimates under regularity conditions. In the composite likelihood approach, a distributional assumption is required to construct the marginal likelihood. To further relax the distributional assumptions, we propose to construct the negative quasi-likelihood as the individual loss function [Wedderburn, 1974] with a much-relaxed condition only on the moments of the response variables. Thus, the proposed composite quasi-likelihood method can provide robust joint sparse estimation without specific distributional assumptions.

The organization of the chapter is as follows. First, we set up the model for multi-task learning with correlated responses in Section 3.2. We provide the main theoretical results about the non-asymptotic error bound and the feature selection consistency of the proposed estimates. The method of numerical optimization is discussed in Section 3.3. The simulation studies are presented to demonstrate the feature selection accuracy and statistical consistency properties in Section 3.4. We provide two examples of multi-task learning on heterogeneous data sets in Section 3.5.

3.2 Methodology

Suppose there are K tasks of related interest, and each task has n_k independent responses $Y_k = (y_{k1}, \cdots, y_{kn_k})^T$, $k = 1, \cdots, K$ (See Table 3.1). In some multi-task learning data sets, the observations across different tasks can be correlated. For example, there could be measurements obtained by different techniques from the same set of experimental subjects, or the observations can be obtained from related subjects. The predictors for different tasks denoted by $(X_1, X_2, \cdots, X_K)$ can be same as the examples shown in Figure 1.1. When the integrated data sets can have different measurements, we assume the predictors to share some similarities. For example, the pth predictor $M_p = (X_{1p}, X_{2p}, \cdots, X_{Kp})$ in Table 3.1 represents the same type of feature in all related studies. For the case with $X_1 = X_2 = \cdots = X_K$, the predictors are obtained from one research to analyze the association with different responses. The parameters $\theta = (\theta_{11}, \theta_{12}, \cdots, \theta_{Kp_n}) \in \mathbb{R}^{Kp_n}$ include the regression coefficients of the predictors across K tasks. The overall effect of each predictor M_p across all tasks is represented by the grouped coefficients $\theta^{(p)} = (\theta_{1p}, \theta_{2p}, \cdots, \theta_{Kp})^T$ for any $p = 1, 2, \cdots, p_n$. The multi-task feature learning aims to select important features whose grouped coefficients have non-zero ℓ_2 norms, i.e., $\|\theta^{(p)}\|_2 \neq 0$. This is equivalent to the support union recovery, a practice that selects the features that have non-zero coefficients in at least one of the tasks [Obozinski et al., 2011, Negahban and Wainwright, 2011, Wang et al., 2015b]. Thus, we define the true union support of the parameter $\mathcal{S} := \{p : \|\theta^{(p)}\|_2 \neq 0\}$ with $|\mathcal{S}| = s$ and its complement can be denoted as $\mathcal{S}^c := \{p : \|\theta^{(p)}\|_2 = 0\}$.

For individual tasks, generalized linear models (GLM) can be applied to model the relationship between the responses and the predictors [McCullagh and Nelder, 1989]

$$g_k(E(y_{ki}|x_{k1i}, \ldots, x_{kp_ni})) = \eta_{ki} = x_{ki}^T \theta_k = \sum_{p=1}^{p_n} x_{kpi}\theta_{kp},$$

Table 3.1: Multiple tasks with a common set of predictors $M_1, M_2, \cdots, M_{p_n}$.

Response		Linear Predictors										
		M_1			M_2			M_p				M_{p_n}
Task 1:	Y_1	θ_{11}	X_{11}	+	θ_{12}	X_{12}	$+\cdots+$	θ_{1p}	X_{1p}	$+\cdots+$	θ_{1p_n}	X_{1p_n}
Task 2:	Y_2	θ_{21}	X_{21}	+	θ_{22}	X_{22}	$+\cdots+$	θ_{2p}	X_{2p}	$+\cdots+$	θ_{2p_n}	X_{2p_n}
								$\cdots$				
Task k:	Y_k	θ_{k1}	X_{k1}	+	θ_{k2}	X_{k2}	$+\cdots+$	θ_{kp}	X_{kp}	$+\cdots+$	θ_{kp_n}	X_{kp_n}
								$\cdots$				
Task K:	Y_K	$\underbrace{\theta_{K1}}_{\theta^{(1)}}$	X_{K1}	+	$\underbrace{\theta_{K2}}_{\theta^{(2)}}$	X_{K2}	$+\cdots+$	$\underbrace{\theta_{Kp}}_{\theta^{(p)}}$	X_{Kp}	$+\cdots+$	$\underbrace{\theta_{Kp_n}}_{\theta^{(p_n)}}$	X_{Kp_n}

In any kth task, the response Y_k consists of n_k observations as $Y_k = (y_{k1}, y_{k2}, \cdots, y_{kn_k})^T$, and the design matrix $X_k = (X_{k1}, X_{k2}, \cdots, X_{kp_n})$ with the pth column denoted as $X_{kp} = (x_{kp1}, x_{kp2}, \cdots, x_{kpn_k})^T$ and the ith row denoted as $x_{ki} = (x_{k1i}, \ldots, x_{kp_ni})^T$.

where each row observation of X_k is denoted by $x_{ki} = (x_{k1i}, \ldots, x_{kp_ni})^T \in \mathbb{R}^{p_n}$ and the regression coefficients are $\theta_k = (\theta_{k1}, \cdots, \theta_{kp_n})$. For each model, a task-specific link function $g_k(\cdot)$ is used.

Under the setting of the high dimensional model, the number of parameters p_n can increase to infinity with the sample size n_k. In order to jointly analyze multiple tasks and recover the correct model for all tasks, the following objective function is proposed:

$$Q(\theta) = \mathcal{L}(\theta) + \mathcal{R}(\theta), \tag{3.1}$$

where the loss function $\mathcal{L}(\theta)$ measures the fitting of multiple tasks, and the penalty function $\mathcal{R}(\theta)$ is a mixed $\ell_{2,1}$ grouped penalization on the parameters.

3.2.1 Composite Quasi Log-likelihood

The joint distribution of responses across multiple tasks can be difficult to model, especially when the responses are correlated across multiple tasks or they are of different types obtained from heterogeneous tasks. Instead of using the joint likelihood, the overall loss function across

multiple tasks can be based on the negative composite log-likelihood function [Godambe, 1960, Lindsay, 1988, Cox and Reid, 2004, Varin, 2008, Gao and Song, 2010, Yi, 2014]:

$$\mathcal{L}(\theta) = -\sum_{k=1}^{K} w_k \ell_k(\theta_k; Y_k),$$

where the positive weights w_k can be assigned based on the relative importance of the tasks and the individual marginal log-likelihood functions $\ell_k(\theta_k; Y_k)$ model the marginal distributions of different types of responses. Using this composite likelihood-based loss function, linear regression, Poisson regression, logistic regression, multinomial regression, and other types of learning tasks can be analyzed together under the multi-task feature learning framework.

When the response variables are assumed to follow distributions in the exponential family, the marginal log-likelihood functions are as follows,

$$\ell_k(\theta_k; Y_k) = \sum_{i=1}^{n_k} \ell_{ki}(\theta_k; y_{ki}) = \sum_{i=1}^{n_k} \frac{y_{ki}\beta_{ki} - b_k(\beta_{ki})}{\phi_k} + c(y_{ki}), \tag{3.2}$$

with the natural canonical parameter β_{ki}, the dispersion parameter $\phi_k > 0$, and the cumulant generating functions $b_k(.)$ assumed to be twice differentiable [McCullagh and Nelder, 1989]. In addition, the natural canonical parameter is related to the predictor values through the relationship: $\partial b_k(\beta_{ki})/\partial\beta_{ki} = E(y_{ki}|x_{k1i}, \ldots, x_{kp_ni}) = \mu_{ki} = g_k^{-1}(\eta_{ki})$, and $\eta_{ki} = \sum_{p=1}^{p_n} x_{kpi}\theta_{kp}$.

In many applications, the response variables may not follow a distribution that belongs to the exponential family. Without specific distributional assumption, the quasi log-likelihood function [Wedderburn, 1974] can be used to model the marginal distribution based on the assumptions of the first two moments:

$$\ell_k(\theta_k; Y_k) = \sum_{i=1}^{n_k} \ell_{ki}(\theta_k; y_{ki}) = \sum_{i=1}^{n_k} \int_{y_{ki}}^{g_k^{-1}(\eta_{ki})} \frac{y_{ki} - \mu}{V_k(\mu)\phi_k} d\mu, , \tag{3.3}$$

where

$$E(y_{ki}|x_{k1i},\ldots,x_{kp_ni}) = \mu_{ki} = g_k^{-1}(\eta_{ki}) = g_k^{-1}(\sum_{p=1}^{p_n} x_{kpi}\theta_{kp}),$$

and

$$\mathrm{Var}(y_{ki}|x_{k1i},\ldots,x_{kp_ni}) = \phi_k V_k(\mu_{ki}).$$

The inverse link function $g_k^{-1}(\eta)$ is a monotone function with respect to the linear predictor η. The variance function $V(u)$ models the relationship between variance and mean. It can take a wide variety of forms, including some polynomial forms u, u^2, u^3, or $u(1-u)$. In comparison with the log-likelihood formulation (3.2), the quasi log-likelihood formulation (3.3) is more general as it does not require knowledge of the specific underlying distribution. The assumptions are only based on the first two moments. It can be shown that (3.2) is equivalent to (3.3) when the underlying distribution indeed belongs to the exponential family. So the quasi-log-likelihood can be applied to a wider range of applications when the exponential family assumption cannot be verified. Throughout this chapter, the overall loss function will be based on the composite quasi log-likelihood as in (3.3), which comprises the log-likelihood (3.2) as a special case.

Assumption 3.1. *The individual quasi log-likelihood $\ell_k(\theta_k; Y_k)$ is a measurable function for all Y_k at any θ_k. It produces distinct values for different θ_k and it is twice differentiable as a function of θ_k. It is assumed that*

$$E_{\theta^*}\{\frac{\partial \ell_k(\theta_k; Y_k)}{\partial \theta_{kp}}\} = \boldsymbol{0},$$

and

$$E_{\theta^*}\Big\{\frac{\partial^2 \ell_k(\theta_k; Y_k)}{\partial \theta_{kp}\partial \theta_{kp'}}\Big\} = -E_{\theta^*}\Big\{\frac{\partial \ell_k(\theta_k; Y_k)}{\partial \theta_{kp}}\frac{\partial \ell_k(\theta_k; Y_k)}{\partial \theta_{kp'}}\Big\}$$

for any $p, p' = 1, 2, \cdots, p_n$, where θ^* is the true parameter vector.

Let $\nabla\mathcal{L}(\theta)$ denote the first derivative of the proposed loss function, where $\nabla\mathcal{L}(\theta)_{kp} = -\sum_{k=1}^{K} w_k \partial\ell_k(\theta_k; Y_k)/\partial\theta_{kp}$ for $k = 1, 2, \cdots, K$ and $p = 1, 2, \cdots, p_n$, which have the zero expectation with respect to θ^* [Yi, 2017]. Let $\nabla^2\mathcal{L}(\theta)$ denote the $Kp_n \times Kp_n$ Hessian matrix with each element of $\nabla^2\mathcal{L}(\theta)$ given by

$$\nabla^2\mathcal{L}(\theta)_{[kp,kp']} = -\sum_{k=1}^{K} w_k \frac{\partial^2\ell_k(\theta_k; Y_k)}{\partial\theta_{kp}\partial\theta_{kp'}},$$
$$\nabla^2\mathcal{L}(\theta)_{[kp,k'p']} = -\sum_{k=1}^{K} w_k \frac{\partial^2\ell_k(\theta_k; Y_k)}{\partial\theta_{kp}\partial\theta_{k'p'}} = 0$$

for any $p = 1, 2, \cdots, p_n$, $k, k^{'} = 1, 2, \cdots, K$ and $k \neq k^{'}$. For simplicity, we set $n_k = n$ across all tasks in the following analysis. The sensitivity matrix and variability matrix are defined as

$$H(\theta) = E\{n^{-1}\nabla^2\mathcal{L}(\theta)\} \text{ and } J(\theta) = \text{Cov}\{n^{-1}\nabla\mathcal{L}(\theta)\}.$$

In addition, the second Bartlett identity does not hold, namely, $H(\theta) \neq J(\theta)$ for the composite quasi log-likelihood function due to the correlations across different tasks.

Assumption 3.2. *The individual quasi log-likelihood functions admit third derivatives, and for any θ in the small neighborhood $\|\theta - \theta^*\|_2 \leq O_p(\sqrt{s\log(p_n)/n})$,*

$$\max_p \Lambda_{\max}(E(-\sum_{k=1}^{K} w_k \frac{\partial^3\ell_k(\theta_k; Y_k)}{\partial\theta\partial\theta^T\partial\theta_{kp}})) \leq \mathcal{W}^*.$$

with some constant $\mathcal{W}^* > 0$.

In addition, we apply the following boundedness condition to the general quasi-likelihood loss, which is similar to van de Geer and Müller [2012].

Assumption 3.3. *For any θ with $\|\theta - \theta^*\|_1 \leq r$, the inverse link function $g_k^{-1}(\eta_{ki})$ with η_{ki}*

evaluated at the point θ is twice differentiable and can satisfy

$$\max_{k,i} \left\{ \left| g_k^{-1}(\eta_{ki}) \right|, \left| \frac{\partial g_k^{-1}(\eta)}{\partial \eta} |_{\eta=\eta_{ki}} \right|, \left| \frac{\partial g_k^{-1}(\eta)}{\partial \eta} |_{\eta=\eta_{ki}} \right|^{-1}, \left| \frac{\partial^2 g_k^{-1}(\eta)}{\partial \eta^2} |_{\eta=\eta_{ki}} \right| \right\} = \mathcal{O}(1). \tag{3.4}$$

3.2.2 Mixed Regularization

The objective function (3.1) includes the penalty function to conduct joint feature selection across all tasks. The proposed penalty function $\mathcal{R}(\theta)$ uses the mixed $\ell_{2,1}$ regularization

$$\mathcal{R}(\theta) = n\lambda_n \|\theta\|_{2,1} = n\lambda_n \sum_{p=1}^{p_n} \|\theta^{(p)}\|_2 = n\lambda_n \sum_{p=1}^{p_n} (\sum_{k=1}^{K} \theta_{kp}^2)^{1/2},$$

with the penalty parameter λ_n. The mixed $\ell_{2,1}$ regularization implements feature selection over p_n elements, and each element is the ℓ_2 norms of the grouped parameters $\theta^{(p)}$ defined in (1.2) across K tasks [Liu et al., 2009, Obozinski et al., 2010, Zhou et al., 2011]. The estimation is different from the method to identify the sparse pattern for the individual feature in each task, in which the LASSO or group LASSO penalty function is commonly used [Tibshirani, 1996, Yuan and Lin, 2006].

Based on the definition proposed by Negahban et al. [2012], the mixed $\ell_{2,1}$ regularization is decomposable in following form,

$$\|\theta\|_{2,1} = \sum_{p=1}^{p_n} \|\theta^{(p)}\|_2 = \sum_{p\in\mathcal{E}} \|\theta^{(p)}\|_2 + \sum_{p\in\mathcal{E}^c} \|\theta^{(p)}\|_2 = \|\theta_{\mathcal{E}}\|_{2,1} + \|\theta_{\mathcal{E}^c}\|_{2,1},$$

where the subset $\mathcal{E} \subseteq \{1, 2, \cdots, p_n\}$. This property is essential for feature learning to construct the grouped norm of θ in different subspaces. In addition, the mixed $\ell_{2,1}$ regularization is a twice-differentiable and convex function with respect to non-zero parameters.

Let the subdifferential of the mixed $\ell_{2,1}$ norm be denoted by $z = (z_{11}, \ldots, z_{Kp_n})^T$, where

$$\begin{cases} z^{(p)} = \frac{\theta^{(p)}}{\|\theta^{(p)}\|_2}, & \text{if } \|\theta^{(p)}\|_2 \neq 0; \\ \|z^{(p)}\|_2 < 1, & \text{if } \|\theta^{(p)}\|_2 = 0, \end{cases}$$

for any $p = 1, 2, \cdots, p_n$. With any element in the subvector $\theta^{(p)}$ not equal to 0, the subdifferential has $\|\theta^{(p)}\|_2 \neq 0$, indicating that this group of features can be important for the learning tasks.

The penalized estimate is the solution of the estimating equation denoted by $\hat{\theta}$, such that for $\|\hat{\theta} - \theta^*\|_1 \leq \tilde{r}$ with some constant $\tilde{r}$,

$$\frac{1}{n}\nabla Q(\hat{\theta}) = \frac{1}{n}\nabla \mathcal{L}(\hat{\theta}) + \lambda_n \hat{z} = 0, \tag{3.5}$$

where $\hat{z}$ is the subdifferential of the mixed $\ell_{2,1}$ norm at the penalized estimate $\hat{\theta}$.

If $\hat{\theta}$ correctly recovers the true union support with $\text{supp}(\hat{\theta}) = \text{supp}(\theta^*)$, then

$$\begin{cases} -\frac{1}{n}\nabla \mathcal{L}(\hat{\theta})^{(p)} = \lambda_n \hat{z}^{(p)}, & \text{for any } p \in \mathcal{S}; \\ \|\frac{1}{n}\nabla \mathcal{L}(\hat{\theta})^{(p)}\|_2 < \lambda_n, & \text{for any } p \in \mathcal{S}^c. \end{cases}$$

Assumption 3.4. *There exist some constants $0 < 3k_1 + k_2 < 1$, such that $s = O(n^{k_1})$ and $\log(p_n) = O(n^{k_2})$. In addition, the true parameter vector $\|\theta^*\|_1 \leq R$ for some constant $R > 0$.*

This assumption about the size of the true model relative to the sample size is commonly imposed in high dimensional regularized estimation [Ravikumar et al., 2010, Li et al., 2021]. In addition, since we set the number of learning tasks K to be finite, the relation with sample size n is not specified in this assumption. Based on Obozinski et al. [2010], the data integration can be improved by increasing the number of related tasks, but it can lead to

longer running time for the computational programming. If the proposed model needs to handle the scenario with divergent K, we need further adjustment to Assumption 3.4 that the number of tasks is proportional to sample size at a polynomial rate.

3.2.3 Sufficient Conditions

In order to obtain the finite sample estimation error bound, we need to assume some concentration conditions. For example, the multivariate regression models are usually assumed with the Gaussian error [Lounici et al., 2011]. To analyze other types of response variables, different methods, such as cumulant boundedness condition or Rademacher complexity analysis, were used by Gao and Carroll [2017], Yousefi et al. [2018], and Fan et al. [2021]. The sub-Gaussian and sub-exponential conditions were imposed on the linear model errors [Negahban et al., 2012, Fang et al., 2020] and random design matrix [van de Geer et al., 2014]. In addition, Ning and Liu [2017] and Li et al. [2021] showed that the concentration conditions hold for high-dimensional generalized linear models. In the following, we assume similar concentration conditions for the multi-task learning problem.

Assumption 3.5. *For any kth task,*

1. *The error terms* $y_{ki} - g_k^{-1}(\eta_{ki}^*)$ *are independent samples from sub-exponential distributions with* ψ_1 *norm (1.3) bounded by some constant* A_0*;*

2. *The covariates in the design matrix satisfy* $\sup_{k,p,i}\{x_{kpi}\} \leq L < \infty$.

Since we aim to use the quasi-likelihood to model multiple heterogeneous data sets without specific distributional assumptions, the sub-exponential error terms are mild conditions to ensure the moments of data sets and tail probabilities are bounded. Thus, the concentration probabilities can be obtained in the following Lemma.

Lemma 3.1. *Under Assumptions 3.1 - 3.5, the composite score vector and Hessian matrix have the following concentration results:*

$$\begin{aligned}
\|\frac{1}{n}\nabla\mathcal{L}(\theta^*)\|_{2,\infty} =& O_p\big(\sqrt{\frac{K}{n}} + \sqrt{\frac{K\log(p_n)}{n}}\big),\\
\|\frac{1}{n}\nabla\mathcal{L}(\theta^*)\|_{\infty} =& O_p(\sqrt{\frac{1}{n}} + \sqrt{\frac{\log(p_n)}{n}}),\\
\sup_{k,p,p^{'}}\{\frac{1}{n}\nabla^2\mathcal{L}(\theta^*) - & H(\theta^*)\}_{[kp,kp']} = O_p(\sqrt{\frac{\log p_n}{n}}),
\end{aligned} \tag{3.6}$$

for any $k = 1, 2, \cdots, K$ and $p, p^{'} = 1, 2, \cdots, p_n$, where $\ell_{2,\infty}$ norm is defined in (1.1).

Proof. First, we need to analyze the distribution of the random variable $\|n^{-1}\nabla\mathcal{L}(\theta^*)^{(p)}\|_2$ for any $p = 1, 2, \cdots, p_n$. In Lemma 3.5, we have

$$\|\frac{1}{\sqrt{n}}\frac{\partial\ell_k(\theta_k^*; Y_k)}{\partial\theta_{kp}}\|_{\psi_1} \leq A_1.$$

with some constant A_1, and by the definition of grouped ℓ_2 norm, we can show that

$$\begin{aligned}
\|n^{-1}\nabla\mathcal{L}(\theta^*)^{(p)}\|_2 &= \big(\sum_{k=1}^{K}(\frac{1}{n}\frac{\partial\ell_k(\theta_k^*; Y_k)}{\partial\theta_{kp}})^2\big)^{1/2}\\
&= \big(\sum_{k=1}^{K}(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\ell_{ki}(\theta_k^*; y_{ki})}{\partial\theta_{kp}})^2\big)^{1/2}\\
&= \big(\frac{1}{n}\sum_{k=1}^{K}(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial\ell_{ki}(\theta_k^*; y_{ki})}{\partial\theta_{kp}})^2\big)^{1/2}.
\end{aligned}$$

This result can be used to bound the sub-exponential norm of $\|n^{-1}\nabla\mathcal{L}(\theta^*)^{(p)}\|_2$ by applying Minkowski's Inequality,

$$\|\|n^{-1}\nabla\mathcal{L}(\theta^*)^{(p)}\|_2\|_{\psi_1} = \|\big(\frac{1}{n}\sum_{k=1}^{K}(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial\ell_{ki}(\theta_k^*; y_{ki})}{\partial\theta_{kp}})^2\big)^{1/2}\|_{\psi_1}$$

$$= \sup_{m\geq 1} \frac{1}{m}\big(E(|\big(\frac{1}{n}\sum_{k=1}^{K}(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial \ell_{ki}(\theta_k^*;y_{ki})}{\partial \theta_{kp}})^2\big)^{1/2}|^m)\big)^{1/m}$$
$$\leq \frac{K}{\sqrt{n}}\|\frac{1}{\sqrt{n}}\frac{\partial \ell_k(\theta_k^*;Y_k)}{\partial \theta_{kp}}\|_{\psi_1} \leq \frac{K}{\sqrt{n}}A_1 < \infty.$$

Furthermore, we can show that

$$E(\|n^{-1}\nabla\mathcal{L}(\theta^*)^{(p)}\|_2) \leq \big\{\frac{1}{n}\sum_{k=1}^{K}E[(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial \ell_{ki}(\theta_k^*;y_{ki})}{\partial \theta_{kp}})^2]\big\}^{1/2} \leq A_1\sqrt{\frac{K}{n}},$$
$$var(\|n^{-1}\nabla\mathcal{L}(\theta^*)^{(p)}\|_2) \leq \frac{1}{n}\sum_{k=1}^{K}E[(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial \ell_{ki}(\theta_k^*;y_{ki})}{\partial \theta_{kp}})^2] \leq 2A_1^2\frac{K}{n}.$$

This implies that $\|n^{-1}\nabla\mathcal{L}(\theta^*)^{(p)}\|_2$ satisfies the sub-exponential property, such that with small δ,

$$P(\|n^{-1}\nabla\mathcal{L}(\theta^*)^{(p)}\|_2 \geq E(\|n^{-1}\nabla\mathcal{L}(\theta^*)^{(p)}\|_2) + \delta) \leq 2\exp\{-\alpha\frac{\delta^2}{2KA_1^2}n\}.$$

Next, we take the supremum of $\|n^{-1}\nabla\mathcal{L}(\theta^*)^{(p)}\|_2$ over $p = 1, 2, \cdots, p_n$,

$$P(\sup_p \|n^{-1}\nabla\mathcal{L}(\theta^*)^{(p)}\|_2 \geq E(\|n^{-1}\nabla\mathcal{L}(\theta^*)^{(p)}\|_2) + \delta) \leq 2p_n\exp\{-\alpha\frac{\delta^2 n}{2KA_1^2}\}.$$

By combining all the results above, with $\delta = A_1\sqrt{2K(1+d)\log(p_n)/(\alpha n)}$ for some constant $d > 1$, we show that with a probability at least $1 - 2p_n^{-d}$,

$$\sup_p \|\frac{1}{n}\nabla\mathcal{L}(\theta^*)^{(p)}\|_2 \leq A_1\big(\sqrt{\frac{K}{n}} + \sqrt{\frac{2K(1+d)\log(p_n)}{\alpha n}}\big).$$

In addition, we showed that the score function can hold the sub-exponential condition from

Lemma 3.5. Therefore, we have

$$P(\|\frac{1}{n}\nabla\mathcal{L}(\theta^*)\|_\infty \geq \varepsilon) \leq Kp_n \max_p P(\frac{1}{n}\frac{\partial \ell_k(\theta_k^*; Y_k)}{\partial \theta_{kp}} \geq \varepsilon) \leq 2Kp_n \exp\{-\frac{\varepsilon^2}{A_1^2}n\},$$

which can imply

$$\|\frac{1}{n}\nabla\mathcal{L}(\theta^*)\|_\infty \leq A_1\big(\sqrt{\frac{1}{n}} + \sqrt{\frac{2(d+1)\log(p_n)}{(\alpha n)}}\big),$$

with a probability at least $1 - 2\exp\{-d\log(p_n) + \log(K)\}$ as claimed in (3.6).

The second part of Lemma 3.1 shows that the difference between the random Hessian and its expectation is bounded. When the tasks are modeled by the canonical link and the response variables are from the exponential family, the Hessian matrix is deterministic, and $n^{-1}\nabla\mathcal{L}(\theta^*) = H(\theta^*)$. For general cases, the entries of the random Hessian of the composite quasi-likelihood are

$$\begin{aligned}\frac{1}{n}\frac{\partial^2}{\partial\theta_{kp}\partial\theta_{kp'}}\mathcal{L}(\theta^*) = &\underbrace{\frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n}\frac{1}{\phi_k V(g_k^{-1}(\eta_{ki}^*))}\left\{\frac{\partial g_k^{-1}(\eta_{ki})}{\partial\eta_{ki}}\frac{\partial\eta_{ki}}{\partial\theta_{kp}}\right\}\left\{\frac{\partial g_k^{-1}(\eta_{ki})}{\partial\eta_{ki}}\frac{\partial\eta_{ki}}{\partial\theta_{kp'}}\right\}}_{\mathcal{I}_1} \\ &- \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n}\underbrace{\left(y_{ki} - g_k^{-1}(\eta_{ki}^*)\right)}_{\mathcal{I}_2}\underbrace{\frac{\partial}{\partial\theta_{kp'}}\left\{\frac{1}{\phi_k V(g_k^{-1}(\eta_{ki}^*))}\frac{\partial g_k^{-1}(\eta_{ki})}{\partial\eta_{ki}}\frac{\partial\eta_{ki}}{\partial\theta_{kp}}\right\}}_{\mathcal{I}_3}.\end{aligned}$$

The component $\mathcal{I}_1$ is equal to the corresponding element in the sensitivity matrix $H(\theta^*)$. With some special link functions, the component $\mathcal{I}_3$ can be equal to zero. For the models with the general quasi-likelihood settings, we can show that the component $\mathcal{I}_3$ can be bounded by universal constant $\mathcal{K} > 0$ across all tasks with similar derivation as Lemma 3.6. Based on Assumption 3.5, the variables $n^{-1}\nabla^2\mathcal{L}(\theta^*)_{[kp,kp']} - H(\theta^*)_{[kp,kp']}$ satisfy the sub-exponential condition with mean zero and the ψ_1 norm bounded by $\mathcal{K}A_0 < A_1$ for some universal constant

A_1. Therefore, we have the concentration result of the random Hessian matrix

$$\sup_{k,p,p^{'}} \{\frac{1}{n}\nabla^2 \mathcal{L}(\theta^*) - H(\theta^*)\}_{[kp,kp']} = O_p(\sqrt{\frac{\log p_n}{n}}),$$

for any $k = 1, 2, \cdots, K$ and $p, p^{'} = 1, 2, \cdots, p_n$.

□

Next, we introduce the restricted eigenvalue (RE) condition for the design matrix, which was commonly used for regularized regression models [Bickel et al., 2009, van de Geer and Bühlmann, 2009, Meinshausen and Yu, 2009].

Assumption 3.6. *Define $S_{n,k} = n^{-1}\sum_{i=1}^{n} x_{ki}x_{ki}^T$ for the kth task. There exist $m = c_0 Ks$ for some $c_0 > 0$ and some positive constants γ, ρ_- and ρ_+, such that the restricted minimum and maximum eigenvalues of the design matrix*

$$\rho_-(m,\gamma) = \inf_k \left\{ u^T S_{n,k} u : u \in \mathcal{C}(m,\gamma) \right\}, \text{ and}$$
$$\rho_+(m,\gamma) = \sup_k \left\{ u^T S_{n,k} u : u \in \mathcal{C}(m,\gamma) \right\}$$

are bounded by

$$0 < \rho_- \leq \rho_-(m,\gamma) < \rho_+(m,\gamma) \leq \rho_+ < \infty,$$

where $\mathcal{C}(m,\gamma) := \{u : \mathcal{S} \subset \mathcal{J}, |\mathcal{J}| < m, \|u_{\mathcal{J}^c}\|_1 \leq \gamma \|u_{\mathcal{J}}\|_1\}$.

Since the Hessian matrix of the proposed loss function depends on the parameter θ, we need to show that the restricted eigenvalue condition holds for the Hessian matrix under the model assumptions.

Lemma 3.2. *Under Assumptions 3.1 - 3.5, there exist positive constants γ, κ_-, κ_+, and m such that the expected Hessian matrix of the composite quasi-likelihood loss function $H(\theta)$ satisfies the restricted eigenvalue (RE) condition*

$$0 < \kappa_- \leq \kappa_-(m,\gamma) < \kappa_+(m,\gamma) \leq \kappa_+ < \infty,$$

where

$$\kappa_+(m,\gamma) = \sup_{u,\theta}\{u^T H(\theta)u : u \in \mathcal{C}(m,\gamma)\};$$
$$\kappa_-(m,\gamma) = \inf_{u,\theta}\{u^T H(\theta)u : u \in \mathcal{C}(m,\gamma)\},$$

with $\mathcal{C}(m,\gamma) \equiv \{u : \mathcal{S} \subset \mathcal{J}, |\mathcal{J}| \leq m, \|u_{\mathcal{J}^c}\|_1 \leq \gamma\|u_{\mathcal{J}}\|_1\}$. Furthermore, there exists some $\tilde{r} > 0$, for any point θ with $\|\theta - \theta^*\|_1 \leq \tilde{r}$, the observed Hessian of the composite quasi-likelihood loss function $n^{-1}\nabla^2\mathcal{L}(\theta)$ satisfies the restricted eigenvalue condition with a probability tending to 1.

Proof. Based on Lemma 3.6, the Hessian matrix of the composite quasi-likelihood is given by

$$\frac{1}{n}\nabla^2\mathcal{L}(\theta) = \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n}\{f_1(\eta_{ki}) - (y_{ki} - g_k^{-1}(\eta_{ki}^*))f_2(\eta_{ki})\}x_{ki}x_{ki}^T,$$

and there exist some positive constants α_0, α_1, and α_2, such that $\alpha_0 < f_1(\eta_{ki}) < \alpha_1$ and $|f_2(\eta_{ki})| < \alpha_2$.

When the parameters are partitioned into subsets of different tasks, the Hessian matrix is in the form of a diagonal block matrix. We show that the minimum and maximum eigenvalues

of the Hessian matrix are given by

$$\min \Lambda(\frac{1}{n}\nabla^2\mathcal{L}(\theta)) = \inf_k \left\{ u^T \frac{1}{n}\sum_{i=1}^{n} \left\{ f_1(\eta_{ki}) - (y_{ki} - g_k^{-1}(\eta_{ki}^*)) f_2(\eta_{ki}) \right\} x_{ki} x_{ki}^T u \right\};$$
$$\max \Lambda(\frac{1}{n}\nabla^2\mathcal{L}(\theta)) = \sup_k \left\{ u^T \frac{1}{n}\sum_{i=1}^{n} \left\{ f_1(\eta_{ki}) - (y_{ki} - g_k^{-1}(\eta_{ki}^*)) f_2(\eta_{ki}) \right\} x_{ki} x_{ki}^T u \right\}.$$

We have

$$u^T \left\{ \frac{1}{n}\sum_{i=1}^{n} f_1(\eta_{ki}) x_{ki} x_{ki}^T \right\} u = \frac{1}{n}\sum_{i=1}^{n} f_1(\eta_{ki}) u^T x_{ki} x_{ki}^T u \geq \alpha_0 \rho_-.$$

We apply Hölder's inequality and get

$$u^T \left\{ \frac{1}{n}\sum_{i=1}^{n} (y_{ki} - g_k^{-1}(\eta_{ki}^*)) f_2(\eta_{ki}) x_{ki} x_{ki}^T \right\} u \leq \max_{k,i} \{(x_{ki}^T u)^2\} \frac{1}{n}\sum_{i=1}^{n} |(y_{ki} - g_k^{-1}(\eta_{ki}^*)) f_2(\eta_{ki})|.$$

Based on Assumption 3.5, $\|x_{ki}\|_\infty \leq L$ across all tasks and $\|u_{\mathcal{J}^c}\|_1 \leq \gamma \|u_{\mathcal{J}}\|_1$ with $|\mathcal{J}| \leq m = c_0 K s$, we have

$$x_{ki}^T u \leq \|x_{ki}\|_\infty \|u\|_1 \leq (1+\gamma) \|x_{ki}\|_\infty \|u_{\mathcal{J}}\|_1 \leq (1+\gamma)\sqrt{|\mathcal{J}|} L.$$

In addition, the variables $y_{ki} - g_k^{-1}(\eta_{ki}^*)$ follow sub-exponential distributions based on Assumption 3.5. We obtain that with a probability at least $1 - 2\exp\{-c\log(p_n)\}$ for some constant $c = (\alpha_2 A_0)^{-2} > 0$,

$$\frac{1}{n}\sum_{i=1}^{n} |y_{ki} - g_k^{-1}(\eta_{ki}^*)| f_2(\eta_{ki}) \leq \sqrt{\frac{2\log p_n}{n}}.$$

Therefore, there exists some $\kappa_- < \alpha_0 \rho_-$. If the sample size is sufficiently large

$$n \geq \left(\frac{c_0 (1+\gamma)^2 L^2 K}{\alpha_0 \rho_- - \kappa_-} \right)^2 2 s^2 \log p_n,$$

then we obtain the lower bound for the minimum eigenvalue of the Hessian matrix

$$u^T \frac{1}{n} \nabla^2 \mathcal{L}(\theta) u \geq \alpha_0 \rho_- - c_0 (1+\gamma)^2 L^2 K s \sqrt{\frac{2\log(p_n)}{n}} \geq \kappa_- > 0.$$

Similarly, the upper bound can be obtained using a similar approach

$$u^T \frac{1}{n} \nabla^2 \mathcal{L}(\theta) u \leq \alpha_1 \rho_+ + c_0 (1+\gamma)^2 L^2 K s \sqrt{\frac{2\log(p_n)}{n}} \leq \kappa_+ < \infty.$$

Combining the results above, the random Hessian matrix satisfies the restricted eigenvalue condition with high probability.

□

Lemma 3.2 guarantees that with a probability tending to 1,

$$(\frac{1}{n} \nabla \mathcal{L}(\theta + \Delta) - \frac{1}{n} \nabla \mathcal{L}(\theta))^T \Delta \geq \kappa_- \|\Delta\|_2^2, \text{ with } \Delta \in \mathcal{C}(m, \gamma).$$

This implies that with high probability, the optimization problem has a stationary point which satisfies the sparsity requirement [Loh and Wainwright, 2015, Fan et al., 2018, Sun et al., 2020].

The mutual incoherence condition is used to control the dependency between the predictors in the true model and the other unimportant predictors, which is a necessary condition to ensure the true support recovery with ℓ_1 regularization [Zhao and Yu, 2006, van de Geer and Bühlmann, 2009, Loh and Wainwright, 2017, Jalali et al., 2010]. For high dimensional models with mixed regularization, the block-wise mutual incoherence was proposed to ensure the group selection consistency [Bach, 2008, Eldar et al., 2010, Hebiri and van de Geer, 2011].

Assumption 3.7. *Let the sub-matrices of the expected Hessian matrix be denoted by* $H^*_{\mathcal{SS}} = E_{\theta^*}[n^{-1}\nabla^2 \mathcal{L}(\theta^*)_{\mathcal{SS}}]$ *and* $H^*_{\mathcal{S}^c\mathcal{S}} = E_{\theta^*}[n^{-1}\nabla^2 \mathcal{L}(\theta^*)_{\mathcal{S}^c\mathcal{S}}]$, *where* $\mathcal{S}$ *is the support of non-zero*

parameters. For some constant $\xi \in (0,1)$*, the inequality holds*

$$\sqrt{K}\left|\left|\left|H^*_{\mathcal{S}^c\mathcal{S}}[H^*_{\mathcal{S}\mathcal{S}}]^{-1}\right|\right|\right|_\infty \leq 1-\xi.$$

When the expected Hessian matrix satisfies the mutual incoherence condition above, it can be shown that the observed Hessian matrix holds a similar condition with a probability tending to one.

Lemma 3.3. *Under Assumptions 3.1 - 3.7, let* $\xi \in (0,1)$*, the following condition*

$$\sqrt{K}\left|\left|\left|\frac{1}{n}\nabla^2\mathcal{L}(\theta^*)_{\mathcal{S}^c\mathcal{S}}(\frac{1}{n}\nabla^2\mathcal{L}(\theta^*)_{\mathcal{S}\mathcal{S}})^{-1}\right|\right|\right|_\infty < 1-\frac{\xi}{2}$$

holds with a probability at least $1-4K\exp\{-C_0\xi^2 n/s^3+2\log(p_n)\}$ *for some universal constant* $C_0 > 0$.

Proof. The proof of lemma 3.3 is analogous to previous work in Ravikumar et al. [2010]. For simplicity, let the sub-matrix of random Hessian $n^{-1}\nabla^2\mathcal{L}(\theta^*)_{\mathcal{S}\mathcal{S}} = \mathcal{H}^*_{\mathcal{S}\mathcal{S}}$, and let the difference of the matrices be denoted by $\Delta H^*_{\mathcal{S}\mathcal{S}} = \mathcal{H}^*_{\mathcal{S}\mathcal{S}} - H(\theta^*)_{\mathcal{S}\mathcal{S}}$. Because the sub-matrices of random Hessian are diagonal block matrices, we show that

$$H^*_{\mathcal{S}\mathcal{S}} = \text{diag}({}_kH^*_{\mathcal{S}\mathcal{S}})_{k=1}^K,$$

where the sub-matrix ${}_kH^*_{\mathcal{S}\mathcal{S}} \in \mathbb{R}^{s\times s}$ represents the kth block in $H^*_{\mathcal{S}\mathcal{S}}$. The difference between sub-matrices is denoted as $\Delta_k H^*_{\mathcal{S}\mathcal{S}} = [{}_k\mathcal{H}^*_{\mathcal{S}\mathcal{S}} -{}_k H(\theta^*)_{\mathcal{S}\mathcal{S}}]$.

We need to obtain the concentration result of the inverse matrix difference $[\mathcal{H}^*_{\mathcal{S}\mathcal{S}}]^{-1} - [H^*_{\mathcal{S}\mathcal{S}}]^{-1}$. Based on Lemma 3.9, we show that the diagonal block matrix

$$\left|\left|\left|[\mathcal{H}^*_{\mathcal{S}\mathcal{S}}]^{-1} - [H^*_{\mathcal{S}\mathcal{S}}]^{-1}\right|\right|\right|_\infty = \sup_k \left|\left|\left|[{}_k\mathcal{H}^*_{\mathcal{S}\mathcal{S}}]^{-1} - [{}_kH^*_{\mathcal{S}\mathcal{S}}]^{-1}\right|\right|\right|_\infty,$$

so that

$$\begin{aligned}
\big|\big|\big|[{}_k\mathcal{H}^*_{\mathcal{SS}}]^{-1} - [{}_kH^*_{\mathcal{SS}}]^{-1}\big|\big|\big|_\infty &= \big|\big|\big|[{}_kH^*_{\mathcal{SS}}]^{-1}\Delta_k H^*_{\mathcal{SS}}[{}_k\mathcal{H}^*_{\mathcal{SS}}]^{-1}\big|\big|\big|_\infty \\
&\overset{(i)}{\leq} \sqrt{s}\big|\big|\big|[{}_kH^*_{\mathcal{SS}}]^{-1}\big|\big|\big|_2 |||\Delta_k H^*_{\mathcal{SS}}|||_2 \big|\big|\big|[{}_k\mathcal{H}^*_{\mathcal{SS}}]^{-1}\big|\big|\big|_2 \\
&\leq \frac{\sqrt{s}}{\kappa_-}|||\Delta_k H^*_{\mathcal{SS}}|||_2 \big|\big|\big|[{}_k\mathcal{H}^*_{\mathcal{SS}}]^{-1}\big|\big|\big|_2.
\end{aligned}$$

In step (i), we apply the inequality between matrix norms and the Cauchy–Schwarz inequality. We have

$$\begin{aligned}
P(\big|\big|\big|[\mathcal{H}^*_{\mathcal{SS}}]^{-1} - [H^*_{\mathcal{SS}}]^{-1}\big|\big|\big|_\infty \geq \varepsilon) &\leq K \sup_k P(\frac{\sqrt{s}}{\kappa_-}|||\Delta_k H^*_{\mathcal{SS}}|||_2 \big|\big|\big|[{}_k\mathcal{H}^*_{\mathcal{SS}}]^{-1}\big|\big|\big|_2 \geq \varepsilon) \\
&\overset{(i)}{\leq} K \sup_k P(\{|||\Delta_k H^*_{\mathcal{SS}}|||_2 \geq \frac{\varepsilon\kappa_-^2}{\sqrt{s}}\} \cup \{|||\Delta_k H^*_{\mathcal{SS}}|||_2 > \varepsilon\} \\
&\leq 2K \exp\big\{ - \frac{\alpha\kappa_-^4\varepsilon^2}{A_1^2 s^3}n + 2\log(s)\big\}.
\end{aligned}$$

The step (i) can be obtained based on the derivation 3.18 and 3.20. This probability is exponentially small as $n > cs^3\log(p_n)$ with some constant c.

We combine all the concentration results and obtain

$$\begin{aligned}
\mathcal{H}^*_{\mathcal{S}^c\mathcal{S}}(\mathcal{H}^*_{\mathcal{SS}})^{-1} &= [H^*_{\mathcal{S}^c\mathcal{S}} + \Delta H^*_{\mathcal{S}^c\mathcal{S}}][[H^*_{\mathcal{SS}}]^{-1} + [\mathcal{H}^*_{\mathcal{SS}}]^{-1} - [H^*_{\mathcal{SS}}]^{-1}] \\
&= \underbrace{H^*_{\mathcal{S}^c\mathcal{S}}(H^*_{\mathcal{SS}})^{-1}}_{\mathcal{I}_1} + \underbrace{H^*_{\mathcal{S}^c\mathcal{S}}([\mathcal{H}^*_{\mathcal{SS}}]^{-1} - [H^*_{\mathcal{SS}}]^{-1})}_{\mathcal{I}_2} \\
&+ \underbrace{\Delta H^*_{\mathcal{S}^c\mathcal{S}}(H^*_{\mathcal{SS}})^{-1}}_{\mathcal{I}_3} + \underbrace{\Delta H^*_{\mathcal{S}^c\mathcal{S}}([\mathcal{H}^*_{\mathcal{SS}}]^{-1} - [H^*_{\mathcal{SS}}]^{-1})}_{\mathcal{I}_4}.
\end{aligned}$$

We have the component $\sqrt{K}|||\mathcal{I}_1|||_\infty \leq (1-\xi)$ based on Assumption 3.7. For the second component $\mathcal{I}_2$, we apply Lemma 3.7 to obtain that with a probability at least $1 - 2K\exp\{-\frac{\alpha\kappa_-^2\varepsilon^2}{A_1^2 s^3}n +$

$2\log(s)\}$,

$$
\begin{aligned}
|||\mathcal{I}_2|||_\infty &\leq \Big|\Big|\Big|H^*_{\mathcal{S}^c\mathcal{S}}(H^*_{\mathcal{S}\mathcal{S}})^{-1}\Big|\Big|\Big|_\infty \Big|\Big|\Big|\Delta H^*_{\mathcal{S}\mathcal{S}}(\mathcal{H}^*_{\mathcal{S}\mathcal{S}})^{-1}\Big|\Big|\Big|_\infty \\
&\leq \Big|\Big|\Big|H^*_{\mathcal{S}^c\mathcal{S}}(H^*_{\mathcal{S}\mathcal{S}})^{-1}\Big|\Big|\Big|_\infty \sup_k\{|||\Delta_k H^*_{\mathcal{S}\mathcal{S}}|||_\infty \Big|\Big|\Big|({}_k\mathcal{H}^*_{\mathcal{S}\mathcal{S}})^{-1}\Big|\Big|\Big|_\infty\} \\
&< \frac{1-\xi}{\sqrt{K}} \times \left\{\frac{\kappa_-\varepsilon}{\sqrt{s}}\right\} \times \left\{\frac{\sqrt{s}}{\kappa_-}\right\} = \frac{1-\xi}{\sqrt{K}} \times \varepsilon^{'}.
\end{aligned}
$$

Based on Lemmas 3.7 and 3.8, the concentration result of the component $\mathcal{I}_3$ and $\mathcal{I}_4$ can be obtained with a probability at least $1-2K\exp\big\{-\frac{\alpha\kappa_-^4\varepsilon^{'2}}{A_1^2 s^3}n+2\log(s)\big\}-2K\exp\big\{-\frac{\alpha\kappa_-^2\varepsilon^2}{A_1^2 s^3}n+\log(s)+\log(p_n-s)\big\}$,

$$
\begin{aligned}
|||\mathcal{I}_3|||_\infty &\leq |||\Delta H^*_{\mathcal{S}^c\mathcal{S}}|||_\infty \sup_k\{\sqrt{s}\Big|\Big|\Big|({}_kH^*_{\mathcal{S}\mathcal{S}})^{-1}\Big|\Big|\Big|_2\} \leq \left\{\frac{\varepsilon\kappa_-}{\sqrt{s}}\right\}\left\{\frac{\sqrt{s}}{\kappa_-}\right\} = \varepsilon \\
|||\mathcal{I}_4|||_\infty &\leq |||\Delta H^*_{\mathcal{S}^c\mathcal{S}}|||_\infty \Big|\Big|\Big|\Delta[H^*_{\mathcal{S}\mathcal{S}}]^{-1}\Big|\Big|\Big|_\infty \leq \varepsilon \times \varepsilon^{'}.
\end{aligned}
$$

We set $\varepsilon \leq \xi/(4\sqrt{K})$ and $\varepsilon^{'} \leq \xi$ that leads to

$$
\begin{aligned}
\sqrt{K}\Big|\Big|\Big|\mathcal{H}^*_{\mathcal{S}^c\mathcal{S}}(\mathcal{H}^*_{\mathcal{S}\mathcal{S}})^{-1}\Big|\Big|\Big|_\infty &< 1-\xi+\sqrt{K}\varepsilon\times\varepsilon^{'}+(1-\xi)\varepsilon+\sqrt{K}\varepsilon \\
&< (1-\xi)+\frac{\xi^2}{4}+\frac{\xi-\xi^2}{4}+\frac{\xi}{4} < 1-\frac{1}{2}\xi
\end{aligned}
$$

with a probability $1-4K\exp\big\{-C_0\xi^2 n/s^3+2\log(p_n)\big\}$ for a universal constant $C_0>0$. □

3.2.4 Selection Consistency and Estimation Error Bound

This section establishes the finite sample estimation error bound and model selection consistency for the penalized estimate.

Lemma 3.4. *Let $\mathcal{E}$ be a subset of $\{1,\ldots,p_n\}$ such that $S\subseteq\mathcal{E}$ and $|\mathcal{E}|=c_1 s$ with some positive constant c_1. Under Assumptions 3.1 - 3.6, suppose with constants $\alpha>0$ and $d>1$,*

the penalty parameter satisfies

$$\lambda_n \geq \frac{4A_1}{\xi}\Big(\sqrt{\frac{K}{n}} + \sqrt{\frac{2(d+1)K\log(p_n)}{\alpha n}}\Big) \tag{3.7}$$

where $A_1 \geq \mathcal{K}A_0$ *for some* $\mathcal{K} > 0$*, and the composite score vector satisfies the inequality* $\|n^{-1}\nabla\mathcal{L}(\theta^*)\|_\infty \leq \lambda_n/(2\sqrt{K})$*, then there exist a optimal solution* $\hat{\theta}$ *of the the estimating equation (3.5) such that* $\|\hat{\theta} - \theta^*\|_1 \leq \tilde{r}$*, and*

$$\|(\hat{\theta} - \theta^*)_{\mathcal{E}^c}\|_1 \leq (2\sqrt{K} + 1)\|(\hat{\theta} - \theta^*)_{\mathcal{E}}\|_1. \tag{3.8}$$

Proof. The first-order partial derivative of the objective function can be expanded by applying the mean value theorem,

$$\mathbf{0} = \nabla Q(\hat{\theta}) = \nabla\mathcal{L}(\theta^*) + \nabla^2\mathcal{L}(\tilde{\theta})(\hat{\theta} - \theta^*) + n\lambda_n\hat{z},$$

where $\tilde{\theta} = \alpha\theta^* + (1-\alpha)\hat{\theta}$ for some $\alpha \in (0,1)$. This entails

$$\nabla Q(\hat{\theta})^T(\hat{\theta} - \theta^*) = (\nabla\mathcal{L}(\theta^*) + n\lambda_n\hat{z})^T(\hat{\theta} - \theta^*) + (\hat{\theta} - \theta^*)^T\nabla^2\mathcal{L}(\tilde{\theta})(\hat{\theta} - \theta^*). \tag{3.9}$$

Based on Lemma 3.6, we can show that with a probability tending to 1,

$$(\hat{\theta} - \theta^*)^T\frac{1}{n}\nabla^2\mathcal{L}(\tilde{\theta})(\hat{\theta} - \theta^*) \geq 0.$$

Thus, we can construct the inequality from 3.9 as follows,

$$\underbrace{\nabla Q(\hat{\theta})^T(\hat{\theta} - \theta^*)}_{\mathcal{I}_1} - \underbrace{\nabla\mathcal{L}(\theta^*)(\hat{\theta} - \theta^*)}_{\mathcal{I}_2} - n\lambda_n\underbrace{\hat{z}^T(\hat{\theta} - \theta^*)}_{\mathcal{I}_3} \geq 0. \tag{3.10}$$

For the exact solution $\hat{\theta}$, all elements in $\nabla Q(\hat{\theta})$ are zero so that the component $\mathcal{I}_1 = 0$. The elements in the vector $(\hat{\theta} - \theta^*) \in \mathcal{R}^{Kp_n}$ can be decomposed into two subsets $\mathcal{E}$ and $\mathcal{E}^c$. By applying Hölder's inequality, the components $\mathcal{I}_2$ from the equation 3.10 can be bounded above as follows

$$\begin{aligned} \mathcal{I}_2 : -\nabla\mathcal{L}(\theta^*)^T(\hat{\theta} - \theta^*) \leq & \|\nabla\mathcal{L}(\theta^*)\|_\infty \|\hat{\theta} - \theta^*\|_1 \\ = & \|\nabla\mathcal{L}(\theta^*)\|_\infty (\|(\hat{\theta} - \theta^*)_\mathcal{E}\|_1 + \|(\hat{\theta} - \theta^*)_{\mathcal{E}^c}\|_1). \end{aligned} \tag{3.11}$$

By the definition, if $\hat{\theta}^{(p)} \neq \mathbf{0}$, $\hat{z}^{(p)} = \hat{\theta}^{(p)}/\|\hat{\theta}^{(p)}\|_2$, and if $\hat{\theta}^{(p)} = \mathbf{0}$, $\|\hat{z}^{(p)}\|_2 < 1$. Since $\mathcal{S} \cap \mathcal{E}^c = \emptyset$, we have $\theta^*_{\mathcal{E}^c} = \mathbf{0}$. First, we decompose the term $\mathcal{I}_3$ into two subsets. In the subset $\mathcal{E}$,

$$-\hat{z}^T_\mathcal{E}(\hat{\theta} - \theta^*)_\mathcal{E} \leq \|\hat{z}^T_\mathcal{E}\|_\infty \|(\hat{\theta} - \theta^*)_\mathcal{E}\|_1 \leq \|(\hat{\theta} - \theta^*)_\mathcal{E}\|_1.$$

In the complement set $\mathcal{E}^c$,

$$\begin{aligned} \hat{z}^T_{\mathcal{E}^c}(\hat{\theta} - \theta^*)_{\mathcal{E}^c} = \hat{z}^T_{\mathcal{E}^c}\hat{\theta}_{\mathcal{E}^c} & \overset{(i)}{=} \sum_{\substack{\hat{\theta}^{(p)} \neq \mathbf{0}; \\ p \subseteq \mathcal{E}^c}} \frac{\|\hat{\theta}^{(p)}\|_2^2}{\|\hat{\theta}^{(p)}\|_2} + \sum_{\substack{\hat{\theta}^{(p)} = \mathbf{0}; \\ p \subseteq \mathcal{E}^c}} (\hat{z}^{(p)})^T\hat{\theta}^{(p)} \\ & \overset{(ii)}{\geq} \sum_{\substack{\hat{\theta}^{(p)} \neq \mathbf{0}; \\ p \subseteq \mathcal{E}^c}} \frac{1}{\sqrt{K}}\|\hat{\theta}^{(p)}\|_1 + \sum_{\substack{\hat{\theta}^{(p)} = \mathbf{0}; \\ p \subseteq \mathcal{E}^c}} 0 = \frac{1}{\sqrt{K}}\|(\hat{\theta} - \theta^*)_{\mathcal{E}^c}\|_1. \end{aligned}$$

In the step (i), We divide the estimator $\hat{\theta}_{\mathcal{E}^c}$ into nonzero and zero subsets. In step (ii), we apply the Cauchy–Schwarz inequality to obtain the result.

From the above derivations, the inequality 3.10 can be expanded as

$$(\lambda_n + \|\frac{1}{n}\nabla\mathcal{L}(\theta^*)\|_\infty)\|(\hat{\theta} - \theta^*)_\mathcal{E}\|_1 \geq (\lambda_n\frac{1}{\sqrt{K}} - \|\frac{1}{n}\nabla\mathcal{L}(\theta^*)\|_\infty)\|(\hat{\theta} - \theta^*)_{\mathcal{E}^c}\|_1. \tag{3.12}$$

Because $\|n^{-1}\nabla\mathcal{L}(\theta^*)\|_\infty \leq \lambda_n/(2\sqrt{K})$ with high probability, we have

$$\|(\hat{\theta}-\theta^*)_{\mathcal{E}^c}\|_1 \leq \frac{\lambda_n + \|n^{-1}\nabla\mathcal{L}(\theta^*)\|_\infty}{\lambda_n/\sqrt{K} - \|n^{-1}\nabla\mathcal{L}(\theta^*)\|_\infty}\|(\hat{\theta}-\theta^*)_{\mathcal{E}}\|_1.$$

In addition, if we plug in the maximum value $\lambda_n/(2\sqrt{K})$ of $\|n^{-1}\nabla\mathcal{L}(\theta^*)\|_\infty$, we obtain

$$\|(\hat{\theta}-\theta^*)_{\mathcal{E}^c}\|_1 \leq (2\sqrt{K}+1)\|(\hat{\theta}-\theta^*)_{\mathcal{E}}\|_1.$$

□

For the special case of $K = 1$, this inequality coincides with the conventional result on the LASSO estimator. Inequality (3.8) is essential to set a bound for the overall estimation error of the penalized estimate.

Theorem 3.1. *Based on Assumptions 3.1 - 3.6, suppose the composite score vector satisfies* $\|n^{-1}\nabla\mathcal{L}(\theta^*)\|_\infty \leq \lambda_n/(2\sqrt{K})$ *with the penalty parameter chosen as (3.7), there exists a penalized estimator* $\hat{\theta}$ *of (3.5), and*

$$\begin{aligned}
\|\hat{\theta}-\theta^*\|_2 &\leq \frac{3\lambda_n\sqrt{s}}{2\kappa_-};\\
\|\hat{\theta}-\theta^*\|_1 &\leq \frac{3\sqrt{K}(\sqrt{K}+1)}{\kappa_-}\lambda_n s;\\
(\frac{1}{n}\nabla\mathcal{L}(\hat{\theta}) - \frac{1}{n}\nabla\mathcal{L}(\theta^*))^T(\hat{\theta}-\theta^*) &\leq \frac{3(\sqrt{K}+1)(2\sqrt{K}+1)}{2\kappa_-}\lambda_n^2 s
\end{aligned}$$

with a probability at least $1 - 2\exp\{-C\log(p_n)\}$ *for some constant* C.

Proof. Lemma 3.4 shows that for the solution $\hat{\theta}$ of the estimating equation (3.5), $(\hat{\theta}-\theta^*) \in \mathcal{C}(m,\gamma)$ with $m = c_0 K s$ and $\gamma = 2\sqrt{K}+1$. Using the results from Lemma 3.2, we can obtain

the following inequality with probability tending to 1,

$$\frac{1}{n}\nabla Q(\hat{\theta})^T(\hat{\theta}-\theta^*) - (\frac{1}{n}\nabla\mathcal{L}(\theta^*) + \lambda_n\hat{z})^T(\hat{\theta}-\theta^*) \geq \kappa_-\|\hat{\theta}-\theta^*\|_2^2. \tag{3.13}$$

We decompose (3.13) into two components and apply Hölder's inequality:

$$\begin{aligned}
-\nabla\mathcal{L}(\theta^*)(\hat{\theta}-\theta^*) &\leq \|\nabla\mathcal{L}(\theta^*)_{\mathcal{E}}\|_2\|(\hat{\theta}-\theta^*)_{\mathcal{E}}\|_2 + \|\nabla\mathcal{L}(\theta^*)_{\mathcal{E}^c}\|_\infty\|(\hat{\theta}-\theta^*)_{\mathcal{E}^c}\|_1; \\
-\hat{z}^T(\hat{\theta}-\theta^*) &\leq \|\hat{z}_{\mathcal{E}}\|_2\|(\hat{\theta}-\theta^*)_{\mathcal{E}}\|_2 - \frac{1}{\sqrt{K}}\|(\hat{\theta}-\theta^*)_{\mathcal{E}^c}\|_1.
\end{aligned}$$

Plugging back into (3.13), we have

$$\begin{aligned}
\kappa_-\|\hat{\theta}-\theta^*\|_2^2 \leq \|\frac{1}{n}\nabla\mathcal{L}(\theta^*)_{\mathcal{E}}\|_2\|(\hat{\theta}-\theta^*)_{\mathcal{E}}\|_2 + \|\frac{1}{n}\nabla\mathcal{L}(\theta^*)_{\mathcal{E}^c}\|_\infty\|(\hat{\theta}-\theta^*)_{\mathcal{E}^c}\|_1 \\
+ \lambda_n\|\hat{z}_{\mathcal{E}}\|_2\|(\hat{\theta}-\theta^*)_{\mathcal{E}}\|_2 - \lambda_n\frac{1}{\sqrt{K}}\|(\hat{\theta}-\theta^*)_{\mathcal{E}^c}\|_1.
\end{aligned}$$

According to Lemma 3.1, $\|n^{-1}\nabla\mathcal{L}(\theta^*)\|_\infty \leq \lambda_n/2\sqrt{K}$ with a probability tending to 1. Therefore, the component $(\|\frac{1}{n}\nabla\mathcal{L}(\theta^*)_{\mathcal{E}^c}\|_\infty - \lambda_n/\sqrt{K})\|(\hat{\theta}-\theta^*)_{\mathcal{E}^c}\|_1 \leq 0$. We simplify the inequality above as follows,

$$\kappa_-\|\hat{\theta}-\theta^*\|_2^2 \leq (\|\frac{1}{n}\nabla\mathcal{L}(\theta^*)_{\mathcal{E}}\|_2 + \lambda_n\|\hat{z}_{\mathcal{E}}\|_2)\|(\hat{\theta}-\theta^*)_{\mathcal{E}}\|_2.$$

Based on the Cauchy–Schwarz inequality, $\|\nabla\mathcal{L}(\theta^*)_{\mathcal{E}}\|_2 \leq \sqrt{K|\mathcal{E}|}\|\nabla\mathcal{L}(\theta^*)_{\mathcal{E}}\|_\infty$. According to the property of mixed $\ell_{2,1}$ norm, $\|\hat{z}_{\mathcal{E}}\|_2 = \sqrt{|\mathcal{E}|}$. Because of the assumption that $|\mathcal{E}| = c_1|\mathcal{S}| = c_1 s$ with some constant $c_1 \geq 1$, the following inequality can be obtained

$$\begin{aligned}
\kappa_-\|\hat{\theta}-\theta^*\|_2^2 &\leq (\lambda_n\sqrt{c_1 s} + \sqrt{K|\mathcal{E}|}\|\frac{1}{n}\nabla\mathcal{L}(\theta^*)_{\mathcal{E}}\|_\infty)\|(\hat{\theta}-\theta^*)_{\mathcal{E}}\|_2, \\
&\leq (\lambda_n\sqrt{c_1 s} + \frac{\lambda_n}{2\sqrt{K}}\sqrt{c_1 K s})\|(\hat{\theta}-\theta^*)_{\mathcal{E}}\|_2 \leq \frac{3\lambda_n\sqrt{c_1 s}}{2}\|(\hat{\theta}-\theta^*)\|_2.
\end{aligned}$$

Therefore, taking the constant $c_1 = 1$, we have

$$\|\hat{\theta} - \theta^*\|_2 \leq \frac{3\lambda_n\sqrt{s}}{2\kappa_-} = O_p(\lambda_n\sqrt{s}).$$

In addition, we derive the following error bounds:

$$\begin{aligned}
\|\hat{\theta} - \theta^*\|_1 \leq & (2\sqrt{K}+2)\|(\hat{\theta} - \theta^*)_{\mathcal{E}}\|_1 \\
\leq & (2\sqrt{K}+2)\sqrt{sK}\|(\hat{\theta} - \theta^*)\|_2 \\
\leq & \frac{3\sqrt{K}(\sqrt{K}+1)}{\kappa_-}\lambda_n s; \\
\frac{1}{n}(\nabla\mathcal{L}(\hat{\theta}) - \nabla\mathcal{L}(\theta^*))^T(\hat{\theta} - \theta^*) \leq & \frac{1}{n}\|\nabla\mathcal{L}(\hat{\theta}) - \nabla\mathcal{L}(\theta^*)\|_\infty\|\hat{\theta} - \theta^*\|_1 \\
\leq & (\|\frac{1}{n}\nabla\mathcal{L}(\hat{\theta}) + \lambda_n\hat{z}\|_\infty + \|\frac{1}{n}\nabla\mathcal{L}(\theta^*)\|_\infty \\
& + \|\lambda_n\hat{z}\|_\infty)\|\hat{\theta} - \theta^*\|_1 \\
= & (\|\frac{1}{n}\nabla\mathcal{L}(\theta^*)\|_\infty + \lambda_n)\|\hat{\theta} - \theta^*\|_1 \\
\leq & \frac{3(\sqrt{K}+1)(2\sqrt{K}+1)}{2\kappa_-}\lambda_n^2 s.
\end{aligned}$$

□

As the size of the penalty parameter is $\mathcal{O}(\sqrt{\log(p_n)/n})$, Theorem 3.1 implies $\|\hat{\theta} - \theta^*\|_2 = O_p(\sqrt{\frac{s\log(p_n)}{n}})$.

Theorem 3.2. *Under Assumptions 3.1 - 3.7, suppose the penalty parameter λ_n is chosen as (3.7) and the minimum non-zero parameter*

$$\min_{k;p\in\mathcal{S}} |\theta_{kp}| \geq \frac{(1+2\sqrt{K})\sqrt{s}}{\kappa_-\sqrt{K}}\lambda_n,$$

then there exists a penalized estimator $\hat{\theta}$ of program (3.5) satisfies $sign(\hat{\theta}) = sign(\theta^*)$ with a probability at least $1 - 2p_n^{-d} - 4K\exp\{-C_0 n/s^3 + \log(p_n)\}$ for the universal constants $d > 1$

and $C_0 > 0$.

Proof. The derivative equation (3.5) can be partitioned into two sets of equations based on the two subspaces of parameters $\mathcal{S}$ and $\mathcal{S}^c$:

$$-\frac{1}{n}\nabla\mathcal{L}(\hat{\theta})_{\mathcal{S}} = \lambda_n \hat{z}_{\mathcal{S}}, \tag{3.14a}$$

$$-\frac{1}{n}\nabla\mathcal{L}(\hat{\theta})_{\mathcal{S}^c} = \lambda_n \hat{z}_{\mathcal{S}^c}. \tag{3.14b}$$

Based on the definition of sub-differential, the sub-differential $\hat{z}_{\mathcal{S}}$ contains grouped subsets $\hat{z}^{(p)} = \hat{\theta}^{(p)}/\|\hat{\theta}^{(p)}\|_2$ with $p \in \mathcal{S}$, and $\max_{p\in\mathcal{S}^c} \|\hat{z}^{(p)}\|_2 < 1$.

According to Lemma 3.2, $\hat{\theta}$ is the optima of the objective function with high probability. Consider an estimator with $\hat{\theta}_{\mathcal{S},0} = (\hat{\theta}_{\mathcal{S}}, \mathbf{0})$, where

$$\hat{\theta}_{\mathcal{S},0} = \underset{\theta=(\theta_{\mathcal{S}},\mathbf{0})}{\arg\min}\{\mathcal{L}(\theta) + n\lambda_n\|\theta\|_{2,1}\}.$$

If the estimator $\hat{\theta}_{\mathcal{S},0}$ satisfies the conditions (3.14a) and (3.14b), then with high probability, $\hat{\theta}_{\mathcal{S},0}$ is the local optimal solution $\hat{\theta}$ to Equation (3.5).

We expand the score function Using the mean value theorem as follows

$$\begin{aligned}\frac{1}{n}\nabla\mathcal{L}(\hat{\theta}_{\mathcal{S},0}) &= \frac{1}{n}\nabla\mathcal{L}(\theta^*) + \frac{1}{n}\nabla\mathcal{L}(\hat{\theta}_{\mathcal{S},0}) - \frac{1}{n}\nabla\mathcal{L}(\theta^*) = \frac{1}{n}\nabla\mathcal{L}(\theta^*) + \frac{1}{n}\nabla^2\mathcal{L}(\tilde{\theta})\hat{\Delta} \\ &= \frac{1}{n}\nabla\mathcal{L}(\theta^*) + \frac{1}{n}\nabla^2\mathcal{L}(\theta^*)\hat{\Delta} + \underbrace{(\frac{1}{n}\nabla^2\mathcal{L}(\tilde{\theta}) - \frac{1}{n}\nabla^2\mathcal{L}(\theta^*))\hat{\Delta}}_{\mathcal{R}},\end{aligned}$$

where $\hat{\Delta} = (\hat{\theta}_{\mathcal{S},0} - \theta^*)$, $\tilde{\theta} = \alpha\theta^* + (1-\alpha)\hat{\theta}_{\mathcal{S}}$ for some $\alpha \in [0,1]$.

Thus, we write the equations (3.14a) and (3.14b) in block format with solution $\hat{\theta}_{\mathcal{S},0}$

$$\frac{1}{n}\begin{bmatrix}\nabla^2\mathcal{L}(\theta^*)_{\mathcal{S}\mathcal{S}} & \nabla^2\mathcal{L}(\theta^*)_{\mathcal{S}\mathcal{S}^c} \\ \nabla^2\mathcal{L}(\theta^*)_{\mathcal{S}^c\mathcal{S}} & \nabla^2\mathcal{L}(\theta^*)_{\mathcal{S}^c\mathcal{S}^c}\end{bmatrix}\begin{pmatrix}\hat{\Delta}_{\mathcal{S}} \\ \mathbf{0}\end{pmatrix} + \frac{1}{n}\begin{pmatrix}\nabla\mathcal{L}(\theta^*)_{\mathcal{S}} \\ \nabla\mathcal{L}(\theta^*)_{\mathcal{S}^c}\end{pmatrix} + \begin{pmatrix}\mathcal{R}_{\mathcal{S}} + \lambda_n\hat{z}_{\mathcal{S}} \\ \mathcal{R}_{\mathcal{S}^c} + \lambda_n\hat{z}_{\mathcal{S}^c}\end{pmatrix} = \mathbf{0}.$$

According to Lemma 3.8, the sub-matrix $n^{-1}\nabla^2\mathcal{L}(\theta^*)_{\mathcal{SS}}$ is invertible with high probability. Thus, we obtain the difference block $\Delta_{\mathcal{S}}$ by solving

$$\Delta_{\mathcal{S}} = \hat{\theta}_{\mathcal{S}} - \theta^*_{\mathcal{S}} = -(\frac{1}{n}\nabla^2\mathcal{L}(\theta^*)_{\mathcal{SS}})^{-1}(\frac{1}{n}\nabla\mathcal{L}(\theta^*)_{\mathcal{S}} + \lambda_n\hat{z}_{\mathcal{S}} + \mathcal{R}_{\mathcal{S}}).$$

Next, we show that the elements of the remainder vector $\mathcal{R}$ can be expanded as follow

$$\mathcal{R}_{kp} = (\frac{\partial^2\ell_k(\tilde{\theta}; Y_k)}{\partial\theta\partial\theta^T\partial\theta_{kp}} - \frac{\partial^2\ell_k(\theta^*; Y_k)}{\partial\theta\partial\theta^T\partial\theta_{kp}})\hat{\Delta} = \tilde{\Delta}^T(\frac{\partial^3\ell_k(\theta^*; Y_k)}{\partial\theta\partial\theta^T\partial\theta_{kp}})\hat{\Delta},$$

with $\tilde{\Delta} = (\tilde{\theta} - \theta^*) = (1-\alpha)\hat{\Delta}$. Let $\nabla_{kp}\mathcal{H}^* = \partial^3\ell_k(\theta^*; Y_k)/\partial\theta\partial\theta^T\partial\theta_{kp}$, where $\nabla_{kp}\mathcal{H}^*$ is a $Kp_n \times Kp_n$ matrix. With similar derivation as in the proof of Lemma 3.6, all elements of $\nabla_{kp}\mathcal{H}^*$ are from sub-exponential distributions. Thus, we show that for any $k = 1, 2, \cdots, K$ and $p = 1, 2, \cdots, p_n$,

$$\begin{aligned}\mathcal{R}_{kp} &= (1-\alpha)\hat{\Delta}^T_{\mathcal{S}}\nabla_{kp}\mathcal{H}^*_{\mathcal{SS}}\hat{\Delta}_{\mathcal{S}} \leq (1-\alpha)\|\hat{\Delta}_{\mathcal{S}}\|_2^2|||\nabla_{kp}\mathcal{H}^*_{\mathcal{SS}}|||_2 \\ &\overset{(i)}{\leq} \mathcal{W}^*\|\hat{\Delta}_{\mathcal{S}}\|_2^2 \overset{(ii)}{\leq} \frac{9(\mathcal{W}^* + \delta)}{4\kappa_-^2}\lambda_n^2 s.\end{aligned}$$

The step (i) is obtained based on the sub-exponential condition for the elements of $\nabla_{kp}\mathcal{H}^*$. For some small δ and a universal constant C,

$$P(|||\nabla_{kp}\mathcal{H}^*_{\mathcal{SS}}|||_2 \geq |||E(\nabla_{kp}\mathcal{H}^*_{\mathcal{SS}})|||_2 + \delta) \leq 2K\exp\{-C\frac{\delta^2 n}{Ks^2} + 2\log(s)\}.$$

According to Assumption 3.2, $\mathcal{W}^* \geq |||E(\nabla_{kp}\mathcal{H}^*_{\mathcal{SS}})|||_2$. Thus, $|||\nabla_{kp}\mathcal{H}^*_{\mathcal{SS}}|||_2 \leq (W^* + \delta)$ with high probability. According to Theorem 3.1, $\|\hat{\Delta}\|_2^2 \leq 9\lambda_n^2 s/(2\kappa_-)^2$. This leads to the result in step (ii).

Combining the results above, we show that with a probability larger than $1 - 2p_n^{-d} -$

$4K\exp\{-C_0\frac{n}{s^3}+\log(p_n)\}$,

$$\begin{aligned}\|\Delta_{\mathcal{S}}\|_\infty &= \|(\frac{1}{n}\nabla^2\mathcal{L}(\theta^*)_{\mathcal{S}\mathcal{S}})^{-1}(\frac{1}{n}\nabla\mathcal{L}(\theta^*)_{\mathcal{S}}+\lambda_n\hat{z}_{\mathcal{S}}+\mathcal{R}_{\mathcal{S}})\|_\infty \\ &\leq \frac{\sqrt{s}}{\kappa_-}\big(\|\frac{1}{n}\nabla\mathcal{L}(\theta^*)_{\mathcal{S}}\|_\infty+\lambda_n\|\hat{z}_{\mathcal{S}}\|_\infty+\|\mathcal{R}_{\mathcal{S}}\|_\infty\big) \leq \frac{(1+2\sqrt{K})\sqrt{s}}{\kappa_-\sqrt{K}}\lambda_n \leq |\theta^*_{\min}|,\end{aligned}$$

for some constant $C_0>0$. This implies $\text{sign}(\hat{\theta}_S)=\text{sign}(\theta^*_S)$.

Next, we show that $\max_{p\subset\mathcal{S}^c}\|\hat{z}^{(p)}\|_2<1$, which satisfies the KKT conditions. The sub-differential $\hat{z}_{\mathcal{S}^c}$ can be calculated from the block equation above,

$$\begin{aligned}\hat{z}_{\mathcal{S}^c} = -\frac{1}{\lambda_n}(\frac{1}{n}\nabla\mathcal{L}(\theta^*)_{\mathcal{S}^c}+\mathcal{R}_{\mathcal{S}^c}-\frac{1}{n}\nabla^2\mathcal{L}(\theta^*)_{\mathcal{S}^c\mathcal{S}}(\frac{1}{n}\nabla^2\mathcal{L}(\theta^*)_{\mathcal{S}\mathcal{S}})^{-1} \\ (\frac{1}{n}\nabla\mathcal{L}(\theta^*)_{\mathcal{S}}+\lambda_n\hat{z}_{\mathcal{S}}+\mathcal{R}_{\mathcal{S}})).\end{aligned} \tag{3.15}$$

The sub-differential $z_{\mathcal{S}^c}$ from (3.15) can be decomposed into three components

$$\begin{aligned}\hat{z}_{\mathcal{S}^c} = \frac{1}{\lambda_n}(&\underbrace{\frac{1}{n}\nabla^2\mathcal{L}(\theta^*)_{\mathcal{S}^c\mathcal{S}}(\frac{1}{n}\nabla^2\mathcal{L}(\theta^*)_{\mathcal{S}\mathcal{S}})^{-1}\frac{1}{n}\nabla\mathcal{L}(\theta^*)_{\mathcal{S}}-\frac{1}{n}\nabla\mathcal{L}(\theta^*)_{\mathcal{S}^c}}_{\mathcal{I}_1} \\ &+\underbrace{\frac{1}{n}\nabla^2\mathcal{L}(\theta^*)_{\mathcal{S}^c\mathcal{S}}(\frac{1}{n}\nabla^2\mathcal{L}(\theta^*)_{\mathcal{S}\mathcal{S}})^{-1}\mathcal{R}_{\mathcal{S}}-\mathcal{R}_{\mathcal{S}^c}}_{\mathcal{I}_2} \\ &+\underbrace{\lambda_n\frac{1}{n}\nabla^2\mathcal{L}(\theta^*)_{\mathcal{S}^c\mathcal{S}}(\frac{1}{n}\nabla^2\mathcal{L}(\theta^*)_{\mathcal{S}\mathcal{S}})^{-1}\hat{z}_{\mathcal{S}}}_{\mathcal{I}_3}).\end{aligned} \tag{3.16}$$

The sub-differential can be grouped as $\hat{z}^{(p)}$ with $p\subset\mathcal{S}^c$.

Based on Lemma 3.3 and Corollary 3.1, the following upper bound can be obtained with a probability at least $1-2\exp\{-d\log(p_n)\}$ for some constant $d>1$,

$$\begin{aligned}\max_{p\subset\mathcal{S}^c}\big\|\mathcal{I}_1^{(p)}\|_2 \leq &\max_{p\subset\mathcal{S}^c}\|\frac{1}{n}\nabla\mathcal{L}(\theta^*)^{(p)}\big\|_2 \\ &+\max_{p\subset\mathcal{S}^c}\big\|\big\{\frac{1}{n}\nabla^2\mathcal{L}(\theta^*)_{\mathcal{S}^c\mathcal{S}}(\frac{1}{n}\nabla^2\mathcal{L}(\theta^*)_{\mathcal{S}\mathcal{S}})^{-1}\frac{1}{n}\nabla\mathcal{L}(\theta^*)_{\mathcal{S}}\big\}^{(p)}\big\|_2\end{aligned}$$

$$
\begin{aligned}
&\leq \max_{p\subset\mathcal{S}^c} \|\frac{1}{n}\nabla\mathcal{L}(\theta^*)^{(p)}\|_2 \\
&+\sqrt{K}\Big|\Big|\Big|\frac{1}{n}\nabla^2\mathcal{L}(\theta^*)_{\mathcal{S}^c\mathcal{S}}(\frac{1}{n}\nabla^2\mathcal{L}(\theta^*)_{\mathcal{S}\mathcal{S}})^{-1}\Big|\Big|\Big|_\infty \|\frac{1}{n}\nabla\mathcal{L}(\theta^*)_{\mathcal{S}}\|_\infty \\
&\leq \frac{\xi}{4}\lambda_n + \frac{\xi}{4}(1-\frac{\xi}{2})\lambda_n < \frac{\xi}{2}\lambda_n.
\end{aligned}
$$

For the remainder component, we have

$$
\begin{aligned}
\max_{p\subset\mathcal{S}^c}\|\mathcal{I}_2^{(p)}\|_2 \leq& \sqrt{K}\|\mathcal{I}_2\|_\infty \\
\leq& \sqrt{K}(\|\mathcal{R}_{\mathcal{S}^c}\|_\infty + \|\mathcal{R}_{\mathcal{S}}\|_\infty\Big|\Big|\Big|\frac{1}{n}\nabla^2\mathcal{L}(\theta^*)_{\mathcal{S}^c\mathcal{S}}(\frac{1}{n}\nabla^2\mathcal{L}(\theta^*)_{\mathcal{S}\mathcal{S}})^{-1}\Big|\Big|\Big|_\infty) \\
\leq& \frac{9\mathcal{W}^*}{4\kappa_-^2}\lambda_n^2 s\sqrt{K} = \mathcal{O}(\frac{s\log(p_n)}{n}) \to o(1).
\end{aligned}
$$

Similarly, we show that the mixed norm of $\mathcal{I}_3$ can be bounded,

$$
\max_{p\subset\mathcal{S}^c}\|\mathcal{I}_3^{(p)}\|_2 = \max_{p\subset\mathcal{S}^c}\lambda_n\big\|\big\{\frac{1}{n}\nabla^2\mathcal{L}(\theta^*)_{\mathcal{S}^c\mathcal{S}}(\frac{1}{n}\nabla^2\mathcal{L}(\theta^*)_{\mathcal{S}\mathcal{S}})^{-1}\hat{z}_{\mathcal{S}}\big\}^{(p)}\big\|_2 \leq \lambda_n(1-\frac{1}{2}\xi).
$$

By adding the three components, we show that

$$
\max_{p\subset\mathcal{S}^c}\|\hat{z}^{(p)}\|_2 \leq \max_{p\subset\mathcal{S}^c}\frac{1}{\lambda_n}(\|\mathcal{I}_1^{(p)}\|_2 + \|\mathcal{I}_2^{(p)}\|_2 + \|\mathcal{I}_3^{(p)}\|_2) < 1 - \frac{\xi}{2} + \frac{\xi}{2} < 1.
$$

Combining the results above, we have $\text{sign}(\hat{\theta}) = \text{sign}(\theta^*)$ with a probability tending to 1. □

Theorem 3.2 indicates that the proposed penalized estimate for multi-task learning achieves sign consistency and recovers the union support across multiple tasks. To measure the performance of selected features for all tasks, we can apply the pseudo-Bayesian information criterion proposed by Gao and Carroll [2017]. This information criterion evaluates the joint model complexity, which also considers the correlation between different related tasks.

3.3 Optimization

From the computational perspective, an iterative algorithm can be developed to solve the optimization problem in (3.5). We set multiple steps of optimization labeled as step t, $t = 1, \ldots, \mathcal{T}$. The parameters updated at the jth iteration of the tth step are denoted by $\theta^{[t,j]}$ with $\theta^{(p),[t,j]}$ representing the pth grouped parameters. In each step, we apply the composite gradient descent (CGD) algorithm [Nesterov, 2013] to update subsequent iterations, which is widely used in high dimensional data analysis [Agarwal et al., 2012a, Loh and Wainwright, 2015]. For simplicity, we use θ^j to denote the update $\theta^{[t,j]}$ and $\theta^{(p),j}$ to denote the grouped parameters $\theta^{(p)[t,j]}$ in step t. To approximate the objective function $Q(\theta)$, we apply the majorize-minimization (MM) method by introducing an isotropic quadratic function

$$Q(\theta|\theta^j) = \frac{1}{n}\mathcal{L}(\theta^j) + \frac{1}{n}\nabla\mathcal{L}(\theta^j)(\theta - \theta^j) + \frac{\gamma_t}{2}\|\theta - \theta^j\|_2^2 + \lambda_n\|\theta\|_{2,1},$$

where the conventional value of the quadratic coefficient γ_t is chosen as the largest eigenvalue of the Hessian matrix. Based on the property of the majorize-minimization (MM) algorithm, we have

$$Q(\theta^{j+1}) \leq Q(\theta^{j+1}|\theta^j) \text{ and } Q(\theta^j|\theta^j) = Q(\theta^j).$$

Therefore, we solve each subsequent optimization sub-problems by minimizing the function $Q(\theta|\theta^j)$

$$\theta^{j+1} = \arg\min_\theta\{Q(\theta|\theta^j)\} = \arg\min_\theta\{\frac{1}{n}\nabla\mathcal{L}(\theta^j)\theta + \frac{\gamma_t}{2}\|\theta - \theta^j\|_2^2 + \lambda_n\|\theta\|_{2,1}\}.$$

In addition, we simplify the solution of the optimization problem by completing the square as

$$\theta^{j+1} = \arg\min_{\theta}\{\frac{1}{2}\|\theta - (\theta^j - \frac{\frac{1}{n}\nabla\mathcal{L}(\theta^j)}{\gamma_t})\|_2^2 + \frac{\lambda_n}{\gamma_t}\|\theta\|_{2,1}\}.$$

We define the thresholding operator $S_{\lambda_n/\eta}$ on grouped parameters as

$$S_{\lambda_n/\eta}(\theta^{(p)}) = (\|\theta^{(p)}\|_2 - \frac{\lambda_n}{\eta})_+ z^{(p)},$$

where the subdifferential is defined as $z^{(p)} = \partial\|\theta^{(p)}\|_2/\partial\theta^{(p)}$. The subsequent update from the jth to the $j+1$th step can be obtained as

$$\theta^{(p),j+1} = S_{\lambda_n/\eta}(\theta^{(p),j} - \frac{1}{n}\frac{\nabla\mathcal{L}(\theta^j)^{(p)}}{\eta}), \tag{3.17}$$

where the value of the step size η can be set equal or proportional to the coefficient γ_t, which can be updated in each step. Within each step, we apply the fast iterative shrinkage thresholding algorithm (FISTA) framework [Beck and Teboulle, 2009] to further accelerate the inner loop updates. Under this setting, we can show that the iterations within each step enjoy a geometric rate of convergence.

3.4 Simulation

In this section, we present the simulation studies to show the empirical performance of the proposed multi-task feature learning algorithm. In section 3.4.1, the joint feature selection is examined for correlated tasks under different scenarios. In section 3.4.2, we evaluate the prediction errors with increasing dimensions of parameters and sample size.

Algorithm 1: The iterative algorithm for the multi-task feature learning

Data: K different platforms of data sets $\{Y_k, X_{k1}, X_{k2}, \cdots, X_{kp_n}\}_{k=1}^{K}$;
Input: the initial parameter $\{\theta^{[1,1]}, \lambda_n, \epsilon_c\}$;
Output: the optimal estimator $\hat{\theta}$.
while $\|\theta^{[t]} - \theta^{[t-1]}\|_2 > \epsilon_c$ **do**
  $\alpha_1 = 1$
  $u^1 = \theta^{[t]}$
  $\eta_t = \max \Lambda(\frac{1}{n}\nabla^2 \mathcal{L}(\theta^{[t]}))$
  while *convergence is not reached* **do**
    $h^j = u^j - \frac{1}{n}\frac{\nabla \mathcal{L}(u^j)}{\eta_t}$
    $\theta^{[t,j]} = S_{\lambda_n/\eta_t}(h^j - \frac{1}{n}\frac{\nabla \mathcal{L}(h^j)}{\eta_t})$
    $\alpha_{j+1} = \frac{1}{2}(1 + \sqrt{1 + 4\alpha_j^2})$
    $u^{j+1} = \theta^{[t,j]} + \frac{\alpha_j - 1}{\alpha_{j+1}}(\theta^{[t,j]} - \theta^{[t,j-1]})$
  end
end

Table 3.2: Coefficients generating process in four tasks with different types of effects

	Coefficient Type	Sampling distribution
Task 1	Large variance	$\theta_1^* \sim N(1, 3)$
Task 2	Small variance	$\theta_2^* \sim N(1, 1)$
Task 3	Strictly positive	$\theta_3^* \sim$Uniform$(1, 2)$
Task 4	No sign constraint	$\theta_4^* \sim$Uniform$(-1, 1)$

3.4.1 Joint Feature Selection

We simulate four different tasks and each task has 200 or 500 observations with $200, 500$, and 1000 predictors. The covariates are generated from a multivariate normal distribution with means zero and variances one. The number of non-zero coefficients is equal to $s = \lfloor p_n^{1/2} \rfloor$. Since the heterogeneous data sets may possess different relationships between the response variables and the predictors, we generate four different types of coefficients for different tasks shown in Table 3.2.

Case 1: Correlated Multivariate Regressions

We consider the case where multiple tasks have correlated responses. We generate the response variables for four different tasks from linear models $y_{ki} = \eta_{ki} + \varepsilon_{ki}$, where the error terms from different tasks are correlated. The error terms $(\varepsilon_{1i}, \ldots, \varepsilon_{4i})^T$ follow a multivariate Gaussian distribution or a heaty-tailed multivariate t distribution with 10 degrees of freedom with all means equal to zero, and a 4×4 covariance matrix Σ is given by

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \rho_{13}\sigma_1\sigma_3 & \rho_{14}\sigma_1\sigma_4 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & \rho_{23}\sigma_2\sigma_3 & \rho_{24}\sigma_2\sigma_4 \\ \rho_{13}\sigma_1\sigma_3 & \rho_{23}\sigma_2\sigma_3 & \sigma_3^2 & \rho_{34}\sigma_3\sigma_4 \\ \rho_{14}\sigma_1\sigma_4 & \rho_{24}\sigma_2\sigma_4 & \rho_{34}\sigma_3\sigma_4 & \sigma_4^2 \end{bmatrix}.$$

The variances of error terms in different platforms are designed as $\sigma_1^2 = 9, \sigma_2^2 = 4, \sigma_3^2 = 4$, and $\sigma_4^2 = 1$. The correlation ρ_{kl}'s are generated from a uniform distribution Unif$(0.4, 0.6)$ for any $k, l = 1, 2, 3, 4$ and $k \neq l$.

The simulation results in Table 3.3 contain the specificity and sensitivity of feature selection obtained from 500 independent replications. In each case, we set the penalty parameter $\lambda_n = 5\sqrt{\log(p_n)/n}$ based on the asymptotic rate from theoretical results. The results in Table 3.3 show that for Gaussian errors and heavy-tailed errors, the proposed feature learning achieves high specificity and low sensitivity.

Case 2: Mixture of Regression and Classification Tasks

In the Case 2 simulation study, we combine two regression tasks and two classification tasks. For continuous response variables, the generating process is identical to the first two tasks in Case 1. The categorical responses are obtained by dichotomizing the continuous responses

Table 3.3: Specificity (SPE %) and sensitivity (SEN %) of the multi-task feature learning (MTL) compared with single-task analysis (SA) for multivariate linear models. The standard errors (%) are provided in parenthesis.

	$n = 200$ $p_n = 200$		$n = 200$ $p_n = 500$		$n = 500$ $p_n = 500$		$n = 500$ $p_n = 1000$	
Model	SPE	SEN	SPE	SEN	SPE	SEN	SPE	SEN
Simulation I: Gaussian Error								
MTL	98 (1)	100 (0)	99 (0)	100 (1)	100 (0)	100 (0)	100 (0)	100 (0)
SA 1	81 (8)	92 (8)	86 (4)	90 (7)	87 (4)	94 (5)	89 (3)	94 (4)
SA 2	83 (6)	89 (9)	87 (5)	86 (9)	87 (4)	96 (6)	90 (3)	92 (5)
SA 3	80 (7)	100 (0)	85 (4)	100 (0)	86 (5)	100 (0)	88 (3)	100 (0)
SA 4	83 (7)	88 (9)	87 (4)	86 (7)	88 (4)	92 (6)	89 (3)	92 (5)
Simulation II: Heavy-tailed Error								
MTL	97 (1)	100 (0)	99 (1)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)
SA 1	82 (8)	91 (8)	87 (5)	89 (7)	87 (4)	94 (5)	89 (3)	93 (5)
SA 2	83 (7)	88 (9)	87 (5)	85 (8)	87 (4)	93 (6)	89 (3)	91 (6)
SA 3	80 (8)	100 (0)	84 (6)	100 (0)	85 (4)	100 (0)	88 (3)	100 (0)
SA 4	82 (8)	88 (10)	87 (5)	84 (9)	88 (4)	91 (6)	89 (3)	91 (6)

[Poon and Lee, 1987a] as follows,

$$\begin{cases} y_{ki} = 1, \text{ if } \eta_{ki} + \varepsilon_{ki} \geq 0, \\ y_{ki} = 0, \text{ if } \eta_{ki} + \varepsilon_{ki} < 0. \end{cases}$$

The error terms ε_{ki} are jointly generated across all tasks from the multivariate distribution same as the simulation in Case 1.

The results in Table 3.4 show that the overall model selection performance of learning from heterogeneous types of tasks is enhanced in comparison with the single task analysis.

Case 3: Mixed Types of Tasks with High Correlation

The proposed multi-task learning algorithm is designed to deal with correlated tasks. We modify the simulation in Case 1 with high correlation errors. An exchangeable correlation

Table 3.4: Specificity (SPE %) and sensitivity (SEN %) of the multi-task feature learning (MTL) compared with single-task analysis (SA) for a mixture of regression and classification tasks. The standard errors (%) are provided in parenthesis.

	$n = 200$ $p_n = 200$		$n = 200$ $p_n = 500$		$n = 500$ $p_n = 500$		$n = 500$ $p_n = 1000$	
Model	SPE	SEN	SPE	SEN	SPE	SEN	SPE	SEN
Simulation I: Gaussian Error								
MTL	98 (1)	97 (4)	99 (1)	94 (5)	100 (0)	98 (3)	100 (0)	96 (3)
SA 1	81 (8)	92 (8)	86 (4)	90 (7)	87 (4)	94 (5)	89 (3)	94 (4)
SA 2	83 (6)	89 (9)	87 (5)	86 (9)	87 (4)	94 (6)	90 (3)	92 (5)
SA 3	79 (5)	98 (4)	90 (3)	82 (12)	80 (4)	100 (0)	87 (2)	99 (2)
SA 4	82 (6)	78 (9)	91 (3)	65 (9)	83 (4)	85 (8)	88 (3)	80 (7)
Simulation II: Heavy-tailed Error								
MTL	97 (1)	97 (4)	99 (1)	93 (5)	100 (0)	98 (3)	100 (0)	96 (3)
SA 1	82 (8)	91 (8)	87 (5)	89 (7)	87 (4)	94 (5)	89 (3)	93 (5)
SA 2	83 (7)	88 (9)	87 (4)	85 (8)	87 (4)	93 (6)	89 (3)	91 (6)
SA 3	81 (5)	98 (4)	91 (3)	80 (10)	81 (3)	100 (1)	87 (2)	98 (2)
SA 4	84 (6)	77 (11)	91 (3)	64 (9)	84 (4)	83 (8)	89 (2)	79 (7)

structure is used with $\rho_{kl} = 0.9$ for $k, l = 1, 2, 3, 4$ and $k \neq l$, and other settings remain the same as the simulation in Case 2. As demonstrated by the simulation result in Table 3.5, the proposed method provides higher Specificity and lower sensitivity than single-task analysis in the presence of high correlation.

Case 4: Tasks with Unbalanced Samples

In this simulation, we examine the performance of the algorithm by analyzing multiple tasks with unbalanced sample sizes across different tasks. The data-generating procedure follows the design in Case 2 with 500 replications, but we randomly delete 50% to 80% of observations in the last three tasks. Thus, the sample sizes are different across the tasks. The multi-task analysis provides high accuracy with unbalanced samples as shown in Table 3.6. For the single-task analysis, the sensitivity rates are much higher in the tasks with smaller sample sizes.

Table 3.5: Specificity (SPE %) and sensitivity (SEN %) of the multi-task feature learning (MTL) compared with single-task analysis (SA) for highly correlated tasks. The standard errors (%) are provided in parenthesis.

	$n = 200$ $p_n = 200$		$n = 200$ $p_n = 500$		$n = 500$ $p_n = 500$		$n = 500$ $p_n = 1000$	
Model	SPE	SEN	SPE	SEN	SPE	SEN	SPE	SEN
Simulation I: Gaussian Error								
MTL	99 (1)	96 (5)	97 (1)	96 (4)	98 (1)	99 (2)	99 (0)	98 (3)
SA 1	81 (8)	92 (8)	87 (4)	90 (7)	87 (4)	94 (5)	89 (3)	94 (4)
SA 2	82 (7)	89 (9)	87 (5)	86 (8)	87 (4)	92 (6)	89 (3)	92 (5)
SA 3	79 (5)	98 (4)	90 (3)	81 (10)	80 (3)	100 (0)	87 (2)	99 (2)
SA 4	83 (6)	77 (12)	91 (3)	64 (10)	83 (4)	85 (8)	88 (2)	79 (7)
Simulation II: Heavy-tail Error								
MTL	98 (1)	95 (5)	97 (1)	96 (4)	98 (1)	99 (2)	99 (0)	98 (3)
SA 1	81 (8)	92 (8)	87 (4)	90 (7)	87 (4)	94 (5)	89 (3)	94 (4)
SA 2	82 (7)	90 (8)	87 (5)	86 (8)	87 (4)	92 (6)	89 (3)	92 (5)
SA 3	80 (5)	98 (4)	90 (3)	81 (10)	80 (3)	100 (0)	87 (2)	99 (2)
SA 4	83 (6)	76 (11)	91 (3)	64 (10)	83 (4)	85 (8)	88 (2)	79 (7)

3.4.2 Estimation Consistency

In this section, we conduct simulations to investigate the statistical consistency property of the proposed penalized estimates. In four tasks, the number of predictors is designed with two levels, such that $p_n = 200$ and 1600. The covariates are simulated from the multivariate Gaussian distribution with means equal to zero and variances equal to one. For important features, the corresponding non-zero coefficients θ^* are generated from uniform distribution $\text{Unif}(0.05, 0.5)$ for all tasks. The true support s is chosen as $\lfloor p^{1/3} \rfloor$. We set sample size satisfying $n = \alpha s \log(p_n)$ for some constant α, where α ranges from 1 to 8. The response variables are generated through a similar process as Case 2. In the first two tasks, the response variables are generated based on the linear predictors $\eta_{ki} + \varepsilon_{ki}$. The response variables in the last two tasks are obtained through dichotomization. The error terms are jointly simulated from a multivariate t distribution with 10 degrees of freedom, which are moderately correlated.

Table 3.6: Specificity (SPE %) and sensitivity (SEN %) of the multi-task feature learning (MTL) compared with single-task analysis (SA) for multiple tasks with unbalanced sample sizes. The standard errors (%) are provided in parenthesis.

		$p_n = 500$		$p_n = 1000$		$p_n = 2000$	
Model	Sample	SPE	SEN	SPE	SEN	SPE	SEN
Simulation I: Gaussian Error							
MTL		99 (0)	99 (2)	100 (0)	97 (3)	99 (0)	96 (4)
SA 1	$n = 500$	87 (4)	97 (4)	89 (3)	98 (3)	91 (2)	92 (4)
SA 2	$n = 250$	87 (4)	88 (5)	89 (3)	85 (6)	92 (2)	77 (6)
SA 3	$n = 100$	97 (3)	27 (17)	99 (1)	10 (10)	100(1)	4 (5)
SA 4	$n = 100$	98 (2)	9 (14)	99 (1)	5 (6)	99 (1)	5 (5)
Simulation II: Heavy-tail Error							
MTL		97 (1)	98 (3)	98 (1)	97 (3)	99 (0)	95 (4)
SA 1	n = 500	87 (4)	93 (6)	89 (3)	93 (5)	91 (2)	91 (4)
SA 2	n = 250	87 (4)	87 (8)	90 (3)	84 (7)	93 (2)	75 (6)
SA 3	n = 100	97 (3)	26 (17)	99 (1)	10 (9)	100(1)	4 (4)
SA 4	n = 100	97 (3)	28 (15)	99 (1)	12 (10)	99 (1)	5 (5)

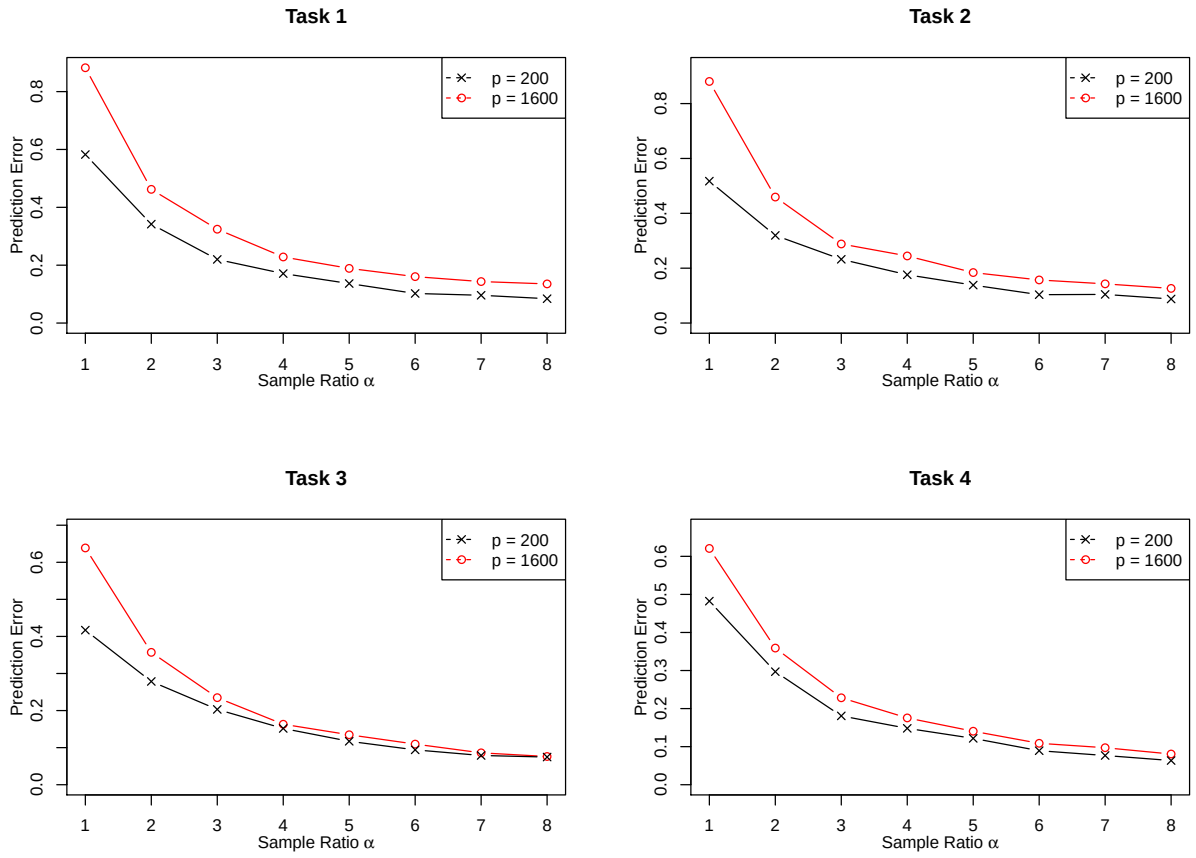
An unstructured correlation matrix is used for the simulation with $\rho_{kl} \sim \text{Unif}(0.4, 0.6)$ for any $k, l = 1, 2, 3, 4$ and $k \neq l$ as previous examples. We conduct 50 independent replications to measure the prediction errors evaluated as $n^{-1}(\nabla\mathcal{L}(\hat{\theta}) - \nabla\mathcal{L}(\theta^*))^T(\hat{\theta} - \theta^*)$ according to Theorem 3.1. The curves shown in Figure 3.1 demonstrate that the prediction error decreases toward zero as the sample size increases.

3.5 Data analysis

3.5.1 Breast Cancer Study

In the first example, we apply the multi-task feature learning on a collection of breast cancer studies from the NCBI (National Centre for Biotechnology Information) database [Hatzis et al., 2011, Itoh et al., 2014, Ivshina et al., 2006, Schmidt et al., 2008, Cadenas et al., 2014, Hellwig et al., 2010, Heimes et al., 2020, Rody et al., 2011, Karn et al., 2014, Gao and Zhong,

Figure 3.1: The prediction error of the multi-tasks feature learning estimator for four correlated tasks.



2019]. The Affymetrix microarray technique was used to provide the gene expression profiles of thousands of genes for the patients. The joint feature selection is applied to obtain the list of genetic biomarkers that are associated with breast cancer disease status. The microarray datasets include 22,283 biomarkers, and we apply the rank aggregation method to obtain 1300 candidate biomarkers for the joint feature selection. In the data sets, there are three levels of histologic grades representing the stages of cancer. However, in some studies, the first and second grades are classified together as the group of early-stage cancer, and the third grade is the group of high risk. Therefore, some studies have binary outcomes, while others contain

Table 3.7: Breast cancer multi-task studies. The performance of the logistic regression models is measured by AUC; the performance of the multinomial regression models is measured by the percentage of correct classification.

Tasks	Logistic regression (AUC)			Multinomial regression (% Classification)	
Data	GSE11121	GSE4922	GSE31519	GSE25055	GSE25066
(n)	(151)	(188)	(46)	(217)	(358)
MTL	0.81	0.81	0.77	66	66
SA	0.74	0.83	0.65	66	64

multinomial outcomes. We perform the five-fold cross-validation 50 times. The multi-task feature learning is applied to the training sets. The selection criterion in the training process is based on the area under the ROC curve (AUC) for binary responses and the percentage of accurate classification for multinomial responses. Based on the selected biomarkers, the logistic models and multinomial regressions are used for prediction in the testing sets. The model performance of AUC and the classification accuracy for the test sets are shown in Table 3.7. The overall classification accuracy is improved by the multi-task feature learning compared to the single-task learning. Especially when the studies have unequal sample sizes, the studies with smaller samples enjoy great improvement by learning from other studies with larger samples.

3.5.2 Community Health Status Research

In the second example, multiple community health status indicators (CHSI) were collected across different counties of the U.S. in 2010 [U.S. Department of Health and Human Services]. There are 428 observations with complete response variables and predictor values in the data set. This research uses the average number of unhealthy days, the death counts, the average life expectancy, and the self-rated health status together to reflect the community

Table 3.8: Community health status results based on five-fold cross-validation

Tasks	Linear Regression (MSE)		Logistic Regression (AUC)	
Responses	Unhealthy Days	Death Counts	Life Expectancy	Health Status
MTL	0.501	0.103	0.963	0.935
SA	0.517	0.196	0.958	0.935

health status. Among these four response variables, the average life expectancy and self-rated health status are dichotomized into two levels based on the median values of all counties. Thus, the responses of interest are mixed with two continuous variables and two categorical variables. To select predictors associated with the responses of interest, we apply the multi-task feature learning method to conduct joint feature selection over 70 predictors. These predictors include overall demographic information, counts of different diseases, different causes of death, environmental conditions, and health-related risk factors. We perform five-fold cross-validation 30 times. For the training set, we apply the multi-task feature learning to combine two linear regression models and two logistic regression models together. The model selection criterion for the linear regression model is the mean squared error between the fitted values and observed values. For the logistic regression, AUC is used to measure classification accuracy. Based on the selected predictors, we fit the model on the testing sets and measure the mean squared errors, and AUCs are shown in Table 3.8. In comparison with the single-task analysis, multi-task feature learning produces smaller prediction errors and higher classification accuracy.

3.6 Technical Lemmas

This section provides some technical Lemmas used in the proofs of Lemmas and Theorem in Section 3.2.

The score function is the first-order derivative of the log quasi-likelihood function for each

task. We can show that the score functions follow the sub-exponential distribution and have a finite ψ_1 norm.

Lemma 3.5. *Based on Assumptions 3.3 - 3.5, the individual score function satisfies the sub-exponential condition (1.3) such that for some universal constant $A_1 > 0$,*

$$\|\frac{1}{\sqrt{n}}\frac{\partial \ell_k(\theta_k^*; Y_k)}{\partial \theta_{kp}}\|_{\psi_1} \leq A_1.$$

for any $k = 1, 2, \cdots, K$ and $p = 1, 2, \cdots, p_n$.

Proof. For each task, the quasi-log-likelihood score function is given by

$$\frac{\partial \ell_k(\theta_k^*; Y_k)}{\partial \theta_{kp}} = \sum_{i=1}^{n} \frac{\partial \ell_{ki}(\theta_k^*; y_{ki})}{\partial \theta_{kp}} = \sum_{i=1}^{n} \underbrace{(y_{ki} - g_k^{-1}(\eta_{ki}^*))}_{\mathcal{I}_1} \underbrace{\frac{1}{\phi_k V(g_k^{-1}(\eta_{ki}^*))} \frac{\partial g_k^{-1}(\eta_{ki})}{\partial \eta_{ki}}}_{\mathcal{I}_2} \underbrace{\frac{\partial \eta_{ki}}{\partial \theta_{kp}}}_{\mathcal{I}_3},$$

for $k = 1, 2, \cdots, K$ and $p = 1, 2, \cdots, p_n$.

From Assumption 3.3, the inverse link functions $g_k^{-1}(\eta_{ki})$ and its derivatives are well-defined and bounded. In addition, the variance function as the polynomial form of $g_k^{-1}(\eta_{ki})$ is also positive and bounded. Thus, we show that the second component $\mathcal{I}_2$ is bounded by some constant. In addition, the derivatives of the linear predictor are $\partial \eta_{ki}/\theta_{kp} = x_{kpi}$, and $\sup_{k,p,i}\{x_{kpi}\} \leq L < \infty$ based on Assumption 3.5. Thus, the component $\mathcal{I}_3$ is bounded by L.

Based on Assumption 3.5, $\mathcal{I}_1 = y_{ki} - g_k^{-1}(\eta_{ki}^*)$ is from a sub-exponential distribution with zero mean and ψ_1 norm bounded above by A_0. Let $\mathcal{K}_{ki} = \mathcal{I}_2 \times \mathcal{I}_3$. The individual score function is given by

$$\frac{\partial \ell_{ki}(\theta_k^*; y_{ki})}{\partial \theta_{kp}} = (y_{ki} - g_k^{-1}(\eta_{ki}^*))\mathcal{K}_{ki},$$

where we have $\mathcal{K}_{ki} < \mathcal{K} < \infty$ for a universal constant $\mathcal{K}$ across all tasks. We obtain that the

ψ_1 norm of the individual score function is as follows

$$\begin{aligned}\|\frac{\partial \ell_{ki}(\theta_k^*; y_{ki})}{\partial \theta_{kp}}\|_{\psi_1} &= \sup_{m\geq 1} \frac{1}{m}(E|\frac{\partial \ell_{ki}(\theta_k^*; y_{ki})}{\partial \theta_{kp}}|^m)^{1/m} \\ &\leq \sup_{m\geq 1} \mathcal{K}_{ki}\frac{1}{m}(E|g_k^{-1}(\eta_{ki}^*) - y_{ki}|^m)^{1/m} \\ &\leq \sup_{p} \mathcal{K}\|g_k^{-1}(\eta_{ki}) - y_{ki}\|_{\psi_1} \leq \mathcal{K}A_0.\end{aligned}$$

Based on the property of sub-exponential distribution [Wainwright, 2019], the ψ_1 norm of $n^{-1/2}\partial \ell_k(\theta_k^*; Y_k)/\partial \theta_{kp}$ can be bounded by some constant $A_1 \geq \mathcal{K}A_0$, such that

$$\|\frac{1}{\sqrt{n}}\frac{\partial \ell_k(\theta_k^*; Y_k)}{\partial \theta_{kp}}\|_{\psi_1} = \sup_{m\geq 1} \frac{1}{m}(E[(\frac{1}{\sqrt{n}}\frac{\partial \ell_k(\theta_k^*; Y_k)}{\partial \theta_{kp}})^m])^{1/m} \leq A_1.$$

□

Based on Lemma 3.1, we can choose the value of the penalty parameter λ_n, which is used in the proof of sign consistency.

Corollary 3.1. *Under Assumptions 3.3 - 3.7, if the penalty parameter is chosen as*

$$\lambda_n \geq \frac{4A_1}{\xi}(\sqrt{\frac{K}{n}} + \sqrt{\frac{2(d+1)K\log(p_n)}{\alpha n}}),$$

then

$$\frac{1}{\lambda_n}\sup_p \|\frac{1}{n}\nabla \mathcal{L}(\theta^*)^{(p)}\|_2 \leq \frac{\xi}{4}$$

with a probability at least $1 - 2\exp\{-d\log(p_n)\}$ *for constant* $d > 0$.

When the quasi log-likelihood is built based on the canonical link with corresponding variance function, the observed Hessian matrix is semi-positive definite under Assumption 3.6

with probability tending to one. However, the general quasi-likelihood can have a variance function set as a polynomial function of the mean. Therefore, we need to analyze the observed Hessian within a local neighborhood.

Lemma 3.6. *Based on Assumption 3.3 and 3.5, let $w_k = 1$, and there exists some $\tilde{r}$, for any $\|\theta - \theta^*\|_1 \leq \tilde{r}$, the observed Hessian can be formulated as follows,*

$$\frac{1}{n}\nabla^2\mathcal{L}(\theta) = \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n}\{f_1(\eta_{ki}) - (y_{ki} - g_k^{-1}(\eta_{ki}^*))f_2(\eta_{ki})\}x_{ki}x_{ki}^T.$$

with $\eta_{ki} = \sum_{p=1}^{p_n} x_{kpi}\theta_{kp}$ and $\eta_{ki}^* = \sum_{p=1}^{p_n} x_{kpi}\theta_{kp}^*$, and the functions of linear predictors $f_1(\eta_{ki})$ and $f_2(\eta_{ki})$ are both bounded. Furthermore, the function $f_1(\eta_{ki}) > 0$.

Proof. First, the observed Hessian can be constructed as follow

$$\begin{aligned}
\frac{1}{n}\nabla^2\mathcal{L}(\theta) =& \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n}\frac{1}{\phi_k V(g_k^{-1}(\eta_{ki}))}\Big\{\big(\frac{\partial g_k^{-1}(\eta_{ki})}{\partial \eta_{ki}}\big)^2 \\
&-(g_k^{-1}(\eta_{ki}^*) - g_k^{-1}(\eta_{ki})) \times \big(\frac{\partial^2 g_k^{-1}(\eta_{ki})}{\partial \eta_{ki}^2} - \frac{V'(g_k^{-1}(\eta_{ki}))}{V(g_k^{-1}(\eta_{ki}))}\big(\frac{\partial g_k^{-1}(\eta_{ki})}{\partial \eta_{ki}}\big)^2\big)\Big\}x_{ki}x_{ki}^T \\
&-\frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n}\frac{y_{ki} - g_k^{-1}(\eta_{ki}^*)}{\phi_k V(g_k^{-1}(\eta_{ki}))}\big(\frac{\partial^2 g_k^{-1}(\eta_{ki})}{\partial \eta_{ki}^2} - \frac{V'(g_k^{-1}(\eta_{ki}))}{V(g_k^{-1}(\eta_{ki}))}\big(\frac{\partial g_k^{-1}(\eta_{ki})}{\partial \eta_{ki}}\big)^2\big)x_{ki}x_{ki}^T.
\end{aligned}$$

Therefore, we can set

$$\begin{aligned}
f_1(\eta_{ki}) =& \frac{1}{\phi_k V(g_k^{-1}(\eta_{ki}))}\Big\{\big(\frac{\partial g_k^{-1}(\eta_{ki})}{\partial \eta_{ki}}\big)^2 \\
&-(g_k^{-1}(\eta_{ki}^*) - g_k^{-1}(\eta_{ki})) \times \big(\frac{\partial^2 g_k^{-1}(\eta_{ki})}{\partial \eta_{ki}^2} - \frac{V'(g_k^{-1}(\eta_{ki}))}{V(g_k^{-1}(\eta_{ki}))}\big(\frac{\partial g_k^{-1}(\eta_{ki})}{\partial \eta_{ki}}\big)^2\big)\Big\}
\end{aligned}$$

and by applying the approximation,

$$g_k^{-1}(\eta_{ki}^*) - g_k^{-1}(\eta_{ki}) = \frac{\partial g_k^{-1}(\eta_{ki})}{\partial \eta_{ki}}(\eta_{ki}^* - \eta_{ki}).$$

We can further show that

$$f_1(\eta_{ki}) = \frac{1}{\phi_k V(g_k^{-1}(\eta_{ki}))}\Big\{\big(\frac{\partial g_k^{-1}(\eta_{ki})}{\partial \eta_{ki}}\big)^2 - \frac{\partial g_k^{-1}(\eta_{ki})}{\partial \eta_{ki}}(\eta_{ki}^* - \eta_{ki}) \times \big(\frac{\partial^2 g_k^{-1}(\eta_{ki})}{\partial \eta_{ki}^2} - \frac{V^{'}(g_k^{-1}(\eta_{ki}))}{V(g_k^{-1}(\eta_{ki}))}\big(\frac{\partial g_k^{-1}(\eta_{ki})}{\partial \eta_{ki}}\big)^2\big)\Big\}.$$

Based on Assumption 3.3, we can set that there exists some positive constants K_1, K_2, K_3, K_4, K_5, and K_6,

$$K_1 \leq \max_{k,i} |\frac{\partial g_k^{-1}(\eta)}{\partial \eta}|_{\eta=\eta_{ki}}| \leq K_2, \text{ and } \max_{k,i} |\frac{\partial^2 g_k^{-1}(\eta)}{\partial \eta^2}|_{\eta=\eta_{ki}}| \leq K_3.$$

Since the variance function has a polynomial form of mean, then

$$K4 \leq V_k(g_k^{-1}(\eta_{ki})) \leq K5 \text{ and } V_k'(g_k^{-1}(\eta_{ki})) \leq K6.$$

Therefore, the function $f_1(\eta_{ki})$ is bounded by

$$\begin{aligned} f_1(\eta_{ki}) \geq & \frac{1}{\phi_k V(g_k^{-1}(\eta_{ki}))}\big(K_1^2 - K_2(K_3 + K_2^2 K_6/K_4)|\eta_{ki}^* - \eta_{ki}|\big) \\ \geq & \frac{1}{\phi_k V(g_k^{-1}(\eta_{ki}))}\big(K_1^2 - K_2(K_3 + K_2^2 K_6/K_4)L\|\theta - \theta^*\|_1\big) \end{aligned}$$

Therefore, as $\|\theta - \theta^*\|_1 \leq r$ and $\|x_{ki}\|_\infty \leq L$, we can set $\tilde{r} = \min\{r, K' K_3^2\}$ with constant $K' = 1/(LK_2(K_3 + K_2^2 K_6/K_4))$, and we can show that $0 < f_1(\eta_{ki}) < \infty$. In addition, we can also set

$$f_2(\eta_{ki}) = \frac{1}{\phi_k V(g_k^{-1}(\eta_{ki}))}\big(\frac{\partial^2 g_k^{-1}(\eta_{ki})}{\partial \eta_{ki}^2} - \frac{V^{'}(g_k^{-1}(\eta_{ki}))}{V(g_k^{-1}(\eta_{ki}))}\big(\frac{\partial g_k^{-1}(\eta_{ki})}{\partial \eta_{ki}}\big)^2\big),$$

which is a bounded function based on Assumption 3.3. □

Next, we can show the concentration of the observed Hessian to the expected value.

Lemma 3.7. *Under Assumptions 3.3 - 3.5, for some positive constants α and ε,*

$$P\Big(\Big|\Big|\Big|\frac{1}{n}\nabla^2\mathcal{L}(\theta)_{\mathcal{SS}} - H(\theta^*)_{\mathcal{SS}}\Big|\Big|\Big|_\infty \le \varepsilon\Big) \ge 1 - 2K\exp\Big\{-\alpha\frac{\varepsilon^2}{(A_1 s)^2}n + 2\log(s)\Big\},$$
$$P\Big(\Big|\Big|\Big|\frac{1}{n}\nabla^2\mathcal{L}(\theta)_{\mathcal{SS}^c} - H(\theta^*)_{\mathcal{SS}^c}\Big|\Big|\Big|_\infty \le \varepsilon\Big) \ge 1 - 2K\exp\Big\{-\alpha\frac{\varepsilon^2}{(A_1 s)^2}n + \log(s(p_n - s))\Big\}.$$

Proof. With the same notation as in the proof of Lemma 3.3, we can show that $\Delta H^*_{\mathcal{SS}} = \mathrm{diag}(\Delta_k H^*_{\mathcal{SS}})_{k=1}^K$. For any $\varepsilon > 0$,

$$\begin{aligned} P(|||\Delta H^*_{\mathcal{SS}}|||_\infty > \varepsilon) &\overset{(i)}{=} P(\sup_k |||\Delta_k H^*_{\mathcal{SS}}|||_\infty > \varepsilon) \\ &\le Ks^2 \sup_{k,p,p'} P(|\Delta_k H^*_{[p,p']}| > \frac{\varepsilon}{s}) \\ &\overset{(ii)}{\le} 2K\exp\big\{-\alpha\min\{\frac{\varepsilon^2}{(A_1 s)^2}, \frac{\varepsilon}{A_1 s}\}n + 2\log(s)\big\}. \end{aligned}$$

In step (i), we apply the result in Lemma 3.9. In step (ii), we apply the concentration result of the Hessian matrix based on Lemma 3.1. Using the same method, we derive that

$$\begin{aligned} P(|||\Delta H^*_{\mathcal{SS}^c}|||_\infty > \varepsilon) &\le K\sup_k P(|||\Delta_k H^*_{\mathcal{SS}^c}|||_\infty > \varepsilon) \\ &\le 2K\exp\big\{-\alpha\min\{\frac{\varepsilon^2}{(A_1 s)^2}, \frac{\varepsilon}{A_1 s}\}n + \log(p_n - s) + \log(s)\big\}. \end{aligned}$$

□

Under Assumptions, we can show that the observed Hessian is invertible on the subspace $\mathcal{S}$ with probability tending to one. The proof of Lemma 3.8 is similar to Ravikumar et al. [2010].

Lemma 3.8. *Under Assumptions 3.3 - 3.6, there exist some positive constants α and ε with*

$\varepsilon < \kappa_-$,

$$P\left(\left|\left|\left|\frac{1}{n}\nabla^2\mathcal{L}(\theta^*)_{\mathcal{SS}}\right|\right|\right|_2 \geq \kappa - \varepsilon\right) \leq 1 - 2K\exp\{-\frac{\alpha\varepsilon^2}{(A_1 s)^2}n + 2\log(s)\}.$$

Proof. With the same notation as in the proofs of Lemmas 3.3 and 3.7, we have the sub-matrix of Hessian denoted by ${}_kH^*_{\mathcal{SS}}$. Lemma 3.2 shows that with high probability, the eigenvalues of $H(\theta^*)$ are bounded and positive. Therefore, for any sub-matrix of Hessian, we have

$$\kappa_- \leq \min\Lambda({}_kH^*_{\mathcal{SS}}). \tag{3.18}$$

Based on Courant–Fischer variational representation [Ravikumar et al., 2010], we have

$$\begin{aligned}
\min\Lambda({}_kH^*_{\mathcal{SS}}) &= \min\Lambda({}_k\mathcal{H}^*_{\mathcal{SS}} + {}_kH^*_{\mathcal{SS}} - {}_k\mathcal{H}^*_{\mathcal{SS}}) \\
&= \min_{\|x\|_2=1} x^T({}_k\mathcal{H}^*_{\mathcal{SS}} + {}_kH^*_{\mathcal{SS}} - {}_k\mathcal{H}^*_{\mathcal{SS}})x \\
&\leq y^T{}_k\mathcal{H}^*_{\mathcal{SS}}y + y^T({}_kH^*_{\mathcal{SS}} - {}_k\mathcal{H}^*_{\mathcal{SS}})y,
\end{aligned}$$

where y is the unit-norm eigenvector of $\mathcal{H}^*_{\mathcal{SS}}$. Using condition 3.18, we can show that

$$\begin{aligned}
y^T{}_k\mathcal{H}^*_{\mathcal{SS}}y \geq \min\Lambda({}_k\mathcal{H}^*_{\mathcal{SS}}) &\geq \min\Lambda({}_kH^*_{\mathcal{SS}}) - y^T({}_kH^*_{\mathcal{SS}} - {}_k\mathcal{H}^*_{\mathcal{SS}})y \\
&\geq \kappa_- - |||{}_kH^*_{\mathcal{SS}} - {}_k\mathcal{H}^*_{\mathcal{SS}}|||_2.
\end{aligned}$$

Next, we have

$$\begin{aligned}
P(|||\Delta_k H^*_{\mathcal{SS}}|||_2 > \varepsilon) &\leq P(|||\Delta_k H^*_{\mathcal{SS}}|||_F > \varepsilon) \\
&\leq s^2 \sup_{k,p,p'} P(|\Delta_k H^*_{[pp']}| > \varepsilon/s) \\
&\leq 2\exp\big\{-\alpha\min\{\frac{\varepsilon^2}{(A_1 s)^2}, \frac{\varepsilon}{A_1 s}\}n + 2\log(s)\big\}.
\end{aligned}$$

As a result, we can show that with a probability at least $1 - 2K\exp\{-\frac{\alpha\varepsilon^2}{(A_1 s)^2}n + 2\log(s)\}$,

$$\Lambda({}_k\mathcal{H}^*_{\mathcal{SS}}) \geq \kappa_- - \varepsilon. \tag{3.19}$$

Furthermore, for $\varepsilon < \kappa_-$ in 3.19, we set the constant $\delta = \kappa_- - \varepsilon > 0$, such that

$$P(\Lambda({}_k\mathcal{H}^*_{\mathcal{SS}}) \leq \delta) = P(\Lambda([{}_k\mathcal{H}^*_{\mathcal{SS}}]^{-1}) \geq \delta^{-1}). \tag{3.20}$$

□

When the matrix is built by diagonal blocks of multiple submatrices, its ℓ_1 and ℓ_∞ norms are also the maximum value among all submatrices.

Lemma 3.9. *Suppose a matrix $A \in \mathbb{R}^{Kd\times Kd}$ consists of diagonal blocks such that $A = diag(A_k)_{k=1}^K$, and each block matrix has the same dimension that $A_k \in \mathbb{R}^{d\times d}$. Then,*

$$|||A|||_1 \leq \sup_k |||A_k|||_1 \text{ and } |||A|||_\infty \leq \sup_k |||A_k|||_\infty$$

Lemma 3.10. *(Bernstein Inequality) Let $X_1, \cdots, X_n$ be independent zero-mean sub-exponential random variables and $A_1 = \max_i \|X_i\|_{\psi_1}$. Then, for any $\varepsilon > 0$, we have*

$$P(\Big|\frac{1}{n}\sum_{i=1}^n X_i\Big| \geq \varepsilon) \leq 2\exp\{-\alpha\min\{\frac{\varepsilon^2}{A_1^2}, \frac{\varepsilon}{A_1}\}\}$$

with a universal constant $\alpha > 0$.

Chapter 4

Robust Multi-task Feature Learning

4.1 Introduction

The multi-task learning utilizes the intrinsic relatedness among the data sets sharing common features [Caruana, 1997, Zhang and Yang, 2017, Thung and Wee, 2018]. When modeling high-dimensional data sets, the approach of combining multiple tasks can alleviate the data scarcity problem. Mixed regularization can be used to recover the union support, which is the set of important features for at least one of the tasks [Liu et al., 2009, Obozinski et al., 2010, Zhou et al., 2011, Rakotomamonjy et al., 2011]. To analyze the estimators obtained from multiple regularized regression tasks, Lounici et al. [2011], Obozinski et al. [2011], and Wang et al. [2015b] established the statistical consistency and the estimation error bounds under regularity conditions. To jointly model different types of learning tasks, Gao and Carroll [2017] used the pseudo log-likelihood to integrate multiple data sets that are correlated and have different types of distributions. In addition, a pseudo log-likelihood based Bayesian information criterion was proposed by Gao and Carroll [2017] to perform model selection on multi-task learning, which can balance the goodness-of-fit and the complexity of the joint model. Maity et al. [2019] extended the high-dimensional data integration approach to jointly

model survival time data and binary data through a Bayesian approach. Existing multi-task learning methods are built under distributional assumptions. For example, the regression models require the data to be normally distributed, and the generalized linear models assume the response variables follow distributions from exponential families. To address this issue, the heterogeneous multi-task feature learning in Chapter 3 was proposed to model multiple data sets without distributional assumptions. Instead, the method requires that the second and higher moments of the error variables are bounded. In practice, the high-dimensional data sets collected from different sources often contain heavy-tailed data and contain large outliers in the measurements. The least squares method or likelihood-based method often performs poorly under this scenario [Catoni, 2012]. To accommodate heavy-tailed distribution and outlier contamination, we propose a robust estimation method for multi-task feature learning.

For single-task learning, the robust regularization methods are extensively studied. The Huber regression proposed by Huber [1964] is widely implemented for the robust M-estimation, and the asymptotic properties have been well established in many studies [Huber, 1964, Yohai and Maronna, 1979, Portnoy, 1985, Mammen, 1989, He and Shao, 1996]. Fan et al. [2017] proposed a penalized Huber loss to deal with heavy-tailed errors in ultra-high dimensional settings, which can provide consistent estimators with a similar optimal rate as the estimators obtained under the normality assumption. Pan et al. [2021] further investigated the optimization property and the asymptotic rate of convergence for estimators using Huber loss with both ℓ_1 regularization and non-convex penalty functions. Wang et al. [2021] proposed a data-driven method to determine the value of the robustification parameters for Huber regression that can be chosen based on the sample size, the dimension of parameters, and the moments of the error terms. To accommodate large outliers with heavy-tailed distribution, Sun et al. [2020] established the theoretical framework based on the adaptive Huber regression with relaxed moments condition for the error variables, such that the errors only need to have

a bounded $(1+\omega)$th moment for any $\omega > 0$. Instead of using the Huber loss function, Bradic et al. [2011] proposed a weighted linear combination of ℓ_1 and ℓ_2 loss functions with a weighted LASSO penalty for robust estimation and feature selection. For the robust M-estimation with folded concave regularization, Loh [2017] showed that the local optimal estimator is a unique stationary point, which coincides with the oracle estimator with a probability tending to one.

Motivated by previous studies, we propose a robust multi-task feature learning by a composite Huber loss function with the mixed $\ell_{2,1}$ regularization. For each Huber regression task, we apply the adaptive method [Sun et al., 2020] to choose the robustification parameter, which can effectively balance the robustness and unbiasedness of the estimator. The non-asymptotic deviation bounds are established based on more relaxed conditions, where the error variables have finite $(1+\omega)$th moment for any $\omega > 0$, and the covariates are independent samples from sub-Gaussian distributions. With $\omega \geq 1$, the estimation properties are similar to those obtained by existing multi-task learning method [Gao and Carroll, 2017]. For $\omega \in (0, 1)$, we obtain a slower convergence rate for the estimator. Based on the notion of the restricted strong convexity (RSC) condition in Agarwal et al. [2012b], Meinshausen and Yu [2009], Loh [2017], we provide a modified RSC condition for the composite Huber loss function that can be used with the mixed regularization. Thus, the proposed loss function is not required to be strongly convex under high dimensional settings. The distributional assumptions required in this thesis are more general and weaker in comparison with existing multi-task learning methods.

In this chapter, we establish the statistical consistency and the asymptotic properties of the estimators obtained by robust multi-task feature learning. The ℓ_2 and ℓ_1 error bounds of the penalized estimates are evaluated, which are in line with the adaptive Huber regression in Sun et al. [2020]. In addition, we show that the regularized estimator can recover the union support correctly, with probability tending to one. The proposed robust multi-task feature

learning method can be regarded as a special group-wise variable selection procedure using the adaptive Huber loss function. To our best knowledge, the properties of the adaptive Huber loss for the multi-task problem have not been explored in the literature. For the optimization algorithm, Liu et al. [2014] proposed to use the block coordinates descent method for the group LASSO-based robust estimation. Li and Sherwood [2021] provided an R package to select group-wise variables for both quantile and robust mean regression. In comparison with previous algorithmic works, we propose to use the composite gradient descent algorithm [Nesterov, 2013], and we prove that the updated iterations converge to the optimal solution at a geometric rate.

To evaluate the model performance based on the selected features, the pseudo Bayesian information criterion [Gao and Carroll, 2017] was developed based on the composite likelihood losses with a model complexity penalty, which was applied to multiple correlated data sets. Dai et al. [2020] extended the information criterion to model multiple quantile regressions and proposed a modified version of the Bayesian information criterion through pooled check loss functions. We treat the Huber loss function as a hybrid of least square loss and Least Absolute Deviation (LAD) and construct the robust Bayesian information criterion with the composite Huber loss based on Gao and Carroll [2017], Wu et al. [2023], and Dai et al. [2020].

The organization of this chapter is as follows. In Section 4.2, we first set up the model for the robust multi-task feature learning with the Huber loss function and the mixed $\ell_{2,1}$ regularization. The main theoretical results are provided in Section 4.3. We establish the non-asymptotic error bounds and the sign recovery consistency of the proposed estimates under the regularity conditions. In Section 4.3.3, the optimization method is provided, and the convergence rate of the optimization algorithm is discussed. The numerical studies are given in Section 4.4 and 4.5 to examine the performance of the proposed method.

4.2 Model Setup

Suppose there are multiple statistical tasks that are used to model K different data sets. In any kth task, there are n_k independent and continuous responses $Y_k = (y_{k1}, \cdots, y_{ki}, \cdots, y_{kn_k})^T$, $k = 1, 2, \cdots, K$. To model the relationship between the response y_{ki} and the covariates x_{ki}, we construct the linear regression model,

$$y_{ki} = x_{ki}^T \theta_k^* + u_{ki}^*, \tag{4.1}$$

where $x_{ki} = (x_{ki1}, \ldots, x_{kip_n})^T$ is the vector of covariates for the ith observation in the kth task, θ_k^* denotes the vector of true regression coefficients, and u_{ki}^* is the random error. This proposed model is aimed at dealing with heavy-tailed error distribution and large outliers. Therefore, we do not require a specific distributional assumption on error term u_{ki}^*. Instead, we make the following assumptions.

Assumption 4.1. *In any kth task,*

1. *The random errors u_{ki}^*'s are independent for $i = 1, 2, \cdots, n_k$, which satisfies $E(u_{ki}^*|x_{ki}) = 0$ and $E(|u_{ki}^*|^{1+\omega}|x_{ki}) = v_{ki} < \infty$ for some $\omega > 0$;*

2. *We consider a random design matrix and the covariates x_{ki}'s are i.i.d vectors from sub-Gaussian distribution such that $P(|x_{ki}^T u| \geq \varepsilon) \leq 2\exp\{-\varepsilon^2/A_0^2\}$ for any $\varepsilon > 0$ and any unit vector $u \in \mathbb{R}^{p_n}$. The covariance matrix $Cov(x_{ki}) = \Sigma_k$ has eigenvalues bounded such that $0 < \alpha_l \leq \Lambda_{\min}(\Sigma_k) \leq \Lambda_{\max}(\Sigma_k) \leq \alpha_u < \infty$.*

When the second moment of u_{ki}^* is bounded with $\omega \geq 1$, we have the standard ordinary regression setting. If $\omega \in (0, 1)$, the second moment of u_{ki}^* is not guaranteed to be bounded. Thus, the model can be used for data with heavy-tailed error distributions or large outliers. In addition, $E(|u_{ki}^*|^{1+\omega}|x_{ki}) = v_{ki}$ can be dependent on x_{ki}, which can be used to model

heteroscedastic regression tasks. For example, we can let the error term $u_{ki}^* = \|\theta_k^*\|_2^{-1}(x_{ki}^T\theta_k^*)\varepsilon_{ki}$, where ε_{ki} satisfies $E(\varepsilon_{ki}) = 0$ and $E(|\varepsilon_{ki}^*|^{1+\omega}) < \infty$.

4.2.1 Huber Loss Function

The multi-task feature learning can combine different types of statistical tasks and implement support union recovery for the high-dimensional problem by penalized M-estimation. The objective function usually contains the loss function and the penalty function as follows,

$$Q(\theta) = \mathcal{L}(\theta) + \mathcal{R}(\theta).$$

The proposed loss function $\mathcal{L}(\theta)$ can combine individual loss functions from different tasks,

$$\mathcal{L}(\theta) = -\sum_{k=1}^{K}\sum_{i=1}^{n_k} w_k \ell_{ki}(\theta_k; y_{ki}, x_{ki}). \tag{4.2}$$

According to the relative importance of the task, users can assign positive weights w_k to each individual loss function [**??**]. In literature, the individual loss function $\ell_{ki}(\theta_k; y_{ki}, x_{ki})$ can be the square loss function [Lounici et al., 2011, Obozinski et al., 2010, 2011], the log-likelihood function [Gao and Carroll, 2017], or quasi log-likelihood function in Chapter 3, which are all sensitive to heavy-tailed error distributions or large outliers in the measurements. In our approach, the individual loss can be modeled by the Huber loss function [Huber, 1964],

$$\ell_{ki}(\theta_k; y_{ki}, x_{ki}) = \begin{cases} u_{ki}^2/2 & \text{if } |u_{ki}| \le \tau_k, \\ \tau_k|u_{ki}| - \tau^2/2 & \text{if } |u_{ki}| > \tau_k, \end{cases}$$

where $u_{ki} = y_{ki} - x_{ki}^T\theta_k$. For each task, the Huber loss uses the robustification parameter τ_k to control the behavior of the loss function. When the error term u_{ki} is smaller than the

value of τ_k, the loss function has a quadratic form, and for large errors, the function becomes linear with respect to u_{ki}. If we let $\tau_k \to \infty$, the loss function becomes the least square loss, and if $\tau_k = 0$, the model is reduced to a Least Absolute Deviation (LAD) regression model. Thus, the balance of the unbiasedness and robustness can be controlled based on the choice of the parameters τ_k.

The adaptive method is commonly used for high-dimension scaling, in which the parameter can be chosen based on the sample size, the dimension of parameters, and the moments of error term [Lepskii, 1992, Sun et al., 2020, Wang et al., 2021]. Without loss of generality, we define $n = n_k$ for all tasks, $v_k = n^{-1} \sum_{i=1}^{n} v_{ki}$, and $v_{\max} = \max_k\{v_k\}$. For $\omega \in (0, 1)$, the range of the parameter τ_k can be set as

$$(v_k)^{1/2} \lesssim \tau_k \lesssim \big(\frac{n}{\log(p_n)}\big)^{\max\{1/2, 1/(1+\omega)\}}. \tag{4.3}$$

Since the relationship between different data sets is unknown, the individual loss functions $\ell_{ki}(\theta_k; y_{ki}, x_{ki})$ can be correlated across K different tasks. Let $\nabla\mathcal{L}(\theta)$ denote the first derivative of the proposed loss, and $\nabla^2\mathcal{L}(\theta)$ denote the observed Hessian matrix. The sensitivity and variability matrix are given by

$$H(\theta) = E(n^{-1}\nabla^2\mathcal{L}(\theta)) \text{ and } J(\theta) = Cov(n^{-1}\nabla\mathcal{L}(\theta)).$$

With correlated data sets across different tasks, the second Bartlett identity no longer holds, i.e., $H(\theta) \neq J(\theta)$. We need to estimate both matrices to perform joint inference on correlated tasks.

4.2.2 Mixed Regularization

The support union recovery under high dimensionality can be achieved by the mixed regularization (1.1) [Obozinski et al., 2010, Gao and Carroll, 2017, Gong et al., 2013]. Let $\mathcal{S}$ denote the true union support $\mathcal{S} := \{p : \|\theta^{(p)}\|_2 \neq 0\}$, and $|\mathcal{S}| = s$. The sample size and the dimension of the parameters are assumed to satisfy the following condition, which is commonly used in the literature of penalized M-estimation Ravikumar et al. [2010], Loh and Wainwright [2017], Li et al. [2021].

Assumption 4.2. *It is assumed that* $n \gtrsim s^2 \log(p_n)$. *In addition, the true parameter vector* $\|\theta^*\|_1 \leq R$ *for some* $R > 0$.

The penalty function $\mathcal{R}(\theta)$ in the mixed $\ell_{2,1}$ regularization is proposed as [Tibshirani, 1996, Yuan and Lin, 2006].

$$\mathcal{R}(\theta) = n\lambda_n \|\theta\|_{2,1},$$

with the penalty parameter λ_n. The subdifferential of the mixed $\ell_{2,1}$ norm is defined as $z^{(p)} = \partial\|\theta^{(p)}\|_2/\partial\theta^{(p)}$, such that

$$\begin{cases} z^{(p)} = \frac{\theta^{(p)}}{\|\theta^{(p)}\|_2}, & \text{if } \|\theta^{(p)}\|_2 \neq 0, \\ \|z^{(p)}\|_2 < 1, & \text{if } \|\theta^{(p)}\|_2 = 0, \end{cases}$$

for any $p = 1, 2, \cdots, p_n$.

The mixed $\ell_{2,1}$ norm has the following properties:

1. For any subset $\mathcal{E} \in \{1, 2, \cdots, p_n\}$, the mixed norm can be decomposed as $\|\theta\|_{2,1} = \|\theta_{\mathcal{E}}\|_{2,1} + \|\theta_{\mathcal{E}^c}\|_{2,1}$;

2. For any two vectors θ_1 and θ_2, $\|\theta_1\|_{2,1} - \|\theta_2\|_{2,1} - y_2^T(\theta_1 - \theta_2) \geq 0$ with y_2 as the

subdifferential of $\|\theta_2\|_{2,1}$.

The first decomposition property of the penalty function can be applied to calculate the estimation error bounds, which is discussed in Negahban et al. [2012]. The second property comes from the definition of subdifferential.

The optimal estimator of the robust multi-task feature learning can be obtained by

$$\tilde{\theta} \in \arg \min_{\|\theta\|_1 \leq R} \{\mathcal{L}(\theta) + \mathcal{R}(\theta)\}. \tag{4.4}$$

The constraint $\|\theta\|_1 \leq R$ is used to ensure the existence of the global optima $\tilde{\theta}$ [Loh and Wainwright, 2015]. In addition, Assumption 4.2 with $\|\theta^*\|_1 \leq R$ can also make the true parameter θ^* a feasible point for the program 4.4.

If the penalized estimate $\hat{\theta}$ satisfies

$$(n^{-1}\nabla\mathcal{L}(\hat{\theta}) + \lambda_n \hat{z})^T(\theta - \hat{\theta}) \geq 0, \tag{4.5}$$

with $\hat{z}$ denoting the subdifferential of $\|\hat{\theta}\|_{2,1}$, and $\hat{\theta}$ denoting stationary point of the program (4.4) [Bertsekas, 1999]. When $\hat{\theta}$ is an interior point of the constraint set, the equality in (4.5) holds. The set of stationary points includes the optimal estimator $\tilde{\theta}$ defined by (4.4).

4.3 Methodology

4.3.1 Theoretical Conditions

The restricted strong convexity (RSC) condition introduced by Agarwal et al. [2012b] and Negahban et al. [2012] can be used to analyze the statistical and optimization properties of penalized M-estimators. Loh and Wainwright [2015] and Loh [2017] imposed the condition on the objective function with non-convex loss functions and regularization. The Huber loss

function with the LASSO penalty can satisfy the RSC condition under both deterministic and random design [Fan et al., 2017, 2018, Sun et al., 2020]. As we use grouped penalty in our model, we introduce a modified version of the RSC condition as follows.

Assumption 4.1. *(Local RSC Condition) There exist constants $\kappa_l > 0$, $\tau_l \geq 0$, and $r > 0$, for all $\theta_1, \theta_2 \in \mathbb{B}_r(\theta^*)$, such that the first-order Taylor error of the loss function satisfies*

$$\begin{aligned} \mathcal{T}(\theta_1, \theta_2) &= (n^{-1}\nabla\mathcal{L}(\theta_1) - n^{-1}\nabla\mathcal{L}(\theta_2))^T(\theta_1 - \theta_2) \\ &\geq \kappa_l\|\theta_1 - \theta_2\|_2^2 - \tau_l\frac{\log(p_n)}{n}\|\theta_1 - \theta_2\|_{2,1}^2. \end{aligned} \tag{4.6}$$

Under high dimensionality, we assume the loss functions satisfy the RSC condition locally in the ℓ_2 ball $\mathbb{B}_r(\theta^*)$ around θ^*. When $\kappa_l > 0$ and $\tau_l = 0$, $\mathcal{T}(\theta_1, \theta_2)$ is bounded below by a positive quadratic term, which implies the loss function is strongly convex. In addition, due to the relation $\|v\|_{2,1} \leq \|v\|_1 \leq \sqrt{K}\|v\|_{2,1}$, the term $\|\theta_1 - \theta_2\|_{2,1}^2$ in (4.6) can be replaced by $\|\theta_1 - \theta_2\|_1^2$, which is the term appeared in the RSC condition defined by Loh [2017].

The irrepresentable condition introduced by Zhao and Yu [2006] is required for the LASSO estimator to recover the correct support, which constrains the strength of dependency between the predictors in the true model and the other unimportant predictors. In addition, van de Geer and Bühlmann [2009] provided a comprehensive analysis for more general conditions imposed to the design matrix.

Assumption 4.2. *The expected Hessian matrix of the proposed loss function is defined as $\Sigma = diag_k\{\Sigma_k\}$. Let $\mathcal{S}$ be denoted as the support of the true parameters and $\mathcal{S}^c$ be the complement. There exists a parameter $\xi < 1$ such that*

$$\sqrt{K}\left|\left|\left|\Sigma_{\mathcal{S}^c\mathcal{S}}\Sigma_{\mathcal{S}\mathcal{S}}^{-1}\right|\right|\right|_\infty \leq 1 - \xi.$$

In the multi-task problem, Assumption 4.2 is needed to recover the union support, which

can be considered as a special case of the block-wise mutual incoherence condition for group LASSO estimation [Bach, 2008, Eldar et al., 2010, Jalali et al., 2010, Hebiri and van de Geer, 2011].

4.3.2 Statistical Consistency

Next, we establish the finite sample error bounds for the proposed estimator. Lemma 4.1 provides the large deviation bound for the ℓ_2-norm of the grouped score functions.

Lemma 4.1. *Based on Assumptions 4.1-4.2,*

$$\|\frac{1}{n}\nabla\mathcal{L}(\theta^*)\|_{2,\infty} \leq (4K\tau_{\max}^{\max\{1-\omega,0\}}A_0^2\upsilon_{\max})^{1/2}(\sqrt{\frac{\log(p_n)}{n}}+\sqrt{\frac{1}{n}})+2\tau_{\max}A_0\frac{K\log(p_n)}{n}, \quad (4.7)$$

with probability at least $1-C_1\exp\{-C_2\log(p_n)\}$ *for some positive constants* C_1 *and* C_2*.*

Proof. We let the score function in the kth task be denoted by

$$U_n(\theta_{kp}) = \sum_{i=1}^{n} U_i(\theta_{kp}) = \sum_{i=1}^{n}\frac{\partial\ell_{ki}(\theta_k; y_{ki}, x_{ki})}{\partial\theta_{kp}},$$

for any $p = 1, 2, \cdots, p_n$. Lemma 4.2 shows that with fixed p,

$$P(|\frac{1}{n}U_n(\theta_{kp})| \geq 2\nu_k\sqrt{\log(p_n)}+2\alpha_k\log(p_n)) \leq 2\exp\{-2\log(p_n)\}, \quad (4.8)$$

where

$$\nu_k = 2\tau_k^{\max\{(1-\omega)/2,0\}}A_0(\frac{\upsilon_k}{n})^{1/2}, \text{ and } \alpha_k = \frac{\tau_k A_0}{n}.$$

We have $\|n^{-1}\nabla\mathcal{L}(\theta^*)^{(p)} - E(n^{-1}\nabla\mathcal{L}(\theta^*)^{(p)})\|_2 = \{\sum_{k=1}^{K}(\frac{1}{n}U_n(\theta_{kp}) - E[\frac{1}{n}U_n(\theta_{kp})])^2\}^{1/2}$.

Thus,

$$P(\|n^{-1}\nabla\mathcal{L}(\theta^*)^{(p)} - E(n^{-1}\nabla\mathcal{L}(\theta^*)^{(p)})\|_2 \geq \varepsilon) \leq \sum_{k=1}^{K} P(|\frac{1}{n}U_n(\theta_{kp})| \geq \frac{\varepsilon}{\sqrt{K}}).$$

If we apply the concentration inequality (4.8), we can show that with $\tau_{\max} = \max_k\{\tau_k\}$ and $v_{\max} = \max_k\{v_k\}$,

$$\|\frac{1}{n}(\nabla\mathcal{L}(\theta^*)^{(p)} - E(\nabla\mathcal{L}(\theta^*)^{(p)}))\|_2 \leq 4\tau_{\max}^{\max\{(1-\omega)/2,0\}}A_0 v_{\max}^{1/2}\sqrt{\frac{K\log(p_n)}{n}} + 2\tau_{\max}A_0\frac{K\log(p_n)}{n}$$

with probability at least $1 - 2K\exp\{-2\log(p_n)\}$.

Furthermore, the expectation of $\|n^{-1}\nabla\mathcal{L}(\theta^*)^{(p)}\|_2$ can be bounded by

$$E(\|n^{-1}\nabla\mathcal{L}(\theta^*)^{(p)}\|_2) \leq \big\{\sum_{k=1}^{K} E\big[(n^{-1}U_n(\theta^*_{kp}))^2\big]\big\}^{1/2} \leq \big(4\tau_{\max}^{1-\omega}A_0^2\frac{Kv_{\max}}{n}\big)^{1/2}.$$

We can conclude that for any p,

$$\sup_p \|n^{-1}\nabla\mathcal{L}(\theta^*)^{(p)}\|_2 \leq (4K\tau_{\max}^{\max\{1-\omega,0\}}A_0^2 v_k)^{1/2}(\sqrt{\frac{\log(p_n)}{n}} + \sqrt{\frac{1}{n}}) + 2\tau_{\max}A_0\frac{K\log(p_n)}{n},$$

with probability at least $1 - C_1\exp\{-C_2\log(p_n)\}$ for some constants C_1 and C_2.

□

Theorem 4.1. *Based on Assumptions 4.1-4.1, if r, R, and λ_n can satisfy*

$$\max\{4\|\frac{1}{n}\nabla\mathcal{L}(\theta^*)\|_{2,\infty}, 8\tau_l R\frac{\log(p_n)}{n}\} \leq \lambda_n \lesssim r/\sqrt{s}, \tag{4.9}$$

then there exists a stationary point $\hat{\theta}$ obtained from (4.5) such that $\|\hat{\theta} - \theta^*\|_2 \leq r$. In addition,

$$\|\hat{\theta} - \theta^*\|_2 \leq \frac{3\lambda_n\sqrt{s}}{2\kappa_l} \text{ and } \|\hat{\theta} - \theta^*\|_1 \leq \frac{6\lambda_n\sqrt{Ks}}{\kappa_l}.$$

Proof. We assume that there exists a stationary point $\hat{\theta} \in \mathbb{B}_r(\theta^*)$, which can satisfy

$$0 \geq \frac{1}{n}\nabla Q(\hat{\theta})^T(\hat{\theta} - \theta^*) = (\frac{1}{n}\nabla \mathcal{L}(\hat{\theta}) + \lambda_n \hat{z})^T(\hat{\theta} - \theta^*).$$

We can show that

$$\begin{aligned} 0 \geq & (\frac{1}{n}\nabla \mathcal{L}(\theta^*) + \frac{1}{n}(\nabla \mathcal{L}(\hat{\theta}) - \nabla \mathcal{L}(\theta^*)) + \lambda_n \hat{z})^T(\hat{\theta} - \theta^*) \\ = & (\frac{1}{n}\nabla \mathcal{L}(\theta^*) + \lambda_n \hat{z})^T(\hat{\theta} - \theta^*) + \mathcal{T}(\hat{\theta}, \theta^*). \end{aligned}$$

Based on Assumption 4.1, we have

$$-(\frac{1}{n}\nabla \mathcal{L}(\theta^*) + \lambda_n \hat{z})^T(\hat{\theta} - \theta^*) \geq \kappa_l \|\hat{\theta} - \theta^*\|_2^2 - \tau_l \frac{\log(p_n)}{n}\|\hat{\theta} - \theta^*\|_{2,1}^2. \tag{4.10}$$

First, we can use (4.10) to obtain

$$\underbrace{\tau_l \frac{\log(p_n)}{n}\|\hat{\theta} - \theta^*\|_{2,1}^2}_{\mathcal{I}_1} - \underbrace{\frac{1}{n}\nabla \mathcal{L}(\theta^*)^T(\hat{\theta} - \theta^*)}_{\mathcal{I}_2} - \underbrace{\lambda_n \hat{z}^T(\hat{\theta} - \theta^*)}_{\mathcal{I}_3} \geq 0. \tag{4.11}$$

We define subspace $\mathcal{E}$, such that $\mathcal{S} \subseteq \mathcal{E}$ and $|\mathcal{E}| = cs$ for some c. The first component $\mathcal{I}_1$ can be bounded by the condition 4.9 for $\|\hat{\theta} - \theta^*\|_1 \leq 2R$,

$$\mathcal{I}_1 : \tau_l \frac{\log(p_n)}{n}\|\hat{\theta} - \theta^*\|_{2,1}^2 \leq \frac{\lambda_n}{4}\|\hat{\theta} - \theta^*\|_{2,1} = \frac{\lambda_n}{4}(\|(\hat{\theta} - \theta^*)_{\mathcal{E}}\|_{2,1} + \|(\hat{\theta} - \theta^*)_{\mathcal{E}^c}\|_{2,1}). \tag{4.12}$$

According to the condition (4.9) that the score function satisfies $\|n^{-1}\nabla \mathcal{L}(\theta^*)\|_{2,\infty} \leq \lambda_n/4$, we can apply Lemma 4.10 to obtain

$$\mathcal{I}_2 : -\nabla \mathcal{L}(\theta^*)^T(\hat{\theta} - \theta^*) \leq \frac{\lambda_n}{4}\|\hat{\theta} - \theta^*\|_{2,1} = \frac{\lambda_n}{4}(\|(\hat{\theta} - \theta^*)_{\mathcal{E}}\|_{2,1} + \|(\hat{\theta} - \theta^*)_{\mathcal{E}^c}\|_{2,1}). \tag{4.13}$$

In addition,

$$\mathcal{I}_3 : \lambda_n \hat{z}^T(\hat{\theta} - \theta^*) = \lambda_n \hat{z}_{\mathcal{E}}^T(\hat{\theta} - \theta^*)_{\mathcal{E}} + \lambda_n \hat{z}_{\mathcal{E}^c}^T(\hat{\theta} - \theta^*)_{\mathcal{E}^c}.$$

In the subspace $\mathcal{E}$, we can apply Lemma 4.10 and get

$$-\lambda_n \hat{z}_{\mathcal{E}}^T(\hat{\theta} - \theta^*)_{\mathcal{E}} \leq \lambda_n \|\hat{z}_{\mathcal{E}}^T\|_{2,\infty} \|(\hat{\theta} - \theta^*)_{\mathcal{E}}\|_{2,1} = \lambda_n \|(\hat{\theta} - \theta^*)_{\mathcal{E}}\|_{2,1}.$$

For the components in the subspace $\mathcal{E}^c$, the true parameters $\theta^*_{\mathcal{E}}$ are all zero. Therefore,

$$\lambda_n \hat{z}_{\mathcal{E}^c}^T(\hat{\theta} - \theta^*)_{\mathcal{E}^c} = \lambda_n \hat{z}_{\mathcal{E}^c}^T \hat{\theta}_{\mathcal{E}^c} = \lambda_n \|(\hat{\theta} - \theta^*)_{\mathcal{E}^c}\|_{2,1}.$$

Thus, we have

$$\mathcal{I}_3 : -\lambda_n \hat{z}^T(\hat{\theta} - \theta^*) \leq \lambda_n (\|(\hat{\theta} - \theta^*)_{\mathcal{E}}\|_{2,1} - \|(\hat{\theta} - \theta^*)_{\mathcal{E}^c}\|_{2,1}). \tag{4.14}$$

By combining (4.12), (4.13), and (4.14), (4.10) can be used to derive

$$\|(\hat{\theta} - \theta^*)_{\mathcal{E}^c}\|_{2,1} \leq 3\|(\hat{\theta} - \theta^*)_{\mathcal{E}}\|_{2,1}. \tag{4.15}$$

Next, we can apply the results above and obtain

$$\begin{aligned}
\kappa_l \|\hat{\theta} - \theta^*\|_2^2 &\leq \frac{3}{2}\lambda_n \|(\hat{\theta} - \theta^*)_{\mathcal{E}}\|_{2,1} - \frac{1}{2}\lambda_n \|(\hat{\theta} - \theta^*)_{\mathcal{E}^c}\|_{2,1} \\
&\leq \frac{3}{2}\lambda_n \sqrt{|\mathcal{E}|} \|\hat{\theta} - \theta^*\|_2 \\
&= \frac{3}{2}\lambda_n \sqrt{s} \|\hat{\theta} - \theta^*\|_2,
\end{aligned}$$

where $c = 1$ without loss of generality. Furthermore, we can conclude

$$\|\hat{\theta} - \theta^*\|_2 \leq \frac{3\lambda_n\sqrt{s}}{2\kappa_l}.$$

For the ℓ_1 norm estimation error, we can show that

$$\|\hat{\theta} - \theta^*\|_1 \leq 4\sqrt{K}\|(\hat{\theta} - \theta^*)_{\mathcal{E}}\|_{2,1} \leq 4\sqrt{Ks}\|(\hat{\theta} - \theta^*)_{\mathcal{E}}\|_2 \leq \frac{6\lambda_n\sqrt{K}s}{\kappa_l}.$$

Since we have $\lambda_n \lesssim r/\sqrt{s}$, the ℓ_2 norm of the estimation error satisfies $\|\hat{\theta} - \theta^*\|_2 \leq r$. Therefore, $\hat{\theta}$ is the interior point of the sphere of radius r around θ^*, which ensures the existence of such local stationary point.

□

Theorem 4.1 shows that there exists a local stationary point $\hat{\theta}$ obtained from (4.5) that has estimation error bounds similar to the results in literature Fan et al. [2017], Sun et al. [2020]. From Lemma 4.1, if $\omega \geq 1$, the grouped score function satisfies $\|n^{-1}\nabla\mathcal{L}(\theta^*)\|_{2,\infty} \lesssim (\log(p_n)/n)^{1/2} + \tau_{\max}\log(p_n)/n$ with probability tending to 1. We set the robustification parameters $\tau_k \lesssim (n/\log(p_n))^{1/2}$ for any kth task and choose the penalty parameter $\lambda_n \asymp (\log(p_n)/n)^{1/2}$ to satisfy the condition (4.9). In this way, the error bounds are identical to those of the group LASSO estimators,

$$\|\hat{\theta} - \theta^*\|_2 \lesssim \big(s\frac{\log(p_n)}{n}\big)^{1/2} \text{ and } \|\hat{\theta} - \theta^*\|_1 \lesssim s\big(\frac{\log(p_n)}{n}\big)^{1/2}.$$

For the cases $\omega \in (0, 1)$, we have

$$\|\frac{1}{n}\nabla\mathcal{L}(\theta^*)\|_{2,\infty} \lesssim (\tau_{\max}^{1-\omega}\frac{\log(p_n)}{n})^{1/2} + \tau_{\max}\frac{\log(p_n)}{n},$$

with probability tending to one. We can set $\tau_k \lesssim (n/\log(p_n))^{1/(1+\omega)}$ based on (4.3) and

choose $\lambda_n \asymp (\log(p_n)/n)^{\omega/(1+\omega)}$. Then, the convergence rates are as follows

$$\|\hat{\theta} - \theta^*\|_2 \lesssim s^{1/2}\big(\frac{\log(p_n)}{n}\big)^{\omega/(1+\omega)} \text{ and } \|\hat{\theta} - \theta^*\|_1 \lesssim s\big(\frac{\log(p_n)}{n}\big)^{\omega/(1+\omega)}.$$

Furthermore, as we choose the penalty $\lambda_n \asymp (\log(p_n)/n)^{\min\{1/2,\omega/(1+\omega)\}}$, then for $R \lesssim s$, the condition $R \lesssim n\lambda_n/\log(p_n)$ in (4.9) can be satisfied based on the sample scaling $n \gtrsim s^2 \log(p_n)$ from Assumption 4.2.

Theorem 4.2. *Based on Assumptions 4.1-4.2, if the condition (4.9) holds and the penalty parameter λ_n is chosen as*

$$\lambda_n \geq \frac{4}{\xi}\{(4K\tau_{\max}^{\max\{1-\omega,0\}}A_0^2 v_{\max})^{1/2}(\sqrt{\frac{\log(p_n)}{n}} + \sqrt{\frac{1}{n}}) + 2\tau_{\max}A_0\frac{K\log(p_n)}{n}\}, \tag{4.16}$$

and the parameters satisfy

$$\min_{k,p\in\mathcal{S}} |\theta_{kp}| \geq \frac{5\sqrt{s}\lambda_n}{4\alpha_l},$$

then there exists a stationary point $\hat{\theta} \in \mathbb{B}_r(\theta^*)$ obtained from (4.5) that satisfies sign$(\hat{\theta})$ = sign(θ^*) with probability at least $1 - C_3\exp\{-C_4\min\{s, \log(p_n)\}\}$ for some positive constants C_3 and C_4.

Proof. The proof follows the primal-dual witness construction. We consider a local point $\hat{\theta}_{\mathcal{S},0} = (\hat{\theta}_{\mathcal{S}}, \mathbf{0}) \in \mathbb{B}_{\mathbf{r}}(\theta^*)$ satisfying

$$\hat{\theta}_{\mathcal{S},0} = \underset{\theta\in\mathbb{R}^{Ks}:\|\theta\|_1\leq R}{\arg\min}\{n^{-1}\mathcal{L}(\theta) + \lambda_n\|\theta\|_{2,1}\}. \tag{4.17}$$

We can apply Theorem 4.1 to this restricted program (4.17), so that $\hat{\theta}_{\mathcal{S},0}$ can satisfy

$$\|\hat{\theta}_{\mathcal{S},0} - \theta^*\|_2 \leq \frac{3\lambda_n\sqrt{s}}{2\kappa_l} \text{ and } \|\hat{\theta}_{\mathcal{S},0} - \theta^*\|_1 \leq \frac{6\lambda_n s}{\kappa_l}.$$

Therefore, $\hat{\theta}_{\mathcal{S},0}$ is an interior point of the restricted program (4.17), and we have the zero-subgradient condition

$$\frac{1}{n}\nabla\mathcal{L}(\hat{\theta}_{\mathcal{S},0}) + \lambda_n \hat{z} = \mathbf{0} \tag{4.18}$$

with $\hat{z}$ denoting the subdifferential of $\|\hat{\theta}_{\mathcal{S},0}\|_{2,1}$. The equations (4.18) can be partitioned into two sets $\mathcal{S}$ and $\mathcal{S}^c$:

$$-\frac{1}{n}\nabla\mathcal{L}(\hat{\theta}_{\mathcal{S},0})_{\mathcal{S}} = \lambda_n \hat{z}_{\mathcal{S}}, \tag{4.19a}$$

$$-\frac{1}{n}\nabla\mathcal{L}(\hat{\theta}_{\mathcal{S},0})_{\mathcal{S}^c} = \lambda_n \hat{z}_{\mathcal{S}^c}, \tag{4.19b}$$

with $\max_{p\in\mathcal{S}^c} \|\hat{z}^{(p)}\|_2 < 1$.

Next, we expand the estimating equation (4.18) as follows

$$\begin{aligned} 0 = \frac{1}{n}\nabla\mathcal{L}(\hat{\theta}_{\mathcal{S},0}) + \lambda_n \hat{z} &= \frac{1}{n}\nabla\mathcal{L}(\theta^*) + \frac{1}{n}\nabla\mathcal{L}(\hat{\theta}_{\mathcal{S},0}) - \frac{1}{n}\nabla\mathcal{L}(\theta^*) + \lambda_n \hat{z} \\ &= \frac{1}{n}\nabla\mathcal{L}(\theta^*) + \frac{X^T X}{n}\hat{\Delta} + \mathcal{R} + \lambda_n \hat{z}. \end{aligned} \tag{4.20}$$

Let $\hat{\Delta} = \hat{\theta}_{\mathcal{S},0} - \theta^*$ and $\tilde{u}_{ki} = y_{ki} - x_{ki}^T(\theta_k^* + t\hat{\Delta}_k) = u_{ki}^* - t x_{ki}^T \hat{\Delta}_k$. Equation (4.20) is obtained by

$$\begin{aligned} \frac{1}{n}(\nabla\mathcal{L}(\hat{\theta}_{\mathcal{S},0}) - \nabla\mathcal{L}(\theta^*)) &= \int_0^1 \frac{1}{n}\nabla^2\mathcal{L}(\theta^* + t\hat{\Delta})\hat{\Delta}dt \\ &= \frac{X^T X}{n}\hat{\Delta} - \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n}\int_0^1 x_{ki}x_{ki}^T\hat{\Delta}_k 1(|\tilde{u}_{ki}| > \tau_k)dt, \end{aligned}$$

where

$$\mathcal{R} = -\frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n}\int_0^1 x_{ki}x_{ki}^T\hat{\Delta}_k 1(|\tilde{u}_{ki}| > \tau_k)dt.$$

For simplicity, we define

$$\hat{\Sigma} = X^T X/n \text{ and } \frac{1}{n}\nabla\mathcal{L}(\theta^*) = X^T\epsilon^*/n$$

with $\epsilon^* = (\epsilon^*_{11}, \cdots, \epsilon^*_{ki}, \cdots, \epsilon^*_{Kn})$ and $\epsilon^*_{ki} = \min\{|u^*_{ki}|, \tau_k\}\text{sign}(u^*_{ki})$ for $k = 1, 2, \cdots, K$ and $i = 1, 2, \cdots, n$.

Therefore, the equations (4.19a) and (4.19b) can be rearranged in block format based on (4.20) as follows,

$$\begin{bmatrix} \hat{\Sigma}_{\mathcal{S}\mathcal{S}} & \hat{\Sigma}_{\mathcal{S}\mathcal{S}^c} \\ \hat{\Sigma}_{\mathcal{S}^c\mathcal{S}} & \hat{\Sigma}_{\mathcal{S}^c\mathcal{S}^c} \end{bmatrix} \begin{pmatrix} \hat{\Delta}_{\mathcal{S}} \\ \mathbf{0} \end{pmatrix} + \begin{pmatrix} X^T_{\mathcal{S}}\epsilon^*/n \\ X^T_{\mathcal{S}^c}\epsilon^*/n \end{pmatrix} + \begin{pmatrix} \mathcal{R}_{\mathcal{S}} + \lambda_n\hat{z}_{\mathcal{S}} \\ \mathcal{R}_{\mathcal{S}^c} + \lambda_n\hat{z}_{\mathcal{S}^c} \end{pmatrix} = \mathbf{0}. \tag{4.21}$$

Next, we need to show that $\|\hat{\Delta}_{\mathcal{S}}\|_\infty$ is bounded and $\|\hat{z}_{\mathcal{S}^c}\|_{2,\infty} < 1$. The difference $\hat{\Delta}_{\mathcal{S}}$ can be obtained by solving equation (4.21)

$$\hat{\Delta}_{\mathcal{S}} = -\hat{\Sigma}^{-1}_{\mathcal{S}\mathcal{S}}(X^T_{\mathcal{S}}\epsilon^*/n + \lambda_n\hat{z}_{\mathcal{S}} + \mathcal{R}_{\mathcal{S}}).$$

We can show that

$$\begin{aligned} \|\hat{\Delta}_{\mathcal{S}}\|_\infty &= \|\hat{\Sigma}^{-1}_{\mathcal{S}\mathcal{S}}(X^T_{\mathcal{S}}\epsilon^*/n + \lambda_n\hat{z}_{\mathcal{S}} + \mathcal{R}_{\mathcal{S}})\|_\infty \\ &\le \left|\left|\left|(\hat{\Sigma}_{\mathcal{S}\mathcal{S}})^{-1}\right|\right|\right|_\infty (\|X^T_{\mathcal{S}}\epsilon^*/n\|_\infty + \|\mathcal{R}_{\mathcal{S}}\|_\infty + \lambda_n) \\ &\le \sqrt{s}\left|\left|\left|(\hat{\Sigma}_{\mathcal{S}\mathcal{S}})^{-1}\right|\right|\right|_2 (\|X^T_{\mathcal{S}}\epsilon^*/n\|_{2,\infty} + \|\mathcal{R}_{\mathcal{S}}\|_\infty + \lambda_n). \end{aligned} \tag{4.22}$$

In addition,

$$\begin{aligned} \hat{z}_{\mathcal{S}^c} =& -\frac{1}{\lambda_n}(X^T_{\mathcal{S}^c}\epsilon^*/n + \mathcal{R}_{\mathcal{S}^c} - \hat{\Sigma}_{\mathcal{S}^c\mathcal{S}}\hat{\Sigma}^{-1}_{\mathcal{S}\mathcal{S}}(X^T_{\mathcal{S}}\epsilon^*/n + \lambda_n\hat{z}_{\mathcal{S}} + \mathcal{R}_{\mathcal{S}}))) \\ =& \hat{\Sigma}_{\mathcal{S}^c\mathcal{S}}\hat{\Sigma}^{-1}_{\mathcal{S}\mathcal{S}}\hat{z}_{\mathcal{S}} - \frac{1}{\lambda_n}(X^T_{\mathcal{S}^c}(I - X^T_{\mathcal{S}}\hat{\Sigma}^{-1}_{\mathcal{S}\mathcal{S}}X_{\mathcal{S}})X^T_{\mathcal{S}^c}\epsilon^*/n + \mathcal{R}_{\mathcal{S}^c} - \hat{\Sigma}_{\mathcal{S}^c\mathcal{S}}\hat{\Sigma}^{-1}_{\mathcal{S}\mathcal{S}}\mathcal{R}_{\mathcal{S}}). \end{aligned}$$

Furthermore, $\max_{p\in\mathcal{S}^c}\|\hat{z}^{(p)}\|_2$ can be bounded as follows,

$$\begin{aligned}\max_{p\in\mathcal{S}^c}\|\hat{z}^{(p)}\|_2 \leq & \max_{p\in\mathcal{S}^c}\{\|(\hat{\Sigma}_{\mathcal{S}^c\mathcal{S}}\hat{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\hat{z}_{\mathcal{S}})^{(p)}\|_2+\frac{1}{\lambda_n}(\|(X_{\mathcal{S}^c}^T\Pi\epsilon^*/n)^{(p)}\|_2+\|\mathcal{R}_{\mathcal{S}^c}^{(p)}\|_2 \\ & +\|(\hat{\Sigma}_{\mathcal{S}^c\mathcal{S}}\hat{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\mathcal{R}_{\mathcal{S}})^{(p)}\|_2)\},\end{aligned}$$

with an orthogonal projection matrix $\Pi = I - X_{\mathcal{S}}^T\hat{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}X_{\mathcal{S}}$.

Next, we need to analyze the upper bound of each element in both $\hat{\Delta}_{\mathcal{S}}$ and $\hat{z}_{\mathcal{S}^c}$. From Lemma 4.1, we have

$$\|\frac{X^T\epsilon^*}{n}\|_{2,\infty} \leq \frac{\xi\lambda_n}{4}, \tag{4.23}$$

with probability at least $1 - C_1\exp\{-C_2\log(p_n)\}$.

By applying (4.50) from Lemma 4.6, we can show that with probability $1 - c_1\exp\{-c_2 s\}$ for some constants c_1, c_2, and c_3,

$$\alpha_u{}^{-1} - c_3\sqrt{\frac{s}{n}} \leq \left|\left|\left|\Sigma_{\mathcal{S}\mathcal{S}}^{-1}\right|\right|\right|_2 - c_3\sqrt{\frac{s}{n}} \leq \left|\left|\left|\hat{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\right|\right|\right|_2 \leq \left|\left|\left|\Sigma_{\mathcal{S}\mathcal{S}}^{-1}\right|\right|\right|_2 + c_3\sqrt{\frac{s}{n}} \leq \alpha_l{}^{-1} + c_3\sqrt{\frac{s}{n}}. \tag{4.24}$$

Thus, the sub-matrix $\hat{\Sigma}_{\mathcal{S}\mathcal{S}}$ is invertible with high probability, and the loss function is strongly convex on the subspace $\mathcal{S}$.

In addition, by fixing p and k, we can apply Cauchy-Schwarz inequality to each element in the vector $\mathcal{R}$ as follows,

$$\begin{aligned}\mathcal{R}_{kp} &= -\int_0^1 \frac{1}{n}\sum_{i=1}^n (x_{ki}x_{ki}^T\hat{\Delta}_k)_p 1(|\tilde{u}_{ki}| > \tau_k)dt \\ &\leq \int_0^1 \{\frac{1}{n}\sum_{i=1}^n((x_{ki}x_{ki}^T\hat{\Delta}_k)_p)^2\}^{1/2}\{\frac{1}{n}\sum_{i=1}^n 1(|\tilde{u}_{ki}| > \tau_k)\}^{1/2}dt.\end{aligned}$$

Since x_{ki}'s are independent sub-Gaussian variables, we can show that with probability

$1 - 2\exp\{-\log(p_n)\}$,

$$\frac{1}{n}\sum_{i=1}^{n}(v^T x_{ki} x_{ki}^T u)^2 \leq \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}((x_{ki}^T v)^4 + (x_{ki}^T u)^4) \lesssim A_0^4,$$

for any unit vector v and u. We can apply Hoeffding's inequality to obtain that with probability at least $1 - \exp\{-2nx^2\}$,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} 1(|\tilde{u}_{ki}| > \tau_k) \leq & \frac{1}{n}\sum_{i=1}^{n} P(|\tilde{u}_{ki}| > \tau_k) + x \\ \leq & \frac{1}{n}\sum_{i=1}^{n} P(|\tilde{u}_{ki}^*| > \frac{\tau_k}{2}) + P(|t x_{ki}^T \Delta_k| > \frac{\tau_k}{2}) + x \\ \leq & v_k\big(\frac{2}{\tau_k}\big)^{1+\omega} + c\big(\frac{A_0 t}{\tau_k}\big)^{1+\omega}\|\Delta_k\|_2^{1+\omega} + x.\end{aligned}$$

Based on Theorem 4.1, $\|\Delta\|_2 \lesssim \sqrt{s}\lambda_n \leq r$, and $\tau_k \lesssim (n/\log(p_n))^{\max\{1/2, 1/(1+\omega)\}}$. For $n \gtrsim s^2\log(p_n)$, we can take $x = \sqrt{\log(p_n)/n}$ to show that with probability at least $1 - \exp\{-2\log(s)\}$,

$$\frac{1}{n}\sum_{i=1}^{n} 1(|\tilde{u}_{ki}| > \tau_k) \lesssim \big(\frac{\log(p_n)}{n}\big)^{\max\{(1+\omega)/2, 1\}} + \big(\frac{\log(p_n)}{n}\big)^{1/2} \lesssim \sqrt{\frac{\log(p_n)}{n}}.$$

Thus, the ℓ_∞ norm of the component $\mathcal{R}$ can be bounded with probability at least $1 - \exp\{-2\log(p_n)\}$,

$$\begin{aligned}\|\mathcal{R}\|_\infty \leq & \max_{kp}\{\frac{1}{n}\sum_{i=1}^{n}((x_{ki}x_{ki}^T\hat{\Delta}_k)_p)^2\}^{1/2}\int_0^1\{\frac{1}{n}\sum_{i=1}^{n} 1(|\tilde{u}_{ki}| > \tau_k)\}^{1/2}dt \\ \lesssim & \big(\frac{s^2\log(p_n)}{n}\big)^{1/4}\lambda_n. \end{aligned} \tag{4.25}$$

Therefore, from results above in (4.23), (4.24), and (4.25), (4.22) can be bounded as

follows

$$\|\hat{\Delta}_{\mathcal{S}}\|_\infty \leq \frac{\sqrt{s}}{\alpha_l}(\frac{\xi}{4}\lambda_n + \big(\frac{s^2\log(p_n)}{n}\big)^{1/4}\lambda_n + \lambda_n) = \frac{\sqrt{s}}{\alpha_l}(\frac{5\lambda_n}{4})$$

with probability $1 - 2\exp\{-\log(p_n)\}$. By taking $\min_{k,p\in\mathcal{S}} |\theta^*_{kp}| \geq \alpha_l^{-1}\sqrt{s}(5\lambda_n/4)$, we have

$$P(\|\hat{\Delta}_{\mathcal{S}}\|_\infty \leq \min_{k,p\in\mathcal{S}} |\theta^*_{kp}|) \geq 1 - 2\exp\{-\log(p_n)\},$$

which provides the sign consistency $\text{sign}(\hat{\theta}_{\mathcal{S},0}) = \text{sign}(\theta^*)$.

In the last part, we can apply (4.51) from Lemma 4.6 to obtain

$$\begin{aligned}
\max_{p\in\mathcal{S}^c} \|\hat{z}^{(p)}\|_2 \leq & \max_{p\in\mathcal{S}^c}\{\|(\hat{\Sigma}_{\mathcal{S}^c\mathcal{S}}\hat{\Sigma}^{-1}_{\mathcal{S}\mathcal{S}}u)^{(p)}\hat{z}^{(p)}_{\mathcal{S}}\|_2 + \frac{1}{\lambda_n}(\|\mathcal{L}(\theta^*)^{(p)}\|_2 \\
& + \|\mathcal{R}^{(p)}\|_2 + \|(\hat{\Sigma}_{\mathcal{S}^c\mathcal{S}}\hat{\Sigma}^{-1}_{\mathcal{S}\mathcal{S}}u)^{(p)}\|_2\|\mathcal{R}_{\mathcal{S}})^{(p)}\|_2)\} \\
\leq & (1 - \frac{\xi}{2}) + \frac{1}{\lambda_n}(\frac{\xi}{4}\lambda_n + \big(\frac{s^2\log(p_n)}{n}\big)^{1/4}\lambda_n) \\
< & (1 - \frac{\xi}{2}) + \frac{\xi}{4} + o(1) < 1 - \frac{\xi}{4}
\end{aligned} \tag{4.26}$$

with probability at least $1 - C_3\exp\{-C_4\min\{s, \log(p_n)\}\}$. □

The following Theorem 4.2 establishes that the robust multi-task learning method can recover the union support with probability tending to one.

Corollary 4.1. *Based on Assumptions 4.1 - 4.2, if the penalty parameter λ_n is chosen as (4.16) and $16R\tau_l\log(p_n)/n \leq \xi\lambda_n$, then the estimator $\hat{\theta} \in \mathbb{B}_r(\theta^*)$ is the unique optimal solution of the program (4.4) that has sign consistency with probability at least $1 - C_3\exp\{-C_4\min\{s, \log(p_n)\}\}$ for some constants C_3 and C_4.*

Proof. From Theorem 4.2, there exists stationary points $\hat{\theta}_{\mathcal{S},0}$ that is a local minimum and has sign consistency. Suppose there exists another stationary point $\tilde{\theta}$, such that $\|\tilde{\theta} - \theta^*\|_2 \leq r$.

For simplicity, let $\hat{\theta}_{\mathcal{S},0}$ in the proof of Theorem 4.2 be denoted by $\hat{\theta}$, and by Assumption 4.1,

$$n^{-1}(\nabla\mathcal{L}(\tilde{\theta}) - \nabla\mathcal{L}(\hat{\theta}))^T(\tilde{\theta} - \hat{\theta}) \geq \kappa_l\|\tilde{\theta} - \hat{\theta}\|_2^2 - \tau_l\frac{\log(p_n)}{n}\|\tilde{\theta} - \hat{\theta}\|_{2,1}^2.$$

The stationary point $\tilde{\theta}$ of program (4.4) can satisfy

$$(n^{-1}\nabla\mathcal{L}(\tilde{\theta}) + \lambda_n\tilde{z})^T(\tilde{\theta} - \hat{\theta}) \leq 0,$$

with $\tilde{z}$ denoting the subdifferential of $\|\tilde{\theta}\|_{2,1}$. In addition, $\hat{\theta}$ is an interior point satisfying

$$(n^{-1}\nabla\mathcal{L}(\hat{\theta}) + \lambda_n\hat{z})^T(\tilde{\theta} - \hat{\theta}) = 0.$$

Therefore, we obtain

$$\lambda_n(\hat{z} - \tilde{z})^T(\tilde{\theta} - \hat{\theta}) \geq \kappa_l\|\tilde{\theta} - \hat{\theta}\|_2^2 - \tau_l\frac{\log(p_n)}{n}\|\tilde{\theta} - \hat{\theta}\|_{2,1}^2. \tag{4.27}$$

Based on (4.26) in the proof of Theorem 4.2, we have $\|\hat{z}_{\mathcal{S}^c}\|_{2,\infty} \leq 1 - \xi/4$, such that $\|\hat{\theta}_{\mathcal{S}^c}\|_{2,1} = 0$, and then

$$\begin{aligned}\hat{z}^T(\tilde{\theta} - \hat{\theta}) =& \|\hat{z}_{\mathcal{S}}\|_{2,\infty}\|(\tilde{\theta} - \hat{\theta})_{\mathcal{S}}\|_{2,1} + \|\hat{z}_{\mathcal{S}^c}\|_{2,\infty}\|(\tilde{\theta} - \hat{\theta})_{\mathcal{S}^c}\|_{2,1} \\ \leq& \|(\tilde{\theta} - \hat{\theta})_{\mathcal{S}}\|_{2,1} + (1 - \frac{\xi}{4})\|(\tilde{\theta} - \hat{\theta})_{\mathcal{S}^c}\|_{2,1},\end{aligned}$$

and

$$\begin{aligned}-\tilde{z}^T(\tilde{\theta} - \hat{\theta}) = \tilde{z}^T\hat{\theta} - \|\tilde{\theta}\|_{2,1} \leq& \|\hat{\theta}\|_{2,1} - \|\tilde{\theta}\|_{2,1} = \|\hat{\theta}_{\mathcal{S}}\|_{2,1} - \|\tilde{\theta}_{\mathcal{S}}\|_{2,1} - \|\tilde{\theta}_{\mathcal{S}^c}\|_{2,1} \\ \leq& \|(\tilde{\theta} - \hat{\theta})_{\mathcal{S}}\|_{2,1} - \|(\tilde{\theta} - \hat{\theta})_{\mathcal{S}^c}\|_{2,1}.\end{aligned}$$

Thus,

$$\begin{aligned}\lambda_n(2\|(\tilde{\theta}-\hat{\theta})_{\mathcal{S}}\|_{2,1}-\frac{\xi}{4}\|(\tilde{\theta}-\hat{\theta})_{\mathcal{S}^c}\|_{2,1}) \geq & \kappa_l\|\tilde{\theta}-\hat{\theta}\|_2^2-\tau_l\frac{\log(p_n)}{n}\|\tilde{\theta}-\hat{\theta}\|_{2,1}^2 \\ \geq & -\tau_l\frac{\log(p_n)}{n}\|\tilde{\theta}-\hat{\theta}\|_{2,1}^2.\end{aligned}$$

With condition $16R\tau_l\log(p_n)/n \leq \xi\lambda_n$, we can show that

$$\tau_l\frac{\log(p_n)}{n}\|\tilde{\theta}-\hat{\theta}\|_{2,1} \leq \tau_l\frac{\log(p_n)}{n}(\|\tilde{\theta}\|_1+\|\hat{\theta}\|_1) \leq 2R\tau_l\frac{\log(p_n)}{n} \leq \frac{\xi}{8}\lambda_n,$$

which can imply the following inequalities

$$\begin{aligned}\lambda_n(2\|(\tilde{\theta}-\hat{\theta})_{\mathcal{S}}\|_{2,1}-\frac{\xi}{4}\|(\tilde{\theta}-\hat{\theta})_{\mathcal{S}^c}\|_{2,1}) \geq & -\frac{\xi}{8}\lambda_n\|\tilde{\theta}-\hat{\theta}\|_{2,1} \\ (2+\frac{\xi}{8})\|(\tilde{\theta}-\hat{\theta})_{\mathcal{S}}\|_{2,1} \geq & \frac{\xi}{8}\|(\tilde{\theta}-\hat{\theta})_{\mathcal{S}^c}\|_{2,1} \\ (\frac{16}{\xi}+1)\|(\tilde{\theta}-\hat{\theta})_{\mathcal{S}}\|_{2,1} \geq & \|(\tilde{\theta}-\hat{\theta})_{\mathcal{S}^c}\|_{2,1} \\ (\frac{16}{\xi}+2)\sqrt{s}\|\tilde{\theta}-\hat{\theta}\|_2 \geq & \|\tilde{\theta}-\hat{\theta}\|_{2,1}.\end{aligned}$$

Based on (4.27), we can show that

$$\lambda_n(\hat{z}-\tilde{z})^T(\tilde{\theta}-\hat{\theta}) \geq \big(\kappa_l-(\frac{16}{\xi}+2)^2\tau_l\frac{s\log(p_n)}{n}\big)\|\tilde{\theta}-\hat{\theta}\|_2^2.$$

Based on Assumption 4.2 that $n \gtrsim s^2\log(p_n)$, the right-hand side is bounded below by 0. In addition,

$$(\hat{z}-\tilde{z})^T(\tilde{\theta}-\hat{\theta}) = \hat{z}^T\tilde{\theta}-\|\hat{\theta}\|_{2,1}-\|\tilde{\theta}\|_{2,1}+\tilde{z}^T\hat{\theta} \leq \hat{z}^T\tilde{\theta}-\|\tilde{\theta}\|_{2,1} \leq 0,$$

with both $\tilde{z}^T\hat{\theta} \leq \|\tilde{z}\|_{2,\infty}\|\hat{\theta}\|_{2,1} \leq \|\hat{\theta}\|_{2,1}$ and $\hat{z}^T\tilde{\theta} \leq \|\hat{z}\|_{2,\infty}\|\tilde{\theta}\|_{2,1} \leq \|\tilde{\theta}\|_{2,1}$.

Therefore, since positive constants κ_l, ξ, and τ_l are not dependent on n, p_n, and s,

$$0 \leq \big(\kappa_l - (\frac{16}{\xi} + 2)^2 \tau_l \frac{s \log(p_n)}{n}\big) \|\tilde{\theta} - \hat{\theta}\|_2^2 = (\kappa_l + o(1)) \|\tilde{\theta} - \hat{\theta}\|_2^2 \leq 0$$

which implies $\tilde{\theta} = \hat{\theta}$.

□

4.3.3 Optimization Property

We propose to apply the composite gradient descent algorithm [Nesterov, 2013] to optimize the objective function (4.4). First, we can use an isotropic quadratic function as the majorizer to approximate the objective function $Q(\theta)$,

$$\frac{1}{n} Q(\theta|\theta^t) = \frac{1}{n} \mathcal{L}(\theta^t) + \frac{1}{n} \nabla \mathcal{L}(\theta^t)(\theta - \theta^t) + \frac{\gamma}{2} \|\theta - \theta^t\|_2^2 + \lambda_n \|\theta\|_{2,1},$$

and the quadratic coefficient γ is chosen as the largest eigenvalue of the Hessian matrix $n^{-1} \nabla^2 \mathcal{L}(\theta)$. The optimization procedure produces a sequence of updated points $\{\theta^t\}_{t=0}^{\infty}$,

$$\theta^{t+1} = \arg \min_{\|\theta - \theta^*\|_1 \leq R} \{\frac{1}{2} \|\theta - (\theta^t - \frac{\frac{1}{n} \nabla \mathcal{L}(\theta^t)}{\gamma})\|_2^2 + \frac{\lambda_n}{\gamma} \|\theta\|_{2,1}\}.$$

The thresholding operator $S_{\lambda_n/\eta}$ on the grouped parameters is defined as

$$S_{\lambda_n/\eta}(\theta^{(p)}) = (\|\theta^{(p)}\|_2 - \frac{\lambda_n}{\eta})_+ z^{(p)}.$$

We use $\theta^{(p),t}$ to represent the pth grouped parameter in the tth update. We can show the subsequent update as follow

$$\theta^{(p),t+1} = S_{\lambda_n/\eta}(\theta^{(p),t} - \frac{1}{n} \frac{\nabla \mathcal{L}(\theta^t)^{(p)}}{\eta}), \tag{4.28}$$

where the step size η can be equal or proportional to γ.

The following theorem shows that the optimization procedure enjoys a geometric rate of convergence.

Theorem 4.3. *Based on Assumptions 4.1 - 4.1, suppose the condition (4.9) holds and the initial point $\theta^0 \in \mathbb{B}_2(r)$, then the optimization procedure can yield $n^{-1}(Q(\theta^t) - Q(\hat{\theta})) \leq \delta^2$ with $\delta^2 = C_5\varepsilon^2 \log(p_n)/n$ for some C_5, and*

$$\|\theta^t - \hat{\theta}\|_2^2 \leq \kappa_l^{-1}(\delta^2 + 16\tau_l\delta^4 + 4\tau_l\frac{\log(p_n)}{n}\varepsilon^2), \text{ for } t \geq T = C_6\frac{\log(1/\delta^2)}{\log(1/\kappa)}$$

with some $\kappa \in (0,1)$ and some constant C_6, where $\varepsilon = 8\sqrt{s}\|\hat{\theta} - \theta^*\|_2$.

Proof. Our proof is analogous with Loh and Wainwright [2015]. The iterations t can be divided into different epochs $t \in [T_j, T_{j+1})$ with different tolerance levels, which can be denoted as $\{\bar{\eta}_j\}_{j=0}^{\infty}$. From Lemmas 4.8 and 4.9, we have

$$\frac{1}{n}Q(\theta^t) - \frac{1}{n}Q(\hat{\theta}) \leq \kappa^{t-T}(\frac{1}{n}Q(\theta^T) - \frac{1}{n}Q(\hat{\theta})) + \frac{\beta}{1-\kappa}(\varepsilon + \epsilon)^2,$$

where $\varepsilon = 8\sqrt{s}\|\hat{\theta} - \theta^*\|_2$ and $\epsilon = \min\{\lambda_n^{-1}2\bar{\eta}, R\}$. In the following iterations, for $t \in [T_j, T_{j+1})$, we can set $n^{-1}Q(\theta^t) - n^{-1}Q(\hat{\theta}) \leq \bar{\eta}_j$ and let ϵ be replaced by $\epsilon_j = \min\{\lambda_n^{-1}2\bar{\eta}_j, R\}$. Therefore, we can show the first epoch $t \in [T_0, T_1)$ with $T_0 = 0$,

$$\frac{1}{n}Q(\theta^t) - \frac{1}{n}Q(\hat{\theta}) \leq \kappa^t(\frac{1}{n}Q(\theta^{T_0}) - \frac{1}{n}Q(\hat{\theta})) + \frac{4\beta}{1-\kappa}\max\{\varepsilon^2, \epsilon_0^2\}.$$

Here, we can set the next precision level as $\bar{\eta}_1 := 8\beta\max\{\varepsilon^2, \epsilon_0^2\}/(1-\kappa)$,

$$\frac{1}{n}Q(\theta^t) - \frac{1}{n}Q(\hat{\theta}) \leq \bar{\eta}_1 \leq \kappa^t\bar{\eta}_0 + \frac{4\beta}{1-\kappa}\max\{\varepsilon^2, \epsilon_0^2\}, \forall t \geq T_1,$$

with $T_1 = \lceil \log(2\bar{\eta}_0/\bar{\eta}_1)/\log(1/\kappa) \rceil$. In the following epochs, the value of $\bar{\eta}_j$ for $j \geq 1$ will

decrease as the contraction factor $\kappa < 1$, and ϵ_j also decreases, which provides the recursive relation

$$\frac{1}{n}Q(\theta^t) - \frac{1}{n}Q(\hat{\theta}) \leq \kappa^{t-T_j}\bar{\eta}_j + \frac{4\beta}{1-\kappa}\max\{\varepsilon^2, \epsilon_j^2\}, \forall t \geq T_j.$$

Through a similar setting, we can show that

$$\bar{\eta}_{j+1} := \frac{8\beta}{1-\kappa}\max\{\varepsilon^2, \epsilon_j^2\} \text{ and } T_{j+1} = \lceil \frac{\log(2\bar{\eta}_j/\bar{\eta}_{j+1})}{\log(1/\kappa)} \rceil + T_j.$$

For the last epoch, we can set the ultimate tolerance level $\bar{\eta} = \delta^2 = C_5\varepsilon^2\log(p_n)/n$ for some constant $C_5 > 0$, such that $n^{-1}Q(\theta^t) - n^{-1}Q(\hat{\theta}) \leq \delta^2$ for $t \geq T$. The total number of iterations can be calculated based on the results above as $T = C_6\log(1/\delta^2)/\log(1/\kappa)$.

Based on Lemma 4.7, if we set the initial point satisfies $\|\theta^0 - \hat{\theta}\|_2 \leq d$, any following updated point also sits in the neighborhood. We can apply Lemma 4.9 to show that the optimization can obtain the ℓ_2-bound of estimators θ^t

$$\begin{aligned}
\|\theta^T - \hat{\theta}\|_2^2 \leq & (\kappa_l - 32\tau_l\frac{s\log(p_n)}{n})^{-1}(\delta^2 + 2\tau_l\frac{\log(p_n)}{n}(\varepsilon+\epsilon)^2) \\
\leq & \kappa_l^{-1}(\delta^2 + 16\tau_l\frac{\log(p_n)}{n}\frac{\delta^4}{\lambda_n^2} + 4\tau_l\frac{\log(p_n)}{n}\epsilon^2) \\
\leq & \kappa_l^{-1}(\delta^2 + 16\tau_l\delta^4 + 4\tau_l\frac{\log(p_n)}{n}\varepsilon^2)
\end{aligned}$$

with $n \gtrsim s^2\log(p_n)$ and $\lambda_n \asymp (\log(p_n)/n)^{\min\{1/2,\omega/(1+\omega)\}}$. □

4.3.4 Robust Information Criterion

The joint feature selection can be based on the aggregation of the information from different tasks. Gao and Carroll [2017] proposed the pseudo Bayesian information criterion by aggregating the pseudo log-likelihoods from different tasks. Dai et al. [2020] proposed the

multiple quantile Bayesian information criterion and replaced the loss function with the pooled check function. Since the Huber loss function is a combination of squared loss and Least Absolute Deviation, we propose the robust version of the information criterion for the multi-task feature learning,

$$\text{robust-BIC} = 2\mathcal{L}(\hat{\theta}_{\mathcal{J}}) + d^*_{\mathcal{J}}\gamma_n. \tag{4.29}$$

The first component in (4.29) is used to measure the goodness-of-fit of the composite Huber loss function for a given union support $\mathcal{J}$ across all related tasks, and the second component is the penalty to control the complexity of the selected model. Following Gao and Carroll [2017], we set the penalty term $\gamma_n = c\log(p_n)$ with some constant c. The effective degrees of freedom $d^*_{\mathcal{J}}$ is defined as $\text{tr}(H^{-1}(\hat{\theta}_{\mathcal{J}})J(\hat{\theta}_{\mathcal{J}}))$. The formulation (4.29) is an extension of the robust information criterion applied to a single learning task with robust loss function [Tharmaratnam and Claeskens, 2013].

4.4 Simulation

In this section, simulation studies are implemented to examine the empirical performance of the robust multi-task feature learning method, which is compared with the multi-task feature learning method without robust loss function and the robust single-task learning method. We use the positive selection rates (PSR), false discovery rates (FDR), and squared ℓ_2 norm estimation error to measure the model performance. The positive selection rate is defined as

$$\text{PSR} = \frac{\sum_p 1(\hat{\theta}^{(p)} \neq \mathbf{0})1(\theta^{*(p)} \neq \mathbf{0})}{\sum_p 1(\theta^{*(p)} \neq \mathbf{0})},$$

which measures the proportion of true features that are accurately identified by the model. The false discovery rate (FDR) is used to demonstrate the proportion of unimportant features

selected by the model, which is defined as

$$\text{FDR} == \frac{\sum_p 1(\hat{\theta}^{(p)} \neq \mathbf{0}) 1(\theta^{*(p)} = \mathbf{0})}{\sum_p 1(\hat{\theta}^{(p)} \neq \mathbf{0})}.$$

Since we compare the multi-task feature learning with the single-task analysis, the squared ℓ_2 norm estimation error is set as $K^{-1}\|\hat{\theta} - \theta^*\|_2^2$.

The penalty parameters are set as $\lambda_n \asymp \sqrt{\log(p_n)/n}$, and we use the robust Bayesian information criterion to determine the value of the penalty parameter. The robustification parameters are updated based on the method proposed by Wang et al. [2021], such that $\tau_k^t = \hat{\sigma}_k^t \sqrt{n/\log(p_n)}/4$, and the algorithm iteratively estimates the standard deviation of the error terms $\hat{\sigma}_k^t = \sqrt{n^{-1}\sum_{i=1}^n (y_{ki} - x_{ki}^T \theta_k^t)^2}$.

4.4.1 Multiple data sets with 10% outliers

In the first simulation, we consider the case that a small portion of data is generated with outliers or heavy-tailed errors for all tasks. The response variables are simulated as follows,

$$y_{ki} = x_{ki}^T \theta_k^* + u_{ki}^*, \text{ and } u_{ki}^* = \zeta_{ki}\varepsilon_{ki}. \tag{4.30}$$

For each task, we generate $n = 500$ observation, and the dimensions of the parameter are set as $p_n = 500$, 1000, and 1500. The true regression parameters are simulated from uniform distribution Unif$(0.05, 0.5)$, and the size of true support in each task is $\lceil p^{1/2} \rceil$.

The predictors for the true model $\mathcal{S}$ are generated from multivariate normal distribution MVN$(\mathbf{0}, \mathbf{\Sigma_k})$, and the $s \times s$ covariance matrix Σ_k is designed with variances $\sigma_p^2 = 2$ and correlation $\rho_{pq} = 0.3$ for any $p, q \in \mathcal{S}$. Other unimportant predictors are generated from a different zero-mean multivariate normal distribution, which has variances $\sigma_p^2 = 0.2$ and correlations $\rho_{pq} = 0.05$ for any $p, q \in \mathcal{S}^c$.

The error terms across all tasks are correlated. We consider two cases, where the errors $(\varepsilon_{1i}, \varepsilon_{2i}, \varepsilon_{3i}, \varepsilon_{4i})$ are generated from

1. Gaussian Mixture Error: Mixture of multivariate Gaussian distribution $0.5\text{MVN}(1, \Sigma) + 0.5\text{MVN}(-1, \Sigma)$;

2. Heavy-tailed Error: Multivariate t distribution with 10 degrees of freedom $t_{10}(\Sigma)$.

For both cases, the 4×4 covariance matrix Σ is set with the variances equal to one and the correlations generated from $\text{Unif}(0.4, 0.65)$. In addition, we randomly introduce 50 outliers among all error terms. For each observation, the error term ε_{ki} multiplies with the scalar ζ_{ki}. Let the vector $\zeta_k = (\zeta_{k1}, \zeta_{k2}, \cdots, \zeta_{kn})^T$. There are 50 randomly selected elements of ζ_k set equal to $\sqrt{n}$ and the remaining elements set to be one.

Table 4.1: Comparison of the robust multi-task learning (RMTL) compared with the multi-task learning without a robust loss (MTL) and the robust single-task analysis (STA) for multiple data sets with 10% outliers. The standard errors (%) are provided in parentheses.

	$p_n = 500$			$p_n = 1000$			$p_n = 1500$		
Model	PSR	FDR	ℓ_2 Err	PSR	FDR	ℓ_2 Err	PSR	FDR	ℓ_2 Err
Simulation I: Gaussian Mixture Error									
RMTL	100 (0)	0 (1)	0.14	100 (0)	0 (1)	0.22	100 (0)	0 (0)	0.30
MTL	98 (3)	0 (1)	0.66	99 (2)	0 (1)	0.83	98 (2)	0 (1)	1.45
STA1	98 (3)	0 (1)	0.14	100 (1)	1 (3)	0.20	100 (1)	7 (12)	0.28
STA2	99 (2)	0 (1)	0.14	100 (1)	2 (3)	0.21	100 (1)	7 (11)	0.28
STA3	99 (2)	0 (1)	0.14	100 (1)	1 (3)	0.21	100 (1)	8 (12)	0.27
STA4	99 (2)	0 (2)	0.14	100 (1)	1 (3)	0.21	100 (1)	7 (11)	0.26
Simulation II: Heavy-tailed Error									
RMTL	100 (0)	0 (2)	0.12	100 (0)	1 (2)	0.20	100 (0)	1 (4)	0.29
MTL	98 (3)	0 (1)	0.66	99 (2)	0 (1)	0.83	99 (1)	0 (0)	0.93
STA1	100 (1)	0 (1)	0.11	100 (0)	2 (9)	0.17	100 (0)	5 (10)	0.26
STA2	99 (2)	0 (1)	0.11	100 (0)	2 (8)	0.18	100 (0)	5 (11)	0.25
STA3	100 (1)	0 (1)	0.10	100 (0)	1 (3)	0.18	100 (0)	5 (12)	0.25
STA4	100 (1)	0 (0)	0.11	100 (0)	2 (8)	0.18	100 (0)	5 (11)	0.25

We conduct 50 independent simulations under each simulation setting. From Table 4.1, we

observe that the proposed model yields high positive selection rates, low false discovery rates, and relatively smaller estimation error $K^{-1}\|\hat{\theta} - \theta^*\|_2^2$. In contrast, the multi-task feature learning method without robust loss is largely affected by the outliers and yields bigger estimation errors. Compared to the robust multi-task method, the robust single-task method yields higher FDR.

4.4.2 Heteroscedastic regression

In the second example, the true regression coefficients and predictors are generated by the same process as the previous example. The random errors are also simulated from asymmetric and heavy-tailed random errors. The data-generating process uses the heteroscedastic model to simulate the response variables,

$$y_{ki} = x_{ki}^T \theta_k^* + u_{ki}^*, \text{ and } u_{ki}^* = \zeta_{ki}\varepsilon_{ki}, \tag{4.31}$$

where the constant $\zeta_{ki} = \|\theta_k^*\|_2^{-1}(x_{ki}^T\theta_k^*)$ can control the noise level of error component u_{ki}^*.

The results in Table 4.2 are summarized from 50 independent simulations. We observe that the overall performance of the multi-task learning is better than that of the single-task analysis. In comparison with the multi-task feature learning method without robust loss, the proposed method produces smaller ℓ_2-norm estimation errors.

4.5 Data Analysis

In this section, we apply the robust multi-task feature learning method to analyze multiple community health status indicators (CHSI), which were collected across different counties of the U.S. in 2010 [U.S. Department of Health and Human Services]. There are 428 observations in the data sets. Four response variables of interest were measured to reflect the community

Table 4.2: Comparison of the robust multi-task learning (RMTL) compared with the multi-task learning without a robust loss (MTL) and the robust single-task analysis (STA) for multiple heteroscedastic regression. The standard errors (%) are provided in parentheses.

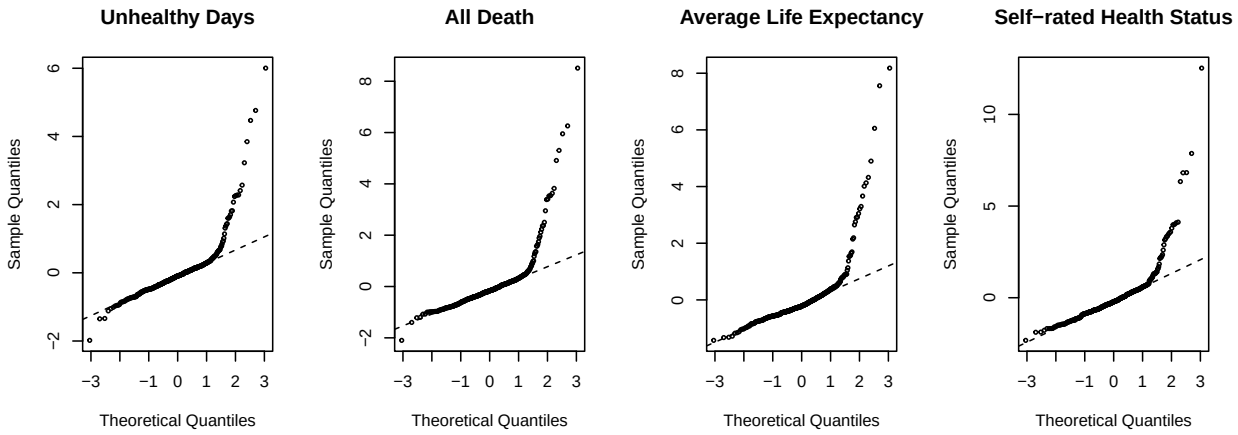
	$p_n = 500$			$p_n = 1000$			$p_n = 1500$		
Model	PSR	FDR	ℓ_2 Err	PSR	FDR	ℓ_2 Err	PSR	FDR	ℓ_2 Err
Simulation I: Asymmetric Error									
RMTL	100 (0)	0 (0)	0.29	100 (0)	0 (1)	0.42	100 (1)	0 (0)	0.54
MTL	100 (1)	0 (0)	0.32	100 (1)	0 (0)	0.53	99 (1)	0 (1)	0.69
STA1	97 (4)	0 (1)	0.33	98 (2)	2 (5)	0.47	99 (2)	5 (10)	0.56
STA2	97 (4)	0 (1)	0.31	99 (2)	2 (4)	0.45	99 (2)	5 (9)	0.56
STA3	97 (3)	0 (1)	0.32	99 (2)	2 (4)	0.46	99 (2)	5 (9)	0.56
STA4	98 (3)	0 (1)	0.31	98 (3)	2 (2)	0.44	99 (2)	6 (10)	0.55
Simulation II: Heavy-tailed Error									
RMTL	100 (0)	0 (0)	0.23	100 (0)	0 (2)	0.34	100 (1)	0 (0)	0.47
MTL	100 (0)	0 (1)	0.25	100 (0)	0 (0)	0.39	100 (0)	0 (0)	0.52
STA1	99 (2)	0 (0)	0.23	100 (1)	2 (4)	0.34	100 (0)	11 (15)	0.46
STA2	100 (2)	0 (1)	0.23	100 (0)	2 (4)	0.34	100 (0)	10 (13)	0.45
STA3	99 (1)	0 (1)	0.22	100 (0)	2 (3)	0.34	100 (0)	10 (13)	0.45
STA4	100 (1)	0 (0)	0.22	100 (1)	2 (5)	0.34	100 (0)	10 (13)	0.45

health status, including the average number of unhealthy days, the death counts, the average life expectancy, and the self-rated health status. By jointly analyzing the four response variables, we select important features from 70 candidate predictors, which include overall demographic information, counts of different diseases, different causes of death, environmental conditions, and health-related risk factors.

To show the performance of the robust regularization method, we randomly select 10% samples and introduce some large outliers. These contaminated samples are the original response variables added with additional terms $\varepsilon_{ki} = c_{ki}|u_{ki}|$ with $c_{ki} \sim$Unif$(0, \sqrt{n})$ and $u_{ki} \sim N(0, 1)$. We model each response variable with all predictors by the ordinary regression model, and the QQ plot in Figure 4.1 shows that the standardized residuals have a heavy right tail and skewed distribution.

We conduct a five-fold cross-validation to examine the model prediction accuracy through

Figure 4.1: QQ plots of the standardized residuals for each regression model.



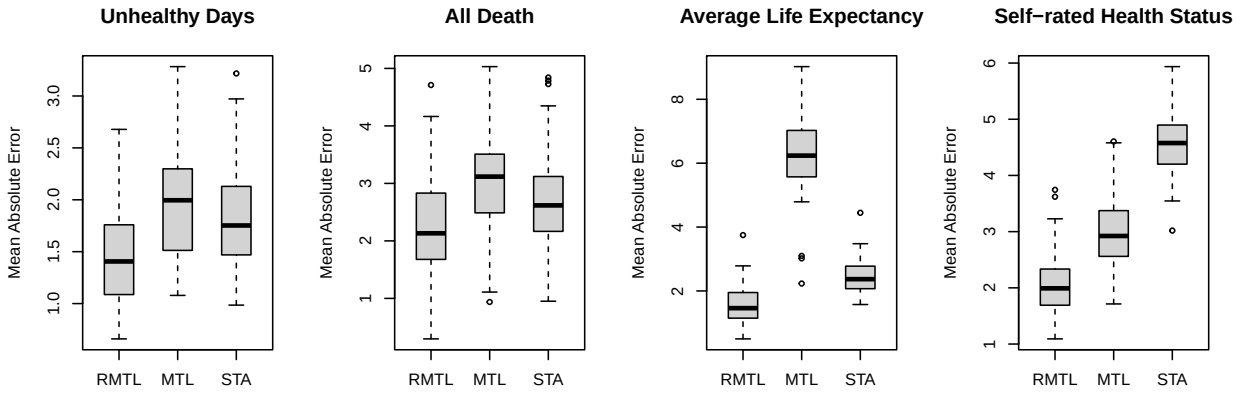
100 independent replications. The training set containing 80% of randomly selected observations is applied to the robust multi-task feature learning. The robustification parameters are chosen based on the adaptive method [Sun et al., 2020, Wang et al., 2021]. We use different values of the penalty parameters to conduct joint feature selection. The robust Bayesian information criterion is evaluated with the selected features, and we choose the joint model with the smallest value of the robust Bayesian information criterion. Then, we fit the adaptive Huber regression on the remaining 20% validation data sets with selected features and use the mean absolute error (MAE) to show the prediction accuracy, where the mean absolute error (MAE) is defined as

$$\mathrm{MAE}(\hat{\theta}_k) = \frac{1}{n}\sum_{i=1}^{n} |y_{ki}^{\mathrm{test}} - x_{ki}^{\mathrm{test}}\hat{\theta}_k|.$$

We use y_{ki}^{test} and x_{ki}^{test} to denote the observations of the response and covariates in the validation data for each task, and $\hat{\theta}_k$ is the estimated regression coefficients for the selected features in the validation data.

From Figure 4.2, the results demonstrate that the proposed method provides much smaller

Figure 4.2: The mean absolute error (MAE) of validation data sets for the community health status based on the robust multi-task learning (RMTL), compared with the multi-task learning without a robust loss (MTL) and the robust single-task analysis (STA).



mean absolute errors (MAE) than the other two comparison methods. The multi-task feature learning method without robust loss is very sensitive to outlier contamination. It has the highest mean absolute errors than the other two methods for three out of the four response variables. For the last response variable, the mean absolute errors from the single-task learning are the highest. We can infer that the data integration process can effectively enhance the model fitting and reduce estimation errors.

4.6 Technical Lemmas

This section provides some technical Lemmas used in the proofs of Lemmas and Theorem in Section 4.3. For each task, we can show the distribution property of the score function, which is similar to the large deviation bound in Sun et al. [2020].

Lemma 4.2. *Based on Assumption 4.1, let the score function of the kth task be denoted by*

$$U_n(\theta_{kp}) = \sum_{i=1}^{n} U_i(\theta_{kp}) = \sum_{i=1}^{n} \frac{\partial \ell_{ki}(\theta_k; y_{ki}, x_{ki})}{\partial \theta_{kp}},$$

for any $p = 1, 2, \cdots, p_n$. *We can show that*

$$\sum_{i=1}^{n} E([U_i(\theta_{kp})]^2) \leq 2\tau_k^{\max\{1-\omega,0\}} A_0^2 v_k n, \tag{4.32}$$

and for any $m \geq 3$

$$\sum_{i=1}^{n} E([U_i(\theta_{kp})]^m) \leq m\Gamma(\frac{m}{2})\tau_k^{\max\{1-\omega,0\}} A_0^2 v_k n(\tau_k A_0)^{m-2}. \tag{4.33}$$

Furthermore, for some $x > 0$

$$P(|\frac{1}{n}U_n(\theta_{kp})| \geq \nu_k\sqrt{2x} + \alpha_k x) \leq 2\exp\{-x\}, \tag{4.34}$$

where

$$\nu_k = 2\tau_k^{\max\{(1-\omega)/2,0\}} A_0(\frac{v_k}{n})^{1/2}, \text{ and } \alpha_k = \frac{\tau_k A_0}{n}.$$

Proof. Define new function $\psi_\tau(u) = \text{sign}(u)\min(|u|, \tau)$ with $E(u) = 0$ and $E(|u|^{1+\omega}) = v < \infty$. We first analyze the cases for $\omega \in (0, 1)$. We can derive the expectation of ψ_τ as follows,

$$\begin{aligned}
E(\psi_\tau(u)) =& E(\text{sign}(u)\min(|u|, \tau)) \\
=& E(\text{sign}(u)\min(|u|, \tau)) - E(u) \\
=& - E(\text{sign}(u)\max(0, |u| - \tau)) \\
=& - E(\text{sign}(u)(|u| - \tau)1(|u| > \tau)),
\end{aligned}$$

which can imply that

$$|E(\psi_\tau(\theta^*;u))| \leq E((|u|-\tau)\times\big(\frac{|u|}{\tau}\big)^\omega 1(|u|>\tau)) \leq v\tau^{-\omega}.$$

This result shows that the expectation of function $\psi_\tau(u)$ is bounded, and the boundedness is dependent on the parameters τ and v. We can apply this result to the absolute expectation of individual score function, such that

$$\begin{aligned}|E(U_i(\theta^*_{kp}))| =&|E(\text{sign}(u^*_{ki})\min(|u^*_{ki}|,\tau_k)x_{kpi})|\\ =&|E(E(\text{sign}(u^*_{ki})\min(|u^*_{ki}|,\tau_k)|x_{kpi})x_{kpi})|\\ <&v_{ki}\tau_k^{-\omega}A_0.\end{aligned}$$

In addition, since we assume x_{ki}'s are zero-mean Sub-Gaussian variables, $E(U_i(\theta_{kp}))=0$.

Next, we consider higher moments of function $\psi_\tau(u)$ for $m\geq 2$,

$$\begin{aligned}|E(\psi_\tau(u)^m)| &= E(\min(|u|^m,\tau^m))\\ &= E(|u|^m 1(|u|\leq\tau)) + E(\tau^m 1(|u|>\tau))\\ &\leq E(|u|^m\big(\frac{\tau}{|u|}\big)^{m-1-\omega}1(|u|\leq\tau)) + E(\tau^m\big(\frac{|u|}{\tau}\big)^{1+\omega}1(|u|>\tau))\\ &\leq E(|u|^{1+\omega}\tau^{m-1-\omega}1(|u|\leq\tau)) + E(|u|^{1+\omega}\tau^{m-1-\omega}1(|u|>\tau))\\ &\leq v\tau^{m-1-\omega}.\end{aligned}$$

Based on Assumption 4.1, we can show that $E((x_{ki}^T u)^m)\leq m\Gamma(m/2)A_0^m$. Thus, the mth moment of the individual score function can be given by

$$E[(U_i(\theta^*_{kp}))^m] = E[(\text{sign}(u^*_{ki})\min(|u^*_{ki}|,\tau_k)x_{kpi})^m] \leq m\Gamma(\frac{m}{2})\tau_k^{m-1-\omega}A_0^m v_{ki}.$$

For cases $\omega \geq 1$, $E(|u|^2) \leq v < \infty$. Therefore, we can derive that

$$\begin{aligned}
|E(\psi_\tau(u)^m)| =& E(\min(|u|^m, \tau^m)) \\
=& E(|u|^m 1(|u| \leq \tau)) + E(\tau^m 1(|u| > \tau)) \\
\leq& E(|u|^m \big(\frac{\tau}{|u|}\big)^{m-2} 1(|u| \leq \tau)) + E(\tau^m \big(\frac{|u|}{\tau}\big)^2 1(|u| > \tau)) \\
\leq& v\tau^{m-2}.
\end{aligned}$$

Thus, the mth moment of the individual score function can be given by

$$E[(U_i(\theta^*_{kp}))^m] = E[(\text{sign}(u^*_{ki}) \min(|u^*_{ki}|, \tau_k) x_{kpi})^m] \leq m\Gamma(\frac{m}{2})\tau_k^{m-2} A_0^m v_{ki},$$

and we combine all results to obtain Conditions (4.32) and (4.33) in the theorem.

Next, we can apply the Bernstein's inequality [Massart and Picard, 2007] to show that for some $x > 0$

$$P(|\frac{1}{n} U_n(\theta_{kp})| \geq \nu_k \sqrt{2x} + \alpha_k x) \leq 2\exp\{-x\}, \tag{4.35}$$

where

$$\nu_k = 2\tau_k^{\max\{(1-\omega)/2, 0\}} A_0 (\frac{v_k}{n})^{1/2}, \text{ and } \alpha_k = \frac{\tau_k A_0}{n}.$$

□

Assumption 4.1 is important to establish the estimation error bound and sign consistency. Thus, we need to show that this condition can be satisfied for the adaptive Huber regression under the high-dimensional random design.

Lemma 4.3. *Based on Assumptions 4.1-4.2, suppose the robustification parameters τ_k's*

satisfy

$$\tau_k \geq (4v_k)^{1/1+\omega} + \tau_0, \text{ for some } \tau_0^2 \geq 32A_0^2 r^2 \log(16A_0^2/\alpha_l).$$

If $R \lesssim \sqrt{s}$*, then, with probability at least* $1 - C^{'} \exp\{-C^{''} n\}$ *for some constant* C*,* $C^{'}$*, and* $C^{''}$*, the proposed loss function* $\mathcal{L}(\theta)$ *satisfies (4.6) with*

$$\kappa_l = \frac{\alpha_l}{8} \text{ and } \tau_l = \frac{CKA_0^2\tau_0^2}{\alpha_l r^2},$$

uniformly over all pairs (θ_1, θ_2)*, where*

$$(\theta_1, \theta_2) \in \{\theta_1, \theta_2 \in \mathbb{B}_r(\theta^*) : \|\theta_1 - \theta^*\|_1 \leq R, \|\theta_2 - \theta^*\|_1 \leq R, \\ \text{and } \frac{\|\theta_1 - \theta_2\|_{2,1}}{\|\theta_1 - \theta_2\|_2} \leq \frac{c\alpha_l r}{A_0\tau_0}\sqrt{\frac{n}{K\log(p_n)}}\}.$$

Proof. The proof is based on a similar approach in Fan et al. [2017], Loh and Wainwright [2015], Loh [2017], Sun et al. [2020]. We first define a new function $\psi_\tau(u) = \text{sign}(u)\min(|u|, \tau)$. The first-order Taylor error $\mathcal{T}(\theta_1, \theta_2)$ can be given by

$$\begin{aligned}\mathcal{T}(\theta_1, \theta_2) &= (n^{-1}\nabla\mathcal{L}(\theta_1) - n^{-1}\nabla\mathcal{L}(\theta_2))^T(\theta_1 - \theta_2) \\ &= \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n}(\psi_{\tau_k}(y_{ki} - x_{ki}^T\theta_{2k}) - \psi_{\tau_k}(y_{ki} - x_{ki}^T\theta_{1k}))x_{ki}^T(\theta_{1k} - \theta_{2k}),\end{aligned}$$

where θ_{1k} and θ_{2k} denote the subsets of the parameters θ_1 and θ_2 in the kth task.

We define an event A_{ki}:

$$A_{ki} := \{|u_{ki}^*| \leq \tau_k - \tau_0\} \cap \{|x_{ki}^T(\theta_{2k} - \theta_k^*)| \leq \frac{\tau_0}{2}\} \cap \{|x_{ki}^T(\theta_{1k} - \theta_{2k})| \leq \frac{\tau_0}{4r}\|\theta_1 - \theta_2\|_2\},$$

with some constant $\tau_0 < \tau_k$. Given the event A_{ki}, we can show that for $\|\theta_1 - \theta^*\|_2 \leq D$ and

$\|\theta_2 - \theta^*\|_2 \leq D$,

$$
\begin{aligned}
|y_{ki} - x_{ki}^T\theta_{2k}| &\leq |u_{ki}^*| + |x_{ki}^T(\theta_k^* - \theta_{2k})| \leq \tau_k - \tau_0 + \frac{\tau_0}{2} < \tau_k \\
|y_{ki} - x_{ki}^T\theta_{1k}| &\leq |u_{ki}^*| + |x_{ki}^T(\theta_k^* - \theta_{2k})| + |x_{ki}^T(\theta_{1k} - \theta_{2k})| \\
&\leq \tau_k - \tau_0 + \frac{\tau_0}{2} + \frac{\tau_0}{4r}(\|\theta_1 - \theta^*\|_2 + \|\theta_2 - \theta^*\|_2) \leq \tau_k.
\end{aligned}
$$

Since $\psi_\tau^{'}(u) = 1$ if $|u| \leq \tau$, the first-order Taylor error $\mathcal{T}(\theta_1, \theta_2)$ can be bounded below as follow,

$$
\mathcal{T}(\theta_1, \theta_2) \geq \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n}(x_{ki}^T(\theta_{1k} - \theta_{2k}))^2 1\{A_{ki}\}. \tag{4.36}
$$

This lower bound in (4.36) is modeled by a non-smooth function. We introduce the following truncation functions to deal with this problem,

$$
\phi_T(x) = \begin{cases} x^2 & \text{if } |x| \leq \frac{T}{2} \\ (T - |x|)^2 & \text{if } \frac{T}{2} < |x| \leq T \\ 0 & \text{if } |x| > T \end{cases} \quad \text{and} \quad \varphi_t(x) = \begin{cases} 1 - (\frac{x}{t})^2 & \text{if } |x| \leq t \\ 0 & \text{if } |x| > t, \end{cases}
$$

where the function $\phi_T(x)$ is T-Lipschitz and $\varphi_t(x)$ is $2/t$-Lipschitz. In addition, the proposed functions are bounded above as

$$
\phi_T(x) \leq x^2 1(|x| \leq T) \quad \text{and} \quad \varphi_t(x) \leq 1(|x| \leq t). \tag{4.37}
$$

Next, we define $\delta = \theta_1 - \theta_2$ and $\Delta = \theta_2 - \theta^*$. For the kth task, $\delta_k = \theta_{1k} - \theta_{2k}$ and $\Delta_k = \theta_{2k} - \theta_k^*$. As a result, the inequality (4.36) can be rearranged with the proposed

truncation functions,

$$\mathcal{T}(\theta_1,\theta_2) \geq g(\delta,\Delta) := \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n} \phi_{\tau_0\|\delta\|_2/4r}(x_{ki}^T\delta_k)\varphi_{\tau_0/2}(x_{ki}^T\Delta_k)\varphi_{\tau_k-\tau_0}(u_{ki}^*)$$
$$\geq E(g(\delta,\Delta)) - \sup_{\delta,\Delta\in\mathcal{B}(\gamma)} |g(\delta,\Delta) - E(g(\delta,\Delta))|, \tag{4.38}$$

where the restricted set is defined as follows,

$$\mathcal{B}(\gamma) := \{(\delta,\Delta) : \|\Delta\|_2 \leq r, \|\Delta\|_1 \leq R, \text{ and } \frac{\|\delta\|_{2,1}}{\|\delta\|_2} \leq \gamma\}$$
$$\text{for } 1 \leq \gamma \leq \frac{c\alpha_l r}{A_0\tau_0}\sqrt{\frac{n}{K\log(p_n)}}.$$

We can first analyze $E(g(\delta,\Delta))$ and obtain the lower bound by applying the properties (4.37) as follows,

$$E(g(\delta,\Delta)) \geq \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n}\{E((x_{ki}^T\delta_k)^2) - E((x_{ki}^T\delta_k)^2 1\{|u_{ki}^*| > \tau_k - \tau_0\})$$
$$-E((x_{ki}^T\delta_k)^2 1\{|x_{ki}^T\Delta_k| > \frac{\tau_0}{2}\}) - E((x_{ki}^T\delta_k)^2 1\{|x_{ki}^T\delta_k| > \frac{\tau_0}{4r}\|\delta\|_2\})\}.$$

Based on Assumption 4.1, we have $E((x_{ki}^T\delta_k)^2) = \delta_k^T\Sigma_k\delta_k \geq \alpha_l$. For the remaining components, we can apply conditional expectation to show that

$$\frac{1}{n}\sum_{i=1}^{n} E((x_{ki}^T\delta_k)^2 1\{|u_{ki}^*| > \tau_k - \tau_0\}) = \frac{1}{n}\sum_{i=1}^{n} E((x_{ki}^T\delta_k)^2)P(|u_{ki}^*| > \tau_k - \tau_0)$$
$$\leq v_k(\tau_k - \tau_0)^{-1-\omega}\delta_k^T\Sigma_k\delta_k,$$

and we can apply Cauchy-Schwarz inequality as follows,

$$\frac{1}{n}\sum_{i=1}^{n} E((x_{ki}^T\delta_k)^2 1\{|x_{ki}^T\Delta_k| > \frac{\tau_0}{2}\}) \leq \frac{1}{n}\sum_{i=1}^{n}\sqrt{E((x_{ki}^T\delta_k)^4)}\sqrt{P(|x_{ki}^T\Delta_k| > \frac{\tau_0}{2})}$$

$$\leq \big(4A_0^4\|\delta_k\|_2^4\big)^{1/2}\big(\exp\{-\frac{\tau_0^2}{4A_0^2\|\Delta_k\|_2^2}\}\big)^{1/2}$$
$$\leq 2A_0^2\|\delta_k\|_2^2\exp\{-\frac{\tau_0^2}{8A_0^2\|\Delta_k\|_2^2}\},$$

and

$$\frac{1}{n}\sum_{i=1}^n E((x_{ki}^T\delta_k)^2 1\{|x_{ki}^T\delta_k| > \frac{\tau_0}{4r}\|\delta\|_2\}) \leq \frac{1}{n}\sum_{i=1}^n \sqrt{E((x_{ki}^T\delta_k)^4)}\sqrt{P(|x_{ki}^T\delta_k| > \frac{\tau_0}{4r}\|\delta\|_2)}$$
$$\leq \big(4A_0^4\|\delta_k\|_2^4\big)^{1/2}\big(\exp\{-\frac{\tau_0^2\|\delta\|_2^2}{16A_0^2r^2\|\delta_k\|_2^2}\}\big)^{1/2}$$
$$\leq 2A_0^2\|\delta_k\|_2^2\exp\{-\frac{\tau_0^2}{32A_0^2r^2}\}.$$

Combining all the inequalities together, we can show that for any unit vector u,

$$E(\frac{g(\delta,\Delta)}{\|\delta\|_2^2}) \geq \inf_k\{u^T\Sigma_k u(1-\frac{v_k}{(\tau_k-\tau_0)^{1+\omega}})$$
$$-2A_0^2(\exp\{-\frac{\tau_0^2}{8A_0^2\|\Delta_k\|_2^2}\}+\exp\{-\frac{\tau_0^2}{32A_0^2r^2}\})\}.$$

We set $\tau_k \geq (4v_k)^{1+\omega}+\tau_0$ and $\tau_0^2 \geq 32A_0^2r^2\log(16A_0^2/\alpha_l)$, and then the expectation $E(g(\delta,\Delta)/\|\delta\|_2^2)$ is further bounded below by

$$E(\frac{g(\delta,\Delta)}{\|\delta\|_2^2}) \geq \alpha_l/2.$$

Next, let the second component in (4.38) be denoted by the random variable $\mathcal{G}(\delta,\Delta)$:

$$\mathcal{G}(\delta,\Delta) := \sup_{(\delta,\Delta)\in\mathcal{B}(\gamma)}\frac{|g(\delta,\Delta)-E(g(\delta,\Delta))|}{\|\delta\|_2^2}.$$

We first introduce a new random variable $\mathcal{Z}(\delta, \Delta)$ as

$$\mathcal{Z}(\delta, \Delta) := \frac{1}{n} \sum_{k=1}^{K} \sum_{i=1}^{n} \frac{y_{ki}}{\|\delta\|_2^2} \phi_{\tau_0 \|\delta\|_2 / 4r}(x_{ki}^T \delta_k) \varphi_{\tau_0/2}(x_{ki}^T \Delta_k) \varphi_{\tau_k - \tau_0}(u_{ki}^*),$$

and y_{ki}'s are i.i.d standard normal random variables, which are independent of x_{ki} and u_{ki}^* across all K tasks. Therefore, given x_{ki}'s and u_{ki}^*'s, $\{\mathcal{Z}(\delta, \Delta)\}$ is a conditional Gaussian process. We apply the inequalities of Gaussian complexity, and Rademacher complexity in Lemmas 12 and 13 from Loh and Wainwright [2015] to the expectation $E(\mathcal{G}(\delta, \Delta))$, which provides that for any $(\delta, \Delta) \in \mathcal{B}(\gamma)$,

$$E(\mathcal{G}(\delta, \Delta)) \leq \sqrt{2\pi} E(\sup_{(\delta, \Delta) \in \mathcal{B}(\gamma)} |\mathcal{Z}(\delta, \Delta)|). \tag{4.39}$$

Based on the results from Ledoux and Talagrand. [1991] and Loh [2017], the right hand side component in (4.39) can be bounded above as follows,

$$E(\sup_{(\delta, \Delta) \in \mathcal{B}(\gamma)} |\mathcal{Z}(\delta, \Delta)|) \leq E(|\mathcal{Z}(\delta', \Delta')|) + 2E(\sup_{(\delta, \Delta) \in \mathcal{B}(\gamma)} \mathcal{Z}(\delta, \Delta)), \tag{4.40}$$

with distinct pairs (δ, Δ) and $(\delta', \Delta') \in \mathcal{B}(\gamma)$. Therefore, we can combine (4.39) and (4.40) to obtain that

$$E(\mathcal{G}(\delta, \Delta)) \leq \sqrt{2\pi} E(|\mathcal{Z}(\delta', \Delta')|) + 2\sqrt{2\pi} E(\sup_{(\delta, \Delta) \in \mathcal{B}(\gamma)} \mathcal{Z}(\delta, \Delta)).$$

Let the conditional expectation given x_{ki}'s and u_{ki}^* be denoted by E^*. We can show that since y_{ki}'s are i.i.d. standard normal variables, the conditional expectation of $|\mathcal{Z}(\delta', \Delta')|$ can be bounded as follows,

$$E^*(|\mathcal{Z}(\delta', \Delta')|) \leq \left\{ \frac{2}{\pi \|\delta'\|_2^4} \frac{1}{n^2} \sum_{k=1}^{K} \sum_{i=1}^{n} \phi^2_{\tau_0 \|\delta_k'\|_2 / 4r}(x_{ki}^T \delta_k') \varphi^2_{\tau_0/2}(x_{ki}^T \Delta_k') \varphi^2_{\tau_k - \tau_0}(u_{ki}^*) \}) \right\}^{1/2},$$

and this inequality can be obtained based on $E(|z|) \leq \sqrt{2\text{var}(z)/\pi}$ for any zero-mean normal variable z. Furthermore, the expectation of $|\mathcal{Z}(\delta^{'}, \Delta^{'})|$ can be derived based on the properties (4.37), such that

$$\begin{aligned}
E(|\mathcal{Z}(\delta^{'}, \Delta^{'})|) \leq & E\big(\{\frac{2}{\pi\|\delta^{'}\|_2^4}\frac{1}{n^2}\sum_{k=1}^{K}\sum_{i=1}^{n}\phi^2_{\tau_0\|\delta^{'}\|_2/4r}(x_{ki}^T\delta_k^{'})\}^{1/2}\big) \\
\leq & \frac{1}{n\|\delta^{'}\|_2^2}\big\{\frac{2}{\pi}E(\sum_{k=1}^{K}\sum_{i=1}^{n}\phi^2_{\tau_0\|\delta_k^{'}\|_2/4r}(x_{ki}^T\delta_k^{'}))\big\}^{1/2} \\
\leq & \frac{1}{n\|\delta^{'}\|_2^2}\big\{\frac{2}{\pi}E(\sum_{k=1}^{K}\sum_{i=1}^{n}(x_{ki}^T\delta_k^{'})^4)\big\}^{1/2} \leq \sqrt{\frac{8}{n\pi}}A_0^2.
\end{aligned}$$

In order to find the upper bound for the component $E(\sup_{(\delta,\Delta)\in\mathcal{B}(\gamma)}\mathcal{Z}(\delta,\Delta))$, we can apply Sudakov–Fernique Theorem [Ledoux and Talagrand., 1991] to construct a new random variable $\mathcal{Y}(\delta,\Delta)$, such that,

$$E((\mathcal{Z}(\delta,\Delta)-\mathcal{Z}(\delta^{'},\Delta^{'}))^2) \leq E((\mathcal{Y}(\delta,\Delta)-\mathcal{Y}(\delta^{'},\Delta^{'}))^2). \tag{4.41}$$

Then we can apply the Gaussian comparison inequality based on (4.41) to derive that $E(\sup_{(\delta,\Delta)\in\mathcal{B}(\gamma)}\mathcal{Z}(\delta,\Delta)) \leq 2E(\sup_{(\delta,\Delta)\in\mathcal{B}(\gamma)}\mathcal{Y}(\delta,\Delta))$. With $E(\mathcal{Z}(\delta,\Delta))=0$, we can show that $E((\mathcal{Z}(\delta,\Delta)-\mathcal{Z}(\delta^{'},\Delta^{'}))^2) = \text{var}(\mathcal{Z}(\delta,\Delta)-\mathcal{Z}(\delta^{'},\Delta^{'}))$. Therefore, we can analyze the second moment through the variance such that

$$\text{var}(\mathcal{Z}(\delta,\Delta)-\mathcal{Z}(\delta^{'},\Delta^{'})) \leq 2\text{var}(\mathcal{Z}(\delta,\Delta)-\mathcal{Z}(\delta^{'},\Delta)) + 2\text{var}(\mathcal{Z}(\delta^{'},\Delta)-\mathcal{Z}(\delta^{'},\Delta^{'}))$$

Let the conditional variance given x_{ki}'s and u^*_{ki} be denoted by var*. Conditioned on x_{ki}'s and u^*_{ki}'s, we can show that the variance of $\mathcal{Z}(\delta,\Delta)-\mathcal{Z}(\delta^{'},\Delta)$ and $\mathcal{Z}(\delta^{'},\Delta)-\mathcal{Z}(\delta^{'},\Delta^{'})$ can

be bounded above based on the Lipschitz continuity,

$$\begin{aligned}
\mathrm{var}^*(\mathcal{Z}(\delta,\Delta)-\mathcal{Z}(\delta^{'},\Delta)) &= \frac{1}{n^2}\sum_{k=1}^{K}\sum_{i=1}^{n}\varphi^2_{\tau_0/2}(x_{ki}^T\Delta_k)\varphi^2_{\tau_k-\tau_0}(u^*_{ki})\left\{\frac{\phi_{\tau_0\|\delta\|_2/4r}(x_{ki}^T\delta_k)}{\|\delta\|_2^2}-\frac{\phi_{\tau_0\|\delta'\|_2/4r}(x_{ki}^T\delta_k^{'})}{\|\delta'\|_2^2}\right\}^2 \\
&\overset{(i)}{\leq} \frac{1}{n^2}\sum_{k=1}^{K}\sum_{i=1}^{n}\left\{\phi_{\tau_0/4r}(x_{ki}^T\frac{\delta_k}{\|\delta\|_2})-\phi_{\tau_0/4r}(x_{ki}^T\frac{\delta_k^{'}}{\|\delta'\|_2})\right\}^2 \\
&\leq \frac{1}{n^2}\sum_{k=1}^{K}\sum_{i=1}^{n}\frac{\tau_0^2}{16r^2}\left\{x_{ki}^T\frac{\delta_k}{\|\delta\|_2}-x_{ki}^T\frac{\delta_k^{'}}{\|\delta'\|_2}\right\}^2,
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{var}^*(\mathcal{Z}(\delta^{'},\Delta)-\mathcal{Z}(\delta^{'},\Delta^{'})) &= \frac{1}{n^2}\sum_{k=1}^{K}\sum_{i=1}^{n}\frac{\phi^2_{\tau_0\|\delta_k'\|_2/4r}(x_{ki}^T\delta_k^{'})}{\|\delta'\|_2^4}\varphi^2_{\tau_k-\tau_0}(u^*_{ki})\left\{\varphi_{\tau_0/2}(x_{ki}^T\Delta_k)-\varphi_{\tau_0/2}(x_{ki}^T\Delta_k^{'})\right\}^2 \\
&\overset{(ii)}{\leq} \frac{1}{n^2}\sum_{k=1}^{K}\sum_{i=1}^{n}\left[\frac{\tau_0}{8r}\right]^4\left\{\frac{4}{\tau_k}x_{ki}^T(\Delta_k-\Delta_k^{'})\right\}^2 \\
&= \frac{1}{n^2}\sum_{k=1}^{K}\sum_{i=1}^{n}(\frac{\tau_0}{16r^2})^2\left\{x_{ki}^T(\Delta_k-\Delta_k^{'})\right\}^2,
\end{aligned}$$

where both steps (i) and (ii) apply the homogeneity property of function $\phi_T(x)$ that $c^2\phi_T(x) = \phi_{cT}(cx)$.

Based on the results above, we can construct a random variable $\mathcal{Y}(\delta,\Delta)$:

$$\mathcal{Y}(\delta,\Delta) = \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n}\frac{1}{2r}\tau_0 y_{ki}^{'}x_{ki}^T\frac{\delta_k}{\|\delta\|_2}+\frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n}\frac{1}{8r^2}\tau_0 y_{ki}^{''}x_{ki}^T\Delta_k, \tag{4.42}$$

where $y_{ki}^{'}$ and $y_{ki}^{''}$ are i.i.d. standard normal random variables. Therefore, we can show that

$$\mathrm{var}(\mathcal{Z}(\delta,\Delta)-\mathcal{Z}(\delta^{'},\Delta^{'}))\leq \mathrm{var}(\mathcal{Y}(\delta,\Delta)-\mathcal{Y}(\delta^{'},\Delta^{'})),$$

which implies $E(\sup_{(\delta,\Delta)\in\mathcal{B}(\gamma)}\mathcal{Z}(\delta,\Delta))\leq 2E(\sup_{(\delta,\Delta)\in\mathcal{B}(\gamma)}\mathcal{Y}(\delta,\Delta))$. By plugging (4.42), we

can show that

$$
\begin{aligned}
E(\sup_{(\delta,\Delta)\in\mathcal{B}(\gamma)} \mathcal{Z}(\delta,\Delta)) \leq & \frac{1}{r} E(\sup_{(\delta,\Delta)\in\mathcal{B}(\gamma)} \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n} \tau_0 y_{ki}^{'} x_{ki}^T \frac{\delta_k}{\|\delta\|_2}) + \frac{1}{4r^2} E(\sup_{(\delta,\Delta)\in\mathcal{B}(\gamma)} \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n} \tau_0 y_{ki}^{''} x_{ki}^T \Delta_k) \\
\overset{(i)}{\leq} & \frac{\tau_0}{r}\frac{\|\delta\|_{2,1}}{\|\delta\|_2} E(\sqrt{K}\sup_k \|\frac{1}{n}\sum_{i=1}^{n} y_{ki}^{'} x_{ki}\|_\infty) + \frac{\tau_0\|\Delta\|_1}{4r^2} E(\sup_k \|\frac{1}{n}\sum_{i=1}^{n} y_{ki}^{''} x_{ki}\|_\infty) \\
\overset{(ii)}{\leq} & \left(\sqrt{K}\gamma + \frac{R}{4r}\right)\frac{2A_0\tau_0}{r}\sqrt{\frac{2\log(p_n)}{n}}.
\end{aligned}
$$

The step (i) is obtained based on Hölder's inequality in Lemma 4.10. In step (ii), we derive the upper bound based on Lemma 4.5.

By combining (4.39) and (4.40) with results above, we can show that with $1/\sqrt{n} = o(1)$,

$$
\begin{aligned}
E(\mathcal{G}(\delta,\Delta)) \leq & \frac{4A_0^2}{\sqrt{n}} + \frac{2\sqrt{\pi}RA_0\tau_0}{r^2}\sqrt{\frac{\log(p_n)}{n}} + 8\sqrt{K\pi}\frac{A_0\tau_0\gamma}{r}\sqrt{\frac{\log(p_n)}{n}} \\
= & \frac{2\sqrt{\pi}RA_0\tau_0}{r^2}\sqrt{\frac{\log(p_n)}{n}} + 8\sqrt{K\pi}\frac{A_0\tau_0\gamma}{r}\sqrt{\frac{\log(p_n)}{n}} + o(1). \qquad (4.43)
\end{aligned}
$$

Next, we analyze the concentration inequality for $\mathcal{G}(\delta,\Delta)$ by a similar method from Sun et al. [2020] and Pan et al. [2021]. Let $g(\delta,\Delta) = n^{-1}\sum_{i=1}^{n} g_i(\delta,\Delta)$, where

$$
g_i(\delta,\Delta) := \sum_{k=1}^{K} \phi_{\tau_0\|\delta\|_2/4r}(x_{ki}^T\delta_k)\varphi_{\tau_0/2}(x_{ki}^T\Delta_k)\varphi_{\tau_k-\tau_0}(u_{ki}^*),
$$

so the random variable $\mathcal{G}(\delta,\Delta)$ can be given by

$$
\mathcal{G}(\delta,\Delta) = \sup_{(\delta,\Delta)\in\mathcal{B}(\gamma)} \frac{|n^{-1}\sum_{i=1}^{n} g_i(\delta,\Delta) - E(g_i(\delta,\Delta))|}{\|\delta\|_2^2}.
$$

Based on the properties (4.37), the function $|\phi_T(x)\varphi_t(x)| \leq T^2/4$ for all pairs (T,t). All $|g_i(\delta,\Delta) - E(g_i(\delta,\Delta))|$'s are uniformly bounded and measurable for all $(\delta,\Delta)\in\mathcal{B}(\gamma)$. Therefore, we can apply Bousquet's version of Talagrand's inequality to $\mathcal{G}(\delta,\Delta)$ based on Bousquet

[2003] and Massart and Picard [2007], such that,

$$\mathcal{G}(\delta,\Delta) - E(\mathcal{G}(\delta,\Delta)) \leq E(\mathcal{G}(\delta,\Delta)) + \sigma_n \frac{\sqrt{2\log(p_n)}}{n} + \big(\frac{\tau_0}{8r}\big)^2 \frac{4K\log(p_n)}{3n}, \tag{4.44}$$

with probability as least $1-\exp\{-\log(p_n)\}$, where σ_n is defined as

$$\begin{aligned}
\sigma_n^2 :=& \sup_{(\delta,\Delta)\in\mathcal{B}(\gamma)} \sum_{i=1}^{n} E[(\frac{g_i(\delta,\Delta) - E(g_i(\delta,\Delta))}{\|\delta\|_2^2})^2] \\
\leq& \sup_{(\delta,\Delta)\in\mathcal{B}(\gamma)} \sum_{i=1}^{n} E[(\sum_{k=1}^{K}(x_{ki}^T \frac{\delta_k}{\|\delta\|_2})^2)^2] \\
\leq& \sup_{(\delta,\Delta)\in\mathcal{B}(\gamma)} \sum_{i=1}^{n} E[\sum_{k=1}^{K}\sum_{k'=1}^{K}(x_{ki}^T \frac{\delta_k}{\|\delta\|_2})^2 (x_{k'i}^T \frac{\delta_{k'}}{\|\delta\|_2})^2] \leq 4KA_0^4 n.
\end{aligned}$$

Based on the assumptions that $n \gtrsim s^2\log(p_n)$ and $R \lesssim s$, we can combine (4.43) and (4.44) to show that with probability at least $1-\exp\{-\log(p_n)\}$,

$$\begin{aligned}
\mathcal{G}(\delta,\Delta) \leq& \frac{4\sqrt{\pi}RA_0\tau_0}{r^2}\sqrt{\frac{\log(p_n)}{n}} + 16\sqrt{\pi}\frac{A_0\tau_0\gamma}{r}\sqrt{\frac{K\log(p_n)}{n}} \\
&+ \big(\frac{\tau_0}{8r}\big)^2 \frac{4K\log(p_n)}{3n} + 2\sqrt{2}A_0^2\sqrt{\frac{K\log(p_n)}{n}} \\
\overset{(i)}{\leq}& \frac{\alpha_l}{4} + 16\sqrt{\pi}\frac{A_0\tau_0\gamma}{r}\sqrt{\frac{K\log(p_n)}{n}} + o(1).
\end{aligned} \tag{4.45}$$

The step (i) is obtained by applying $n \gtrsim s\log(p_n)$, which leads to $\log(p_n)/n = o(1)$, and

$$\frac{4\sqrt{\pi}RA_0\tau_0}{r^2}\sqrt{\frac{\log(p_n)}{n}} \lesssim \sqrt{\frac{s\log(p_n)}{n}} \leq \frac{\alpha}{4}.$$

Combining (4.38) and (4.45), we can show that the first order-Taylor error can satisfy

$$\frac{\mathcal{T}(\theta_1,\theta_2)}{\|\delta\|_2^2} \geq \frac{\alpha_l}{4} - \frac{C_0A_0\tau_0\gamma}{r}\sqrt{\frac{K\log(p_n)}{n}} \tag{4.46}$$

with probability at least $1-\exp\{-\log(p_n)\}$ and $C_0 = 16\sqrt{\pi}$,.

The last step is to extend the result above to be bounded uniformly over the ratio $\|\delta\|_{2,1}/\|\delta\|_2$, and we apply a peeling argument similar to Loh [2017]. Define the functions

$$h(\theta_1,\theta_2) := \frac{\alpha_l}{4} - \frac{\mathcal{T}(\theta_1,\theta_2)}{\|\delta\|_2^2}, g(\gamma) := \frac{C_0 A_0 \tau_{\max}}{2r}\gamma\sqrt{\frac{K\log(p_n)}{n}},$$
$$\text{and } \Gamma(\theta_1,\theta_2) = \frac{\|\theta_1-\theta_2\|_{2,1}}{\|\theta_1-\theta_2\|_2},$$

and the event

$$\mathcal{E} = \Big\{\frac{\mathcal{T}(\theta_1,\theta_2)}{\|\delta\|_2^2} \geq \frac{\alpha_l}{4} - \frac{C_0 A_0 \tau_0}{r}\frac{\|\delta\|_{2,1}}{\|\delta\|_2}\sqrt{\frac{K\log(p_n)}{n}}, \forall(\delta,\Delta)\in\mathcal{B}(\gamma)\Big\}$$
$$\text{for } 1 \leq \gamma \leq C_0\frac{\alpha_l r}{A_0\tau_0}\sqrt{\frac{n}{K\log(p_n)}}.$$

Based on (4.46), we can derive

$$\begin{aligned}
P(\sup_{(\delta,\Delta)\in\mathcal{B}(\gamma)} h(\theta_1,\theta_2) \geq 2g(\gamma)) &= P(\sup_{(\delta,\Delta)\in\mathcal{B}(\gamma)} \frac{\alpha_l}{4} - \frac{\mathcal{T}(\theta_1,\theta_2)}{\|\delta\|_2^2} \geq \frac{C_0 A_0\tau_0}{r}\gamma\sqrt{\frac{K\log(p_n)}{n}}) \\
&\leq P(\sup_{(\delta,\Delta)\in\mathcal{B}(\gamma)} \frac{|g(\delta,\Delta)-E(g(\delta,\Delta))|}{\|\delta\|_2^2} - \frac{E(g(\delta,\Delta))}{\|\delta\|_2^2} + \frac{\alpha_l}{4} \geq \frac{C_0 A_0\tau_0}{r}\gamma\sqrt{\frac{K\log(p_n)}{n}}) \\
&\leq P(\mathcal{G}(\delta,\Delta) \geq \frac{C_0 A_0\tau_0}{r}\gamma\sqrt{\frac{K\log(p_n)}{n}} - \frac{\alpha_l}{4}) \leq \exp\{-\log(p_n)\}.
\end{aligned}$$

Then define the set for integer $m \geq 1$,

$$V_m := \{(\theta_1,\theta_2) : 2^{m-1}\mu \leq g(\Gamma(\theta_1,\theta_2)) \leq 2^m\mu\} \cap \mathcal{B}(\gamma), \text{ with } \mu = C_0\frac{A_0\tau_0}{r}\sqrt{\frac{K\log(p_n)}{n}}.$$

We can derive a union bound with the index m ranging up to $M = \lceil \log(c\sqrt{n/\log(p_n)})\rceil$

for some c,

$$\begin{aligned}
P(\mathcal{E}^c) \leq & \sum_{m=1}^{M} P(\exists(\theta_1,\theta_2) \in V_m : h(\theta_1,\theta_2) \geq 2g(\Gamma(\theta_1,\theta_2))) \\
\leq & \sum_{m=1}^{M} P(\sup_{\substack{(\delta,\Delta)\in\mathcal{B}(\gamma) \\ \frac{\|\delta\|_1}{\|\delta\|_2}\leq g^{-1}(2^m\mu)}} h(\theta_1,\theta_2) \geq 2^m\mu) \\
\leq & C_1 \exp\{-C_2\log(p_n) + C_3 \log\log(\frac{n}{\log(p_n)})\}
\end{aligned}$$

for some positive constant C_1, C_2, and C_3.

Therefore, we can show that with probability at least $1 - C^{'}\exp\{-C^{''}\log(p_n)\}$ for some constant C, $C^{'}$, and $C^{''}$

$$\begin{aligned}
\mathcal{T}(\theta_1,\theta_2) \geq & \frac{\alpha_l}{4}\|\theta_1-\theta_2\|_2^2 - \frac{C_0A_0\tau_0}{r}\|\theta_1-\theta_2\|_{2,1}\|\theta_1-\theta_2\|_2\sqrt{\frac{K\log(p_n)}{n}} \\
\overset{(i)}{\geq} & \frac{\alpha_l}{8}\|\theta_1-\theta_2\|_2^2 - \frac{CKA_0^2\tau_0^2}{\alpha_l r^2}\frac{\log(p_n)}{n}\|\theta_1-\theta_2\|_{2,1}^2,
\end{aligned}$$

where the step (i) is obtained by

$$\begin{aligned}
\frac{C_0A_0\tau_0}{r}\|\theta_1-\theta_2\|_{2,1}\|\theta_1-\theta_2\|_2\sqrt{\frac{K\log(p_n)}{n}} \leq & \frac{\alpha_l}{8}\|\theta_1-\theta_2\|_2^2 \\
& + \frac{C}{\alpha_l}\big(\frac{A_0\tau_0}{r}\big)^2\frac{K\log(p_n)}{n}\|\theta_1-\theta_2\|_{2,1}^2.
\end{aligned}$$

□

Lemma 4.4. *Based on Assumptions 4.1-4.2, the proposed loss function $\mathcal{L}(\theta)$ satisfies the restricted smoothness (RSM) condition, such that with probability at least $1-2\exp\{-\log(p_n)\}$,*

$$n^{-1}(\mathcal{L}(\theta_1) - \mathcal{L}(\theta_2) - \nabla\mathcal{L}(\theta_1)^T(\theta_1-\theta_2)) \leq \kappa_u\|\theta_1-\theta_2\|_2^2 + \tau_u\frac{\log(p_n)}{n}\|\theta_1-\theta_2\|_{2,1}^2, \quad (4.47)$$

where $\kappa_u = K\alpha_u$ *and* $\tau_u = 4KA_0^2$.

Proof. We can show that the proposed loss $\mathcal{L}(\theta)$ satisfies the restricted smoothness (RSM) based on a similar approach from some previous works [Agarwal et al., 2012a, Loh and Wainwright, 2015, Fan et al., 2017], which can be used to analyze the optimization properties in section 4.3.3. We can apply Taylor expansion with $\tilde{\theta} = \alpha\theta_1 + (1-\alpha)\theta_2$ for $\alpha \in (0,1)$,

$$\begin{aligned} n^{-1}\mathcal{L}(\theta_1) - n^{-1}\mathcal{L}(\theta_2) - n^{-1}\nabla\mathcal{L}(\theta_1)^T(\theta_1-\theta_2) \leq & \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n}(x_{ki}^T(\theta_{k1}-\theta_{k2}))^2 \\ = & \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n}E((x_{ki}^T(\theta_{k1}-\theta_{k2}))^2) \\ & + |x_{ki}^T(\theta_{k1}-\theta_{k2}))^2 - E((x_{ki}^T(\theta_{k1}-\theta_{k2}))^2| \\ \leq & K\alpha_u\|\theta_1-\theta_2\|_2^2 + 4KA_0^2\frac{\log(p_n)}{n}\|\theta_1-\theta_2\|_{2,1}. \end{aligned}$$

The last step is obtained by applying the concentration probability of the sub-exponential variable $(x_{ki}^T u)^2$, which holds with probability at least $1 - 2\exp\{-\log(p_n)\}$. □

Lemma 4.5. *Let* $\{y_i\}_i^n$ *be i.i.d standard Gaussian variables and* $\{X_i\}_{i=1}^n$ *be i.i.d sub-Gaussian vectors with* $X_i = (x_{i1}, x_{i2}, \cdots, x_{ip}, \cdots, x_{ip_n})^T$. *For some* $m \geq 1$, $E(x_{ip}^m) \leq mA_0^m\Gamma(m/2)$. *Then with* $n > \log(p_n)$,

$$E(\|n^{-1}\sum_{i=1}^{n} y_i X_i\|_\infty) \leq 2\sqrt{2}A_0\sqrt{\frac{\log(p_n)}{n}}.$$

Proof. We set variable $Z = \|n^{-1}\sum_{i=1}^n y_i X_i\|_\infty$, and use Jensen's inequality to show that for some $0 < t < n/A_0$,

$$\begin{aligned} \exp\{tE(Z)\} \leq & E(\exp\{tZ\}) \\ = & E(\sup_p \exp\{tn^{-1}\sum_{i=1}^{n} y_i x_{ip}\}) \end{aligned}$$

$$\leq \sum_{p=1}^{p_n} \prod_{i=1}^{n} E(\exp\{n^{-1} t y_i x_{ip}\}). \tag{4.48}$$

Next, we can derive an upper bound for $E(\exp\{tn^{-1}y_i x_{ip}\})$, such that

$$\begin{aligned}
E(\exp\{n^{-1}ty_i x_{ip}\}) =& E(E(\exp\{n^{-1}ty_i x_{ip}\}|x_{ip})) \\
\leq& E(\exp\{t^2 x_{ip}^2/(2n^2)\}) \\
\leq& 1 + 2\sum_{m=1}^{\infty} \big(\frac{t^2}{2n^2}\big)^m A_0^{2m} \\
\leq& \exp\{\frac{(tA_0)^2/(n^2)}{1-(t^2A_0^2/(2n^2))}\} \\
\leq& \exp\{2\big(\frac{tA_0}{n}\big)^2\}.
\end{aligned} \tag{4.49}$$

We combine (4.48) and (4.49) to show that with $0 < t \leq n/A_0$,

$$\begin{aligned}
\exp\{tE(Z)\} &\leq p_n \exp\{t^2 \frac{2A_0^2}{n}\}, \\
E(Z) &\leq \frac{\log(p_n)}{t} + t\frac{2A_0^2}{n},
\end{aligned}$$

which can be minimized at $t = \sqrt{n\log(p_n)/(2A_0^2)}$, such that

$$E(Z) \leq \frac{\log(p_n)}{\sqrt{n\log(p_n)/(2A_0^2)}} + \sqrt{n\log(p_n)/(2A_0^2)}\frac{2A_0^2}{n} \leq 2\sqrt{2}A_0\sqrt{\frac{\log(p_n)}{n}}.$$

□

Lemma 4.6. *Let the sample covariance be denoted as $\hat{\Sigma} = X^T X/n$. Based on Assumptions 4.1 - 4.2, there exist some positive constant sets $\{c_j\}_{j=1}^3$ such that*

$$P\big(\Big|\Big|\Big|\hat{\Sigma}_{\mathcal{SS}}^{-1} - \Sigma_{\mathcal{SS}}^{-1}\Big|\Big|\Big|_2 \geq c_3\big(\frac{Ks}{n} + \sqrt{\frac{Ks}{n}}\big)\big) \leq c_1 \exp\{-c_2 ns\}\}. \tag{4.50}$$

Furthermore, if the parameter ξ from Assumption 4.2 satisfies $\xi \in (0,1)$, then for some vector $e \in \mathbb{R}^{Ks}$ and $\|e\|_{2,\infty} \leq 1$,

$$P\big(\sup_{p\in\mathcal{S}^c} \|((\hat{\Sigma}_{\mathcal{S}^c\mathcal{S}}\hat{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1} - \Sigma_{\mathcal{S}^c\mathcal{S}}\Sigma_{\mathcal{S}\mathcal{S}}^{-1})e^T)^{(p)}\|_2 \geq \frac{\xi}{2}\big) \leq C_3 \exp\{-C_4 \min\{s, \log(p_n)\}\}, \tag{4.51}$$

with some constants $C_3 > 0$ and $C_4 > 0$.

Proof. Based on Assumption 4.1, each row of covariates X is sampled from a sub-Gaussian vector with parameter A_0. We can apply Lemma 4.11 to show that with probability $1 - c_1\exp\{-c_2 s\}$

$$\Big|\Big|\Big|\hat{\Sigma}_{\mathcal{S}\mathcal{S}} - \Sigma_{\mathcal{S}\mathcal{S}}\Big|\Big|\Big|_2 \leq c_0(\sqrt{\frac{Ks}{n}} + \frac{Ks}{n})$$

for a set of constants $\{c_j\}_{j=0}^2$. According to Lemma 11 from Loh and Wainwright [2017], if $|||\Sigma_{\mathcal{S}\mathcal{S}}^{-1}|||_2 \Big|\Big|\Big|\hat{\Sigma}_{\mathcal{S}\mathcal{S}} - \Sigma_{\mathcal{S}\mathcal{S}}\Big|\Big|\Big| \leq 1/2$, then $\Big|\Big|\Big|\hat{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1} - \Sigma_{\mathcal{S}\mathcal{S}}^{-1}\Big|\Big|\Big|_2 = \mathcal{O}(|||\Sigma_{\mathcal{S}\mathcal{S}}^{-1}|||_2^2 \Big|\Big|\Big|\hat{\Sigma}_{\mathcal{S}\mathcal{S}} - \Sigma_{\mathcal{S}\mathcal{S}}\Big|\Big|\Big|_2)$. Since $\Big|\Big|\Big|\Sigma_{\mathcal{S}\mathcal{S}}^{-1}\Big|\Big|\Big|_2 \leq \alpha_l^{-1}$, we can further derive that for c_3,

$$P\big(\Big|\Big|\Big|\hat{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1} - \Sigma_{\mathcal{S}\mathcal{S}}^{-1}\Big|\Big|\Big|_2 \geq c_3(\frac{Ks}{n} + \sqrt{\frac{Ks}{n}})\big) \leq c_1\exp\{-c_2 s\}.$$

Next, we need to analyze the sub-matrix $\hat{\Sigma}_{\mathcal{S}^c\mathcal{S}}$. Based the argument from Loh and Wainwright [2017],

$$\sup_{p\in\mathcal{S}^c} \|u_p^T(\hat{\Sigma}_{\mathcal{S}^c\mathcal{S}} - \Sigma_{\mathcal{S}^c\mathcal{S}})\|_2 \lesssim \max\{\sqrt{\frac{Ks}{n}}, \sqrt{\frac{\log(Kp_n)}{n}}\},$$

with probability at least $1 - b_1\exp\{-b_2\min\{s, \log(p_n)\}\}$ for some constants b_1 and b_2. In addition,

$$\|((\hat{\Sigma}_{\mathcal{S}^c\mathcal{S}}\hat{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1} - \Sigma_{\mathcal{S}^c\mathcal{S}}\Sigma_{\mathcal{S}\mathcal{S}}^{-1})e)\|_\infty \leq \|(\hat{\Sigma}_{\mathcal{S}^c\mathcal{S}} - \Sigma_{\mathcal{S}^c\mathcal{S}})(\hat{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1} - \Sigma_{\mathcal{S}\mathcal{S}}^{-1})e\|_\infty$$

$$+\|(\hat{\Sigma}_{\mathcal{S}^c\mathcal{S}} - \Sigma_{\mathcal{S}^c\mathcal{S}})\Sigma_{\mathcal{S}\mathcal{S}}^{-1})e\|_\infty + \|\Sigma_{\mathcal{S}^c\mathcal{S}}(\hat{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1} - \Sigma_{\mathcal{S}\mathcal{S}}^{-1})e\|_\infty.$$

Each component can be bounded as

$$\begin{aligned}
\|(\hat{\Sigma}_{\mathcal{S}^c\mathcal{S}} - \Sigma_{\mathcal{S}^c\mathcal{S}})(\hat{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1} - \Sigma_{\mathcal{S}\mathcal{S}}^{-1})e\|_\infty &\leq \sup_{p\in\mathcal{S}^c} \|u_p^T(\hat{\Sigma}_{\mathcal{S}^c\mathcal{S}} - \Sigma_{\mathcal{S}^c\mathcal{S}})\|_2 \left|\left|\left|\hat{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1} - \Sigma_{\mathcal{S}\mathcal{S}}^{-1}\right|\right|\right|_2 \|e\|_2 \\
&\lesssim \max\{\sqrt{\frac{Ks}{n}}, \sqrt{\frac{K\log(p_n)}{n}}\}(\frac{Ks}{n} + \sqrt{\frac{Ks}{n}})\sqrt{s} \\
&\lesssim \max\{\sqrt{\frac{s^3}{n^2}}, \sqrt{\frac{s^2\log(p_n)}{n^2}}\}, \\
\|(\hat{\Sigma}_{\mathcal{S}^c\mathcal{S}} - \Sigma_{\mathcal{S}^c\mathcal{S}})\Sigma_{\mathcal{S}\mathcal{S}}^{-1})e\|_\infty &\leq \sup_{p\in\mathcal{S}^c} \|u_p^T(\hat{\Sigma}_{\mathcal{S}^c\mathcal{S}} - \Sigma_{\mathcal{S}^c\mathcal{S}})\|_2 \left|\left|\left|\Sigma_{\mathcal{S}\mathcal{S}}^{-1}\right|\right|\right|_2 \|e\|_2 \\
&\lesssim \max\{\sqrt{\frac{s^2}{n}}, \sqrt{\frac{s\log(p_n)}{n}}\}, \\
\|\Sigma_{\mathcal{S}^c\mathcal{S}}(\hat{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1} - \Sigma_{\mathcal{S}\mathcal{S}}^{-1})e\|_\infty &\leq \|\Sigma_{\mathcal{S}^c\mathcal{S}}\Sigma_{\mathcal{S}\mathcal{S}}^{-1}\|_\infty \left|\left|\left|\hat{\Sigma}_{\mathcal{S}\mathcal{S}} - \Sigma_{\mathcal{S}\mathcal{S}}\right|\right|\right|_2 \left|\left|\left|\hat{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\right|\right|\right|_2 \|e\|_2 \\
&\lesssim \sqrt{\frac{s^2}{n}} + \sqrt{\frac{s^3}{n^2}}.
\end{aligned}$$

Also, we can apply the relationship between ℓ_∞ norm and ℓ_2 norm as follows,

$$\sup_{p\in\mathcal{S}^c} \|(\hat{\Sigma}_{\mathcal{S}^c\mathcal{S}}\hat{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}e)^{(p)}\|_2 \leq \sqrt{K}\|\hat{\Sigma}_{\mathcal{S}^c\mathcal{S}}\hat{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}e\|_\infty.$$

Thus, with $n \gtrsim s^2\log(p_n)$, we can combine the results above to show that

$$\begin{aligned}
\sup_{p\in\mathcal{S}^c} \|(\hat{\Sigma}_{\mathcal{S}^c\mathcal{S}}\hat{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}e)^{(p)}\|_2 \leq& \sqrt{K}\{\|\Sigma_{\mathcal{S}^c\mathcal{S}}\Sigma_{\mathcal{S}\mathcal{S}}^{-1}e\|_\infty + \|(\hat{\Sigma}_{\mathcal{S}^c\mathcal{S}}\hat{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1} - \Sigma_{\mathcal{S}^c\mathcal{S}}\Sigma_{\mathcal{S}\mathcal{S}}^{-1})e\|_\infty\} \\
\lesssim& \sqrt{K}\left|\left|\left|\Sigma_{\mathcal{S}^c\mathcal{S}}\Sigma_{\mathcal{S}\mathcal{S}}^{-1}\right|\right|\right|_\infty \|e\|_\infty + \max\{\sqrt{\frac{s^3}{n^2}}\sqrt{\frac{s^2\log(p_n)}{n^2}}\} \\
&+ \max\{\sqrt{\frac{s^2}{n}}, \sqrt{\frac{s\log(p_n)}{n}}\} \\
\lesssim& (1-\xi) + o(1) < 1 - \frac{\xi}{2}
\end{aligned}$$

with probability at least $1 - C_3 \exp\{-C_4 \min\{s, \log(p_n)\}\}$ for some constant C_3 and C_4,

□

Lemma 4.7. *Based on Assumptions 4.1 - 4.2, let the initial precision level be set as $d \leq r$, then we have*

$$\|\theta^t - \hat{\theta}\|_2 \leq d, \forall t \geq 1.$$

Proof. The proof of Lemma 4.7 can be shown by induction. Suppose $\|\theta^t - \hat{\theta}\|_2 \leq d$, we need to obtain that $\|\theta^{t+1} - \hat{\theta}\|_2 \leq d$. Let the distance be denoted by $\hat{\Delta}^t = \theta^t - \hat{\theta}$. Since $Q(\theta^{t+1}|\theta^t) \geq Q(\theta^{t+1})$,

$$\begin{aligned}
&\frac{1}{n}Q(\theta^{t+1}) - \frac{1}{n}Q(\hat{\theta}) \leq \frac{1}{n}Q(\theta^{t+1}|\theta^t) - \frac{1}{n}Q(\hat{\theta}) \\
&\quad = \frac{1}{n}\mathcal{L}(\theta^t) + \frac{1}{n}\nabla\mathcal{L}(\theta^t)^T(\theta^{t+1} - \theta^t) - \frac{1}{n}\mathcal{L}(\hat{\theta}) + \frac{\gamma}{2}\|\theta^{t+1} - \theta^t\|_2^2 + \lambda_n\|\theta^{t+1}\|_{2,1} - \lambda_n\|\hat{\theta}\|_{2,1} \\
&\quad \overset{(i)}{\leq} \frac{1}{n}\mathcal{L}(\theta^t) + \frac{1}{n}\nabla\mathcal{L}(\theta^t)^T(\theta^{t+1} - \theta^t) + \frac{1}{n}\mathcal{L}(\theta^t) + \tau_l\frac{\log(p_n)}{n}\|\hat{\Delta}^t\|_{2,1}^2 + \lambda_n(\|\theta^{t+1}\|_{2,1} - \|\hat{\theta}\|_{2,1}) \\
&\quad - \frac{1}{n}\nabla\mathcal{L}(\theta^t)^T(\hat{\theta} - \theta^t) + \frac{\gamma}{2}\|\theta^{t+1} - \theta^t\|_2^2 - \kappa_l\|\hat{\Delta}^t\|_2^2 \\
&\quad \leq \frac{1}{n}\nabla\mathcal{L}(\theta^t)^T\hat{\Delta}^{t+1} + \frac{\gamma}{2}\|\theta^{t+1} - \theta^t\|_2^2 - \kappa_l\|\hat{\Delta}^t\|_2^2 + \tau_l\frac{\log(p_n)}{n}\|\hat{\Delta}^t\|_{2,1}^2 + \lambda_n(\|\theta^{t+1}\|_{2,1} - \|\hat{\theta}\|_{2,1}) \\
&\quad \overset{(ii)}{\leq} (\frac{1}{n}\nabla\mathcal{L}(\theta^t) + \lambda_n z^{t+1})^T\hat{\Delta}^{t+1} + \frac{\gamma}{2}\|\theta^{t+1} - \theta^t\|_2^2 - \kappa_l\|\hat{\Delta}^t\|_2^2 + \tau_l\frac{\log(p_n)}{n}\|\hat{\Delta}^t\|_{2,1}^2 \\
&\quad \overset{(iii)}{=} -\gamma(\theta^{t+1} - \theta^t)^T\hat{\Delta}^{t+1} + \frac{\gamma}{2}\|\theta^{t+1} - \theta^t\|_2^2 - \kappa_l\|\hat{\Delta}^t\|_2^2 + \tau_l\frac{\log(p_n)}{n}\|\hat{\Delta}^t\|_{2,1}^2 \\
&\quad = \frac{\gamma - 2\kappa_l}{2}\|\hat{\Delta}^t\|_2^2 - \frac{\gamma}{2}\|\hat{\Delta}^{t+1}\|_2^2 + \tau_l\frac{\log(p_n)}{n}\|\hat{\Delta}^t\|_{2,1}^2.
\end{aligned}$$

Based on the argument from Loh [2017], the following term in the step (i) of the above derivation can also satisfy the RSC condition through similar derivation as the proof of Lemma 4.3,

$$\frac{1}{n}\mathcal{L}(\hat{\theta}) - \frac{1}{n}\mathcal{L}(\theta^t) - \frac{1}{n}\nabla\mathcal{L}(\theta^t)^T(\hat{\theta} - \theta^t) \geq \kappa_l\|\hat{\Delta}^t\|_2^2 - \tau_l\frac{\log(p_n)}{n}\|\hat{\Delta}^t\|_{2,1}^2.$$

In step (ii), we apply the property of the subdifferential to obtain

$$\|\hat{\theta}\|_{2,1} - \|\theta^{t+1}\|_{2,1} \geq -(\hat{z}^{t+1})^T \hat{\Delta}^{t+1}.$$

The step (iii) applies the optimization $\theta^{t+1} = \arg\min Q(\theta|\theta^t)$, such that

$$\frac{1}{n}\nabla\mathcal{L}(\theta^t) + \gamma(\theta^{t+1} - \theta^t) + \lambda_n z^{t+1} = \mathbf{0}.$$

In addition, we can show that

$$\begin{aligned}
\frac{1}{n}Q(\theta^{t+1}) - \frac{1}{n}Q(\hat{\theta}) =& \frac{1}{n}\mathcal{L}(\theta^{t+1}) - \frac{1}{n}\mathcal{L}(\hat{\theta}) + \lambda_n(\|\theta^{t+1}\|_{2,1} - \|\hat{\theta}\|_{2,1}) \\
\geq& \kappa_l\|\hat{\Delta}^t\|_2^2 - \tau_l\frac{\log(p_n)}{n}\|\hat{\Delta}^t\|_{2,1}^2 + (\frac{1}{n}\nabla\mathcal{L}(\hat{\theta})^T + \lambda_n\hat{z})^T\hat{\Delta}^{t+1} \\
\geq& \kappa_l\|\hat{\Delta}^t\|_2^2 - \tau_l\frac{\log(p_n)}{n}\|\hat{\Delta}^t\|_{2,1}^2
\end{aligned}$$

with $\nabla Q(\hat{\theta}) = \nabla\mathcal{L}(\hat{\theta})^T + n\lambda_n\hat{z} = \mathbf{0}$.

By combining the results above, we can show that

$$\begin{aligned}
2\kappa_l\|\hat{\Delta}^{t+1}\|_2^2 - 2\tau_l\frac{\log(p_n)}{n}\|\hat{\Delta}^{t+1}\|_{2,1}^2 \leq& (\gamma - 2\kappa_l)\|\hat{\Delta}^t\|_2^2 - \gamma\|\hat{\Delta}^{t+1}\|_2^2 + 2\tau_l\frac{\log(p_n)}{n}\|\hat{\Delta}^t\|_{2,1}^2 \\
(2\kappa_l + \gamma)\|\hat{\Delta}^{t+1}\|_2^2 \leq& (\gamma - 2\kappa_l)\|\hat{\Delta}^t\|_2^2 + 8\tau_l R^2\frac{\log(p_n)}{n} \\
\|\hat{\Delta}^{t+1}\|_2^2 \leq& \frac{\gamma - 2\kappa_l}{\gamma + 2\kappa_l}\|\hat{\Delta}^t\|_2^2 + \frac{8\tau_l R^2}{\gamma + \kappa_l}\frac{\log(p_n)}{n},
\end{aligned}$$

which leads to $\|\hat{\Delta}^{t+1}\|_2 \leq d$ based on the sample size assumption $n \gtrsim s^2\log(p_n)$. □

Lemma 4.8. *Based on Assumptions 4.1-4.2, suppose*

$$\max\{4\|\frac{1}{n}\nabla\mathcal{L}(\theta^*)\|_{2,\infty}, 8\tau_l R\frac{\log(p_n)}{n}\} \leq \lambda_n, \tag{4.52}$$

and then there exists a pair ($\bar{\eta}$, T) such that

$$\frac{1}{n}Q(\theta^t) - \frac{1}{n}Q(\hat{\theta}) \leq \bar{\eta}, \forall t \geq T. \tag{4.53}$$

Furthermore, for any $t \geq T$, we can show that

$$\|\theta^t - \hat{\theta}\|_{2,1} \leq 4\sqrt{s}\|\theta^t - \hat{\theta}\|_2 + 8\sqrt{s}\|\hat{\theta} - \theta^*\|_2 + \min(2\frac{\bar{\eta}}{\lambda_n}, R).$$

Proof. Suppose there exists a pair ($\bar{\eta}$, T) such that the algorithm reaches the precision level $\bar{\eta}$ after T iterations,

$$\frac{1}{n}Q(\theta^t) - \frac{1}{n}Q(\hat{\theta}) \leq \bar{\eta}, \forall t \geq T. \tag{4.54}$$

In addition, based on the optimality of $\hat{\theta}$ analyzed in Theorem 4.2, $n^{-1}Q(\hat{\theta}) \leq n^{-1}Q(\theta^*)$. Therefore, we have

$$\frac{1}{n}Q(\theta^t) - \frac{1}{n}Q(\theta^*) \leq \bar{\eta}, \forall t \geq T. \tag{4.55}$$

Next, we can use (4.55) to show that,

$$\begin{aligned}
\frac{1}{n}\mathcal{L}(\theta^t) + \lambda_n\|\theta^t\|_{2,1} \leq & \frac{1}{n}\mathcal{L}(\theta^*) + \lambda_n\|\theta^*\|_{2,1} + \bar{\eta} \\
\frac{1}{n}\mathcal{L}(\theta^t) - \frac{1}{n}\mathcal{L}(\theta^*) - \frac{1}{n}\nabla\mathcal{L}(\theta^*)^T(\theta^t - \theta^*) + \lambda_n\|\theta^t\|_{2,1} - \lambda_n\|\theta^*\|_{2,1} \leq & -\frac{1}{n}\nabla\mathcal{L}(\theta^*)^T(\theta^t - \theta^*) + \bar{\eta} \\
\kappa_l\|\theta^t - \theta^*\|_2^2 - \tau_l\frac{\log(p_n)}{n}\|\theta^t - \theta^*\|_{2,1}^2 + \lambda_n\|\theta^t\|_{2,1} - \lambda_n\|\theta^*\|_{2,1} \leq & \frac{1}{4}\lambda_n\|\theta^t - \theta^*\|_{2,1} + \bar{\eta} \\
\lambda_n\|\theta^t\|_{2,1} - \lambda_n\|\theta^*\|_{2,1} \leq & \frac{1}{2}\lambda_n\|\theta^t - \theta^*\|_{2,1} + \bar{\eta}.
\end{aligned}$$

By the decomposition property of the mixed norm, we can further derive that for subspace

$|\mathcal{E}| \leq s$,

$$\|\theta^t_{\mathcal{E}^c}\|_{2,1} = \|(\theta^t - \theta^*)_{\mathcal{E}^c}\|_{2,1} \leq 3\|(\theta^t - \theta^*)_{\mathcal{E}}\|_{2,1} + 2\frac{\bar{\eta}}{\lambda_n}.$$

If we replace $\hat{\theta}$ to θ^t, we have the identical result in Theorem 4.1,

$$\|(\hat{\theta} - \theta^*)_{\mathcal{E}^c}\|_{2,1} \leq 3\|(\hat{\theta} - \theta^*)_{\mathcal{E}}\|_{2,1}.$$

By applying the inequality of the mixed $\ell_{2,1}$ norm and ℓ_2 norm, we can show that

$$\|\theta^t - \theta^*\|_{2,1} \leq 4\|(\theta^t - \theta^*)_{\mathcal{E}}\|_{2,1} + 2\frac{\bar{\eta}}{\lambda_n} \leq 4\sqrt{s}\|(\theta^t - \theta^*)_{\mathcal{E}}\|_2 + \min(2\frac{\bar{\eta}}{\lambda_n}, R),$$

where we have $\|\theta^t - \theta^*\|_{2,1} \leq R$.

Therefore, we can show that

$$\|\theta^t - \hat{\theta}\|_{2,1} \leq 4\sqrt{s}\|\theta^t - \hat{\theta}\|_2 + 8\sqrt{s}\|\hat{\theta} - \theta^*\|_2 + \min(2\frac{\bar{\eta}}{\lambda_n}, R).$$

In general, we can show that the difference $n^{-1}Q(\theta^t) - n^{-1}Q(\theta^*)$ is bounded above as follows,

$$\begin{aligned}
\frac{1}{n}Q(\theta^t) - \frac{1}{n}Q(\theta^*) =& \frac{1}{n}\mathcal{L}(\theta^t) - \frac{1}{n}\mathcal{L}(\theta^*) + \lambda_n\|\theta^t\|_{2,1} - \lambda_n\|\theta^*\|_{2,1} \\
=& \frac{1}{n}\mathcal{L}(\theta^t) - \frac{1}{n}\mathcal{L}(\theta^*) - \frac{1}{n}\nabla\mathcal{L}(\theta^*)^T(\theta^t - \theta^*) \\
&+ \frac{1}{n}\nabla\mathcal{L}(\theta^*)^T(\theta^t - \theta^*) + \lambda_n\|\theta^t\|_{2,1} - \lambda_n\|\theta^*\|_{2,1} \\
\overset{(i)}{\leq}& \kappa_u\|\theta^t - \theta^*\|_2^2 + \tau_u\frac{\log(p_n)}{n}\|\theta^t - \theta^*\|_{2,1}^2 + \frac{5}{4}\lambda_n\|\theta^t - \theta^*\|_{2,1} \\
\leq& \kappa_u r^2 + 4\tau_u\frac{\log(p_n)}{n}R^2 + \frac{5}{4}\lambda_n R.
\end{aligned}$$

The step (i) can be obtained based on Lemma 4.4 as $\theta^t \in \mathbb{B}_r(\theta^*)$ and condition 4.52. Based on $n^{-1}Q(\hat{\theta}) \leq n^{-1}Q(\theta^*)$, we can choose $\bar{\eta} \geq \kappa_u r^2 + 4\tau_u \log(p_n)R^2/n + 5\lambda_n R/4$, such that $n^{-1}Q(\theta^t) - n^{-1}Q(\hat{\theta}) \leq \bar{\eta}$.

□

Lemma 4.9. *Define parameters κ and β,*

$$\kappa = 1 - \frac{\kappa_l}{2\gamma} + 16\tau_l \frac{s\log(p_n)}{\gamma n} \text{ and } \beta = (2 + \kappa_l + 32\tau_l \frac{s\log(p_n)}{n})\frac{\tau_l}{\gamma}\frac{\log(p_n)}{n},$$

with $\gamma > 2\kappa_l$, and the estimation error $\varepsilon = 8\sqrt{s}\|\hat{\theta} - \theta^*\|_2$ with precision level $\epsilon = \min(\lambda_n^{-1}2\bar{\eta}, R)$.

Based on Assumptions 4.1-4.2, for any $t \in [T_j, T_{j+1})$, we have

$$\frac{1}{n}Q(\theta^t) - \frac{1}{n}Q(\hat{\theta}) \leq \kappa^{t-T_j}(\frac{1}{n}Q(\theta^{T_j}) - \frac{1}{n}Q(\hat{\theta})) + \frac{\beta}{1-\kappa}(\varepsilon + \epsilon)^2,$$

and

$$\|\theta^t - \hat{\theta}\|_2^2 \leq (\kappa_l - 32\frac{\tau_l s\log(p_n)}{n})^{-1}(\frac{1}{n}Q(\theta^t) - \frac{1}{n}Q(\hat{\theta}) + 2\tau_l\frac{\log(p_n)}{n}(\varepsilon + \epsilon)^2).$$

Proof. Let the distance be denoted by $\hat{\Delta}^t = \theta^t - \hat{\theta}$. First, we apply the RSC condition and the property of subdifferential,

$$\begin{aligned}
\kappa_l\|\hat{\Delta}^t\|_2^2 - \tau_l\frac{\log(p_n)}{n}\|\hat{\Delta}^t\|_{2,1}^2 \leq & \frac{1}{n}\mathcal{L}(\theta^t) - \frac{1}{n}\mathcal{L}(\hat{\theta}) - \frac{1}{n}\nabla\mathcal{L}(\hat{\theta})^T\hat{\Delta}^t \\
= & \frac{1}{n}Q(\theta^t) - \frac{1}{n}Q(\hat{\theta}) - \lambda_n\|\theta^t\|_{2,1} + \lambda_n\|\hat{\theta}\|_{2,1} - \frac{1}{n}\nabla\mathcal{L}(\hat{\theta})^T\hat{\Delta}^t \\
\leq & \frac{1}{n}Q(\theta^t) - \frac{1}{n}Q(\hat{\theta}) - (\frac{1}{n}\nabla\mathcal{L}(\hat{\theta}) + \lambda_n\hat{z})^T\hat{\Delta}^t = \frac{1}{n}Q(\theta^t) - \frac{1}{n}Q(\hat{\theta}).
\end{aligned}$$

From Lemma 4.8, the upper bound of $\|\hat{\Delta}^t\|_{2,1}$ can be applied as follows,

$$\begin{aligned}
\kappa_l\|\hat{\Delta}^t\|_2^2 - \tau_l\frac{\log(p_n)}{n}(32s\|\hat{\Delta}^t\|_2^2 + 2(\varepsilon+\epsilon)^2) \leq & \frac{1}{n}Q(\theta^t) - \frac{1}{n}Q(\hat{\theta}) \\
(\kappa_l - 32\tau_l\frac{s\log(p_n)}{n})\|\hat{\Delta}^t\|_2^2 \leq & \frac{1}{n}Q(\theta^t) - \frac{1}{n}Q(\hat{\theta}) + 2\tau_l\frac{\log(p_n)}{n}(\varepsilon+\epsilon)^2 \\
\|\hat{\Delta}^t\|_2^2 \leq & \kappa_0^{-1}(\frac{1}{n}Q(\theta^t) - \frac{1}{n}Q(\hat{\theta}) + 2\tau_l\frac{\log(p_n)}{n}(\varepsilon+\epsilon)^2),
\end{aligned}$$

with $\kappa_0 = \kappa_l - 32\tau_l s\log(p_n)/n$.

Given $\theta' = \alpha\hat{\theta} + (1-\alpha)\theta^t$ for some $\alpha \in (0,1)$, the optimization function can hold the following inequality based on the MM method,

$$\begin{aligned}
\frac{1}{n}Q(\theta^{t+1}) \leq & \frac{1}{n}Q(\theta^{t+1}|\theta^t) \leq \frac{1}{n}Q(\theta'|\theta^t) \\
= & \frac{1}{n}\mathcal{L}(\theta^t) + \frac{1}{n}\nabla\mathcal{L}(\theta^t)^T(\theta' - \theta^t) + \frac{\gamma}{2}\|\theta' - \theta^t\|_2^2 + \lambda\|\theta'\|_{2,1}^2 \\
\leq & \frac{1}{n}\mathcal{L}(\theta^t) + \alpha\frac{1}{n}\nabla\mathcal{L}(\theta^t)^T\hat{\Delta}^t + \frac{\gamma\alpha^2}{2}\|\hat{\Delta}^t\|_2^2 + \lambda_n(\alpha\|\hat{\theta}\|_{2,1} + (1-\alpha)\|\theta^t\|_{2,1}) \\
\leq & \frac{1}{n}\mathcal{L}(\theta^t) + \alpha(\frac{1}{n}\mathcal{L}(\hat{\theta}) - \frac{1}{n}\mathcal{L}(\theta^t) + \tau_l\frac{\log(p_n)}{n}\|\hat{\Delta}^t\|_{2,1}^2) + \frac{\gamma\alpha^2}{2}\|\hat{\Delta}^t\|_2^2 + \lambda_n(\alpha\|\hat{\theta}\|_{2,1} + (1-\alpha)\|\theta^t\|_{2,1}) \\
= & (1-\alpha)(\frac{1}{n}Q(\theta^t) - \frac{1}{n}Q(\hat{\theta})) + \frac{1}{n}Q(\hat{\theta}) + \frac{\gamma\alpha^2}{2}\|\hat{\Delta}^t\|_2^2 + \alpha\tau_l\frac{\log(p_n)}{n}\|\hat{\Delta}^t\|_{2,1}^2 \\
\leq & (1-\alpha)(\frac{1}{n}Q(\theta^t) - \frac{1}{n}Q(\hat{\theta})) + \frac{1}{n}Q(\hat{\theta}) + \frac{\gamma\alpha^2}{2}\|\hat{\Delta}^t\|_2^2 + \alpha\tau_l\frac{\log(p_n)}{n}(32s\|\hat{\Delta}^t\|_2^2 + 2(\varepsilon+\epsilon)^2) \\
\leq & (1-\alpha)(\frac{1}{n}Q(\theta^t) - \frac{1}{n}Q(\hat{\theta})) + \frac{1}{n}Q(\hat{\theta}) + (\frac{\gamma\alpha^2}{2} + 32\alpha\tau_l\frac{s\log(p_n)}{n})\|\hat{\Delta}^t\|_2^2 \\
& + 2\alpha\tau_l\frac{\log(p_n)}{n}(\varepsilon+\epsilon)^2.
\end{aligned}$$

We set the difference to be denoted by $\bar{\eta}_t = n^{-1}(Q(\theta^t) - Q(\hat{\theta}))$ and combine previous inequalities to obtain the recursive relation as follows,

$$\begin{aligned}
\bar{\eta}_{t+1} \leq & (1-\alpha)\bar{\eta}_t + \kappa_0^{-1}(\frac{\gamma\alpha^2}{2} + 32\alpha\tau_l\frac{s\log(p_n)}{n})(\bar{\eta}_t + 2\tau_l\frac{\log(p_n)}{n}(\varepsilon+\epsilon)^2) + 2\alpha\tau_l\frac{\log(p_n)}{n}(\varepsilon+\epsilon)^2 \\
= & (1 - \alpha + \frac{\gamma\alpha^2}{2\kappa_0} + 32\alpha\tau_l\frac{s\log(p_n)}{\kappa_0 n})\bar{\eta}_t + \kappa_0^{-1}(2\alpha + \gamma\alpha^2 + 64\alpha\tau_l\frac{s\log(p_n)}{n})\tau_l\frac{\log(p_n)}{n}(\varepsilon+\epsilon)^2.
\end{aligned}$$

Since $\gamma > 2\kappa_l$, we can take

$$\alpha = \frac{\kappa_0}{\gamma}(1 - 32\tau_l \frac{s\log(p_n)}{\kappa_0 n}) \in (0,1), \text{ and } 1 - 32\tau_l \frac{s\log(p_n)}{\kappa_0 n} = 1 - o(1),$$

then

$$\bar{\eta}_{t+1} \leq (1 - \frac{\kappa_l}{2\gamma} + 16\tau_l \frac{s\log(p_n)}{\gamma n})\bar{\eta}_t + (2 + \kappa_l + 32\tau_l \frac{s\log(p_n)}{n})\frac{\tau_l}{\gamma}\frac{\log(p_n)}{n}(\varepsilon + \epsilon)^2,$$

with

$$\kappa = 1 - \frac{\kappa_l}{2\gamma} + 16\tau_l \frac{s\log(p_n)}{\gamma n} \text{ and } \beta = (2 + \kappa_l + 32\tau_l \frac{s\log(p_n)}{n})\frac{\tau_l}{\gamma}\frac{\log(p_n)}{n}.$$

we can conclude that

$$\bar{\eta}_{t+1} \leq \kappa\bar{\eta}_t + \beta(\varepsilon + \epsilon)^2.$$

Since $\gamma > 2\kappa_l$ and $n \gtrsim s^2\log(p_n)$, the coefficient $\kappa \in (0,1)$. Therefore, if the iterative algorithm set steps $t \in [T_j, T_{j+1})$, we can show that

$$\bar{\eta}_{j+1} \leq \kappa^{t-T_j}\bar{\eta}_j + \frac{\beta}{1-\kappa}(\varepsilon + \epsilon)^2$$

as claimed. □

Lemma 4.10. *Consider vectors u and $v \in R^{Kp_n}$ double-indexed as $u = (u_{11}, \cdots, u_{kp}, \cdots, u_{Kp_n})$ and $v = (v_{11}, \cdots, v_{kp}, \cdots, v_{Kp_n})$ for $k = 1, 2, \cdots, K$ and $p = 1, 2, \cdots, p_n$. Then*

$$uv \leq \|u\|_{2,1}\|v\|_{2,\infty}.$$

Proof. We apply Hölder's inequality to show that

$$uv \leq \sum_{p=1}^{p_n} \sum_{k=1}^{K} u_{kp} v_{kp} \leq \sum_{p=1}^{p_n} \|u^{(p)}\|_2 \|v^{(p)}\|_2 \leq \|u\|_{2,1} \|v\|_{2,\infty}.$$

□

Lemma 4.11. *(Theorem 6.5 Wainwright 2019) There exist universal constants* $\{c_j\}_{j=0}^2$ *such that for any row-wise sub-Gaussian random matrix* $X \in \mathbb{R}^{n \times p}$ *with parameter* A_0*, the sample covariance matrix* $\hat{\Sigma} = X^T X/n$ *satisfies that*

$$P\big(A_0^{-1} \Big|\Big|\Big| \hat{\Sigma} - \Sigma \Big|\Big|\Big|_2 \geq c_0 \big(\sqrt{\frac{p}{n}} + \frac{p}{n}\big) + \varepsilon\big) \leq c_1 \exp\{-c_2 n \min(\varepsilon, \varepsilon^2)\}.$$

Chapter 5

Discussions and Future Work

In this thesis, we focus on the development and implementation of statistical methods for data integration. To improve the performance of data integration in biomedical research, we apply the pairwise composite likelihood to conduct the joint inference for multiple correlated data sets, which have responses mixed with continuous and discrete variables. We show that the maximum composite likelihood estimators are consistent and asymptotically normally distributed, and the composite statistics provide increased statistical power for joint hypothesis testing.

Multi-task feature learning is commonly used to recover the union support when integrated data sets have divergent dimensionality. Since the existing multi-task learning methods are built under distributional assumptions, we propose to use the composite quasi log-likelihood with mixed $\ell_{2,1}$ regularization to combine tasks of different natures and perform joint learning on multiple correlated heterogeneous tasks. In chapter 3, the method is shown to achieve estimation consistency and model selection consistency in high-dimensional settings.

In chapter 4, we propose the adaptive Huber regression with group-wise feature selection to solve the multi-task learning problem with heavy-tailed error distribution and outlier contamination. The model is different from previous work proposed by Gong et al. [2012].

They focused on the scenario that among the multiple tasks, some outlier tasks possess different sparsity patterns than the other tasks. Besides the mixed $\ell_{2,1}$-norm penalty across all tasks, their objective function includes an additional LASSO penalty to select features in each task. To accommodate outlier tasks in addition to outlier contamination, we can combine the adaptive Huber loss function with the penalty functions used in Gong et al. [2012].

In this thesis, the theoretical conditions for the data are similar to previous high-dimension M-estimation for single-task analysis. In real applications, distinctive measurement methods can produce data with much more complex patterns. From the computational perspective, the proposed multi-task feature learning can apply standardization to the data and improve the selection accuracy. Obozinski et al. [2011] proposed to use a sparsity-overlap function to reflect the sample complexity for the multivariate regression learning, which can be extended to the proposed methods in this thesis. In addition, the data structure can also be more complicated than the settings considered in the chapters above. For example, some predictors can be collected from some but not all tasks. Consequently, the learning tasks have some features not shared by all the tasks. Based on Jalali et al. [2010], Yang et al. [2017], we can use different penalty functions to conduct both group-wise and element-wise feature selection simultaneously. Gao and Carroll [2017] proposed to modify the penalty function by scaling the grouped norm based on the cardinality of the grouped parameters such that

$$\mathcal{R}(\theta) = n\lambda_n \sum_{p=1}^{p_n} \{ \frac{K}{K_p} \sum_{k=1}^{K_p} \theta_{kp}^2 \}^{\frac{1}{2}},$$

where K_p is the number of tasks from which the pth predictor was collected. When the predictor is only collected in one task, the mixed $\ell_{2,1}$ norm is reduced to an ℓ_1 norm.

Since we relax the assumptions of the distribution and moments for the data, the proposed model can be theoretically suitable for a wide variety of real applications, and the integration

is achieved through the composite form of weighted marginal loss functions across all related tasks. With some adjustments to the theoretical conditions, the loss functions can be further extended to estimate the correlation between learning tasks, such that a composite form of generalized estimating equations (GEE) or the pairwise composite form as shown in Chapter 2. In future research, we can extend the multi-task learning to model repeated measurements collected from different biomedical experiments. As a result, the learning tasks need to analyze multiple longitudinal data simultaneously. In addition, when the data are zero-inflated or have a mixture distribution, we can extend the group-wise structure of the parameters based on the research design.

Furthermore, the proposed methods are established based on locally strong convexity or restricted strong convexity. Based on the folded concave regularization framework proposed by [Loh and Wainwright, 2017, Loh, 2017], we can apply mixed regularization to other types of non-convex penalty functions. For example, Gao and Carroll [2017] proposed the group smoothly clipped absolute deviations (SCAD) to perform the joint sparse estimation. In this case, the objective function may not hold the strong convexity condition, which can lead to further investigation.

The multi-task learning algorithm that is publicly available can solve multi-task problems with different types of penalty functions. For example, the R package called Regularized Multi-Task Learning was established by Cao and Schwarz [2022]. However, the heterogeneous tasks are not handled by the existing multi-task learning algorithm, and the robust estimation function is not available for multivariate analysis. Our next step is to develop an R package based on the proposed data integration methods, which can provide an efficient estimation algorithm to solve both heterogeneous and data-contaminated tasks.

Bibliography

A. Agarwal, S. Negahban, and M. J. Wainwright. Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. *The Annals of Statistics*, 40(2):1171–1197, 04 2012a.

A. Agarwal, S. Negahban, and M. J. Wainwright. Fast global convergence of gradient methods for high-dimensional statistical recovery. *The Annals of Statistics*, 40(5):2452 – 2482, 2012b.

J. A. Anderson and J. D. Pemberton. The grouped continuous model for multivariate ordered categorical variables and covariate adjustment. *Biometrics*, 41(4):875–885, 1985.

R. K. Ando and T. Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6(61):1817–1853, 2005.

A. Argyriou, T. Evgeniou, and M. Pontil. Multi-task feature learning. In *Proceedings of the 19th International Conference on Neural Information Processing Systems*, NIPS'06, page 41–48, Cambridge, MA, USA, 2006. MIT Press.

F. R. Bach. Consistency of the group lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9(40):1179–1225, 2008.

H. Bai, Y. Zhong, X. Gao, and W. Xu. Multivariate mixed response model with pairwise composite-likelihood method. *Stats*, 3(3):203–220, 2020.

T. Barrett, D. B. Troup, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, et al. Ncbi geo: archive for functional genomics data sets—10 years on. *Nucleic acids research*, 39(suppl_1):D1005–D1010, 2010.

T. Barrett, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, M. Holko, et al. Ncbi geo: archive for functional genomics data sets—update. *Nucleic acids research*, 41(D1):D991–D995, 2012.

A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Img. Sci.*, 2(1):183–202, Mar. 2009.

D. P. Bertsekas. *Nonlinear Programming*. Athena Scientific, Belmont, MA, 1999.

P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 08 2009.

O. Bousquet. Concentration inequalities for sub-additive functions using the entropy method. In E. Giné, C. Houdré, and D. Nualart, editors, *Stochastic Inequalities and Applications*, pages 213–247, Basel, 2003. Birkhäuser Basel. ISBN 978-3-0348-8069-5.

J. Bradic, J. Fan, and W. Wang. Penalized composite quasi-likelihood for ultrahigh dimensional variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3):325–349, 2011.

C. Cadenas, L. van de Sandt, K. Edlund, M. Lohr, B. Hellwig, R. Marchan, M. Schmidt, J. Rahnenführer, H. Oster, and J. G. Hengstler. Loss of circadian clock gene expression is associated with tumor progression in breast cancer. *Cell Cycle*, 13(20):3282–3291, 2014. PMID: 25485508.

H. Cao and E. Schwarz. *RMTL: Regularized Multi-Task Learning*, 2022. URL `https://CRAN.R-project.org/package=RMTL`. R package version 0.9.9.

R. Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997.

P. J. Catalano. Bivariate modelling of clustered continuous and ordered categorical outcomes. *Statistics in Medicine*, 16(8):883–900, 2022/09/05 1997.

P. J. Catalano and L. M. Ryan. Bivariate latent variable models for clustered discrete and continuous outcomes. *Journal of the American Statistical Association*, 87(419):651–658, 1992. doi: 10.1080/01621459.1992.10475264.

O. Catoni. Challenging the empirical mean and empirical variance: A deviation study. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 48(4):1148 – 1185, 2012. URL `10.1214/11-AIHP454`.

N. R. Council et al. Steps toward large-scale data integration in the sciences: Summary of a workshop. 2010.

D. Cox and D. Hinkley. *Theoretical Statistics*. Chapman and Hall/CRC, 1974.

D. R. Cox. The analysis of multivariate binary data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 21(2):113–120, 1972. ISSN 00359254, 14679876.

D. R. Cox and N. Reid. A note on pseudolikelihood constructed from marginal densities. *Biometrika*, 91(3):729–737, 09 2004.

D. R. Cox and N. Wermuth. Response models for mixed binary and quantitative variables. *Biometrika*, 79(3):441–461, 09 1992. doi: 10.1093/biomet/79.3.441.

G. Dai, U. U. Müller, and R. J. Carroll. Data integration in high dimension with multiple quantiles, 2020. URL `https://arxiv.org/abs/2006.16357`.

A. R. De Leon. Pairwise likelihood approach to grouped continuous model and its extension. *Statistics & Probability Letters*, 75(1):49–57, 2005. doi: https://doi.org/10.1016/j.spl.2005.05.017.

A. R. De Leon and K. C. Carrièger. General mixed-data model: Extension of general location and grouped continuous models. *Canadian Journal of Statistics*, 35(4):533–548, 2022/09/05 2007. doi: https://doi.org/10.1002/cjs.5550350405.

D. B. Dunson. Bayesian latent variable models for clustered mixed outcomes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(2):355–366, 2022/09/05 2000. doi: https://doi.org/10.1111/1467-9868.00236.

R. Edgar, M. Domrachev, and A. E. Lash. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic acids research*, 30(1):207–210, 2002.

A. Eklund, T. E. Nichols, and H. Knutsson. Cluster failure: Why fmri inferences for spatial extent have inflated false-positive rates. *Proceedings of the national academy of sciences*, 113(28):7900–7905, 2016.

K. O. Ekvall and A. J. Molstad. Mixed-type multivariate response regression with covariance estimation. *Statistics in Medicine*, 41(15):2768–2785, 2022. doi: https://doi.org/10.1002/sim.9383. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.9383`.

Y. C. Eldar, P. Kuppinger, and H. Bolcskei. Block-sparse signals: Uncertainty relations and efficient recovery. *IEEE Transactions on Signal Processing*, 58(6):3042–3054, 2010.

C. Faes, M. Aerts, G. Molenberghs, H. Geys, G. Teuns, and L. Bijnens. A high-dimensional joint model for longitudinal outcomes of different nature. *Statistics in Medicine*, 27(22):4408–4427, 2022/09/05 2008. doi: https://doi.org/10.1002/sim.3314.

J. Fan, Q. Li, and Y. Wang. Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(1):247–265, 2017.

J. Fan, H. Liu, Q. Sun, and T. Zhang. I-lamm for sparse learning: Simultaneous control of algorithmic complexity and statistical error. *The Annals of Statistics*, 46 2:814–841, 2018.

J. Fan, W. Wang, and Z. Zhu. A shrinkage principle for heavy-tailed data: High-dimensional robust low-rank matrix recovery. *Annals of statistics*, 49(3):1239–1266, 06 2021. doi: 10.1214/20-aos1980.

E. X. Fang, Y. Ning, and R. Li. Test of significance for high-dimensional longitudinal data. *The Annals of Statistics*, 48(5):2622–2645, 10 2020.

X. Gao and R. J. Carroll. Data integration with high dimensionality. *Biometrika*, 104(2): 251–272, 05 2017.

X. Gao and P. X.-K. Song. Composite likelihood bayesian information criteria for model selection in high-dimensional data. *Journal of the American Statistical Association*, 105 (492):1531–1540, 2010.

X. Gao and Y. Zhong. Fusionlearn: a biomarker selection algorithm on cross-platform data. *Bioinformatics*, 35(21):4465–4468, 2019.

H. Geys, G. Molenberghs, and L. M. Ryan. Pseudolikelihood modeling of multivariate outcomes in developmental toxicology. *Journal of the American Statistical Association*, 94 (447):734–745, 09 1999. doi: 10.1080/01621459.1999.10474176.

V. P. Godambe. An optimum property of regular maximum likelihood estimation. *Ann. Math. Statist.*, 31(4):1208–1211, 12 1960.

D. Gomez-Cabrero, I. Abugessaisa, D. Maier, A. Teschendorff, M. Merkenschlager, A. Gisel, E. Ballestar, E. Bongcam-Rudloff, A. Conesa, and J. Tegnér. Data integration in the era of omics: current and future challenges. *BMC Systems Biology*, 8(2):I1, 2014. doi: 10.1186/1752-0509-8-S2-I1. URL `https://doi.org/10.1186/1752-0509-8-S2-I1`.

P. Gong, J. Ye, and C. Zhang. Robust multi-task feature learning. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 895–903, 2012.

P. Gong, J. Ye, and C. Zhang. Multi-stage multi-task feature learning. *Journal of Machine Learning Research*, 14(55):2979–3010, 2013.

R. V. Gueorguieva and A. Agresti. A correlated probit model for joint modeling of clustered binary and continuous responses. *Journal of the American Statistical Association*, 96(455): 1102–1112, 09 2001. doi: 10.1198/016214501753208762.

C. Hatzis, L. Pusztai, V. Valero, D. J. Booser, L. Esserman, A. Lluch, T. Vidaurre, F. Holmes, E. Souchon, H. Wang, M. Martin, J. Cotrina, H. Gomez, R. Hubbard, J. I. Chacón, J. Ferrer-Lozano, R. Dyer, M. Buxton, Y. Gong, Y. Wu, N. Ibrahim, E. Andreopoulou, N. T. Ueno, K. Hunt, W. Yang, A. Nazario, A. DeMichele, J. O'Shaughnessy, G. N. Hortobagyi, and W. F. Symmans. A genomic predictor of response and survival following taxane-anthracycline chemotherapy for invasive breast cancer. *JAMA*, 305(18):1873–1881, 05 2011.

X. He and Q.-M. Shao. A general bahadur representation of m-estimators and its application to linear regression with nonstochastic designs. *The Annals of Statistics*, 24(6):2608 – 2630, 1996.

M. Hebiri and S. van de Geer. The Smooth-Lasso and other 1+2-penalized methods. *Electronic Journal of Statistics*, 5(none):1184 – 1226, 2011.

A.-S. Heimes, F. Härtner, K. Almstedt, S. Krajnak, A. Lebrecht, M. J. Battista, K. Edlund, W. Brenner, A. Hasenburg, U. Sahin, M. Gehrmann, J. G. Hengstler, and M. Schmidt. Prognostic significance of interferon- and its signaling pathway in early breast cancer depends on the molecular subtypes. *International Journal of Molecular Sciences*, 21(19), 2020. ISSN 1422-0067.

B. Hellwig, J. G. Hengstler, M. Schmidt, M. C. Gehrmann, W. Schormann, and J. Rahnenführer. Comparison of scores for bimodality of gene expression distributions and genome-wide evaluation of the prognostic relevance of high-scoring genes. *BMC Bioinformatics*, 11(1):276, 2010.

P. J. Huber. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 35(1):73 – 101, 1964.

M. Itoh, T. Iwamoto, J. Matsuoka, T. Nogami, T. Motoki, T. Shien, N. Taira, N. Niikura, N. Hayashi, S. Ohtani, K. Higaki, T. Fujiwara, H. Doihara, W. F. Symmans, and L. Pusztai. Estrogen receptor (er) mrna expression and molecular subtype distribution in er-negative/progesterone receptor-positive breast cancers. *Breast Cancer Research and Treatment*, 143(2):403–409, 2014.

A. V. Ivshina, J. George, O. Senko, B. Mow, T. C. Putti, J. Smeds, T. Lindahl, Y. Pawitan, P. Hall, H. Nordgren, J. E. Wong, E. T. Liu, J. Bergh, V. A. Kuznetsov, and L. D. Miller. Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. *Cancer Research*, 66(21):10292–10301, 2006.

A. Jalali, S. Sanghavi, C. Ruan, and P. Ravikumar. A dirty model for multi-task learning. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010.

T. Karn, A. Rody, V. Müller, M. Schmidt, S. Becker, U. Holtrich, and L. Pusztai. Control of dataset bias in combined affymetrix cohorts of triple negative breast cancer. *Genomics Data*, 2:354–356, 2014.

J. T. Kent. Robust properties of likelihood ratio test. *Biometrika*, 69(1):19–27, 1982. ISSN 00063444. URL `http://www.jstor.org/stable/2335849`.

M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes.* Springer-Verlag, New York,, 1991.

O. V. Lepskii. Asymptotically minimax adaptive estimation. i: Upper bounds. optimally adaptive estimates. *Theory of Probability & Its Applications*, 36(4):682–697, 1992.

S. Li and B. Sherwood. *hrqglas: Group Variable Selection for Quantile and Robust Mean Regression*, 2021. URL `https://CRAN.R-project.org/package=hrqglas`. R package version 1.0.1.

Y. Li, W. Xu, and X. Gao. Graphical-model based high dimensional generalized linear models. *Electronic Journal of Statistics*, 15(1):1993 – 2028, 2021.

B. Lindsay. Composite likelihood methods. *Contemporary Mathematics*, 80:220–239, 1988.

B. G. Lindsay, G. Y. Yi, and J. Sun. Issues and strategies in the selection of composite likelihoods. *Statistica Sinica*, 21(1):71–105, 2011.

J. Liu, S. Ji, and J. Ye. Multi-task feature learning via efficient $l_{2,1}$-norm minimization. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, page 339–348, Arlington, Virginia, USA, 2009. AUAI Press.

L.-Z. Liu, F.-X. Wu, and W.-J. Zhang. A group lasso-based method for robustly inferring gene regulatory networks from multiple time-course datasets. *BMC Systems Biology*, 8(3): S1, 2014. doi: 10.1186/1752-0509-8-S3-S1.

Q. Liu, Q. Xu, V. W. Zheng, H. Xue, Z. Cao, and Q. Yang. Multi-task learning for cross-platform sirna efficacy prediction: an in-silico study. *BMC bioinformatics*, 11(1):1–16, 2010.

P.-L. Loh. Statistical consistency and asymptotic normality for high-dimensional robust m-estimators. *The Annals of Statistics*, 45(2):866–896, 04 2017.

P.-L. Loh and M. J. Wainwright. Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *Journal of Machine Learning Research*, 16(19): 559–616, 2015.

P.-L. Loh and M. J. Wainwright. Support recovery without incoherence: A case for nonconvex regularization. *The Annals of Statistics*, 45(6):2455–2482, 12 2017.

K. Lounici, M. Pontil, S. van de Geer, and A. B. Tsybakov. Oracle inequalities and optimal inference under group sparsity. *The Annals of Statistics*, 39(4):2164–2204, 2011.

A. K. Maity, R. J. Carroll, and B. K. Mallick. Integration of survival and binary data for variable selection and prediction: a bayesian approach. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 68(5):1577–1595, 2019. doi: 10.1111/rssc.12377.

E. Mammen. Asymptotics with Increasing Dimension for Robust Regression with Applications to the Bootstrap. *The Annals of Statistics*, 17(1):382 – 400, 1989.

P. Massart and J. Picard. In *Concentration inequalities and model selection*. Springer Berlin, Heidelberg, 2007.

P. McCullagh and J. Nelder. *Generalized Linear Models, Second Edition*. Chapman and Hall/CRC Monographs on Statistics and Applied Probability Series. Chapman & Hall, 1989.

N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*, 37(1):246–270, 02 2009.

C. Meng, B. Kuster, A. C. Culhane, and A. M. Gholami. A multivariate approach to the integration of multi-omics datasets. *BMC bioinformatics*, 15(1):1–13, 2014.

J. S. Najita, Y. Li, and P. J. Catalano. A novel application of a bivariate regression model for binary and continuous outcomes to studies of fetal toxicity. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 58(4):555–573, 2022/09/05 2009.

S. N. Negahban and M. J. Wainwright. Simultaneous support recovery in high dimensions: Benefits and perils of block ℓ_1/ℓ_∞-regularization. *IEEE Transactions on Information Theory*, 57(6):3841–3863, 2011.

S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statist. Sci.*, 27 (4):538–557, 11 2012.

Y. Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161, 2013.

Y. Ning and H. Liu. A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *The Annals of Statistics*, 45(1):158–195, 02 2017.

G. Obozinski, B. Taskar, and M. I. Jordan. Joint covariate selection and joint subspace selection for multiple classification problems. *Statistics and Computing*, 20(2):231–252, 2010.

G. Obozinski, M. J. Wainwright, and M. I. Jordan. Support union recovery in high-dimensional multivariate regression. *The Annals of Statistics*, 39(1):1–47, 02 2011.

I. Olkin and R. F. Tate. Multivariate Correlation Models with Mixed Discrete and Continuous Variables. *The Annals of Mathematical Statistics*, 32(2):448 – 465, 1961. doi: 10.1214/aoms/1177705052.

L. Pace, A. Salvan, and N. Sartori. Adjusting composite likelihood ratio statistics. *Statistica Sinica*, 21(1):129–148, 2011. ISSN 10170405, 19968507.

X. Pan, Q. Sun, and W.-X. Zhou. Iteratively reweighted ℓ_1-penalized robust regression. *Electronic Journal of Statistics*, 15(1):3287 – 3348, 2021.

W.-Y. Poon and S.-Y. Lee. Maximum likelihood estimation of multivariate polyserial and polychoric correlation coefficients. *Psychometrika*, 52(3):409–430, 1987a.

W.-Y. Poon and S.-Y. Lee. Maximum likelihood estimation of multivariate polyserial and polychoric correlation coefficients. *Psychometrika*, 52(3):409–430, 1987b. doi: 10.1007/BF02294364.

S. Portnoy. Asymptotic behavior of m estimators of p regression parameters when p2 / n is large; ii. normal approximation. *The Annals of Statistics*, 13(4):1403–1417, 1985.

A. Rakotomamonjy, R. Flamary, G. Gasso, and S. Canu. $\ell_p - \ell_q$ penalty for sparse linear and sparse multiple kernel multitask learning. *IEEE Transactions on Neural Networks*, 22 (8):1307–1320, 2011.

P. Ravikumar, M. J. Wainwright, and J. D. Lafferty. High-dimensional Ising model selection using 1-regularized logistic regression. *The Annals of Statistics*, 38(3):1287 – 1319, 2010. doi: 10.1214/09-AOS691.

A. Rody, T. Karn, C. Liedtke, L. Pusztai, E. Ruckhaeberle, L. Hanker, R. Gaetje, C. Solbach, A. Ahr, D. Metzler, M. Schmidt, V. Müller, U. Holtrich, and M. Kaufmann. A clinically

relevant gene signature in triple negative and basal-like breast cancer. *Breast Cancer Research*, 13(5):R97, 2011.

A. Rotnitzky and N. P. Jewell. Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. *Biometrika*, 77(3):485–497, 09 1990. ISSN 0006-3444. doi: 10.1093/biomet/77.3.485.

M. Sammel, X. Lin, and L. Ryan. Multivariate linear mixed models for multiple outcomes. *Statistics in Medicine*, 18(17-18):2479–2492, 2022/09/04 1999.

M. D. Sammel, L. M. Ryan, and J. M. Legler. Latent variable models for mixed discrete and continuous outcomes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(3):667–678, 1997. doi: https://doi.org/10.1111/1467-9868.00090.

M. Schmidt, D. Böhm, C. von Törne, E. Steiner, A. Puhl, H. Pilch, H.-A. Lehr, J. G. Hengstler, H. Kölbl, and M. Gehrmann. The humoral immune system has a key prognostic impact in node-negative breast cancer. *Cancer Research*, 68(13):5405–5413, 2008.

A. Skrondal and S. Rabe-Hesketh. Latent variable modelling: A survey*. *Scandinavian Journal of Statistics*, 34(4):712–745, 2007.

Q. Sun, W.-X. Zhou, and J. Fan. Adaptive huber regression. *Journal of the American Statistical Association*, 115(529):254–265, 2020.

A. Teixeira-Pinto and S.-L. T. Normand. Correlated bivariate continuous and binary outcomes: Issues and applications. *Statistics in Medicine*, 28(13):1753–1773, 2009. doi: https://doi.org/10.1002/sim.3588.

K. Tharmaratnam and G. Claeskens. A comparison of robust versions of the aic based on m-, s- and mm-estimators. *Statistics*, 47(1):216–235, 2013. doi: 10.1080/02331888.2011.568120.

K.-H. Thung and C.-Y. Wee. A brief review on multi-task learning. *Multimedia Tools and Applications*, 77(22):29705–29725, 2018.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

U.S. Department of Health and Human Services. Community Health Status Indicators (CHSI) to Combat Obesity, Heart Disease and Cancer 1996-2003. Washington, D.C., USA: US Department of Health and Human Services, 2010.

S. van de Geer and P. Müller. Quasi-likelihood and/or robust estimation in high dimensions. *Statist. Sci.*, 27(4):469–480, 11 2012.

S. van de Geer, P. Bühlmann, Y. Ritov, and R. Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 06 2014.

S. A. van de Geer and P. Bühlmann. On the conditions used to prove oracle results for the lasso. *Electron. J. Statist.*, 3:1360–1392, 2009.

C. Varin. On composite marginal likelihoods. *AStA Advances in Statistical Analysis*, 92(1):1, 2008.

M. J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

L. Wang, B. Peng, and R. Li. A high-dimensional nonparametric multivariate test for mean vector. *Journal of the American Statistical Association*, 110(512):1658–1669, 2015a.

L. Wang, C. Zheng, W. Zhou, and W.-X. Zhou. A new principle for tuning-free huber regression. *Statistica Sinica*, 31(4):2153–2177, 2021.

W. Wang, Y. Liang, and E. P. Xing. Collective support recovery for multi-design multi-response linear regression. *IEEE Transactions on Information Theory*, 61(1):513–534, 2015b.

R. W. M. Wedderburn. Quasi-likelihood functions, generalized linear models, and the gauss-newton method. *Biometrika*, 61(3):439–447, 1974.

S. Wu, X. Gao, and R. J. Carroll. Model selection of generalized estimating equation with divergent model size. *Statistica Sinica*, pages 1–22, 2023. doi: 10.5705/ss.202020.0197.

X. Xu and N. Reid. On the robustness of maximum composite likelihood estimate. *Journal of Statistical Planning and Inference*, 141(9):3047 – 3054, 2011.

P. Yang, P. Zhao, and X. Gao. Robust online multi-task learning with correlative and personalized structures. *IEEE Transactions on Knowledge and Data Engineering*, 29(11): 2510–2521, 2017. doi: 10.1109/TKDE.2017.2703106.

Y. Yang, J. Kang, K. Mao, and J. Zhang. Regression models for mixed poisson and continuous longitudinal data. *Statistics in Medicine*, 26(20):3782–3800, 2007. doi: https://doi.org/10.1002/sim.2776.

G. Y. Yi. Composite likelihood/pseudolikelihood. *Wiley StatsRef: Statistics Reference Online*, pages 1–14, 2014.

G. Y. Yi. *Statistical analysis with measurement error or misclassification: strategy, method and application*. Springer, 2017.

V. J. Yohai and R. A. Maronna. Asymptotic Behavior of M-Estimators for the Linear Model. *The Annals of Statistics*, 7(2):258 – 268, 1979.

N. Yousefi, Y. Lei, M. Kloft, M. Mollaghasemi, and G. C. Anagnostopoulos. Local rademacher complexity-based learning guarantees for multi-task learning. *Journal of Machine Learning Research*, 19(38):1–47, 2018.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B*, 68(1):49–67, 2006.

H. Zhang, D. Liu, J. Zhao, and X. Bi. Modeling hybrid traits for comorbidity and genetic studies of alcohol and nicotine co-dependence. *The Annals of Applied Statistics*, 12(4): 2359–2378, 12 2018. doi: 10.1214/18-AOAS1156.

J. Z. Zhang, W. Xu, and P. Hu. Tightly integrated multiomics-based deep tensor survival model for time-to-event prediction. *Bioinformatics*, 38(12):3259–3266, 2022.

K. Zhang, J. W. Gray, and B. Parvin. Sparse multitask regression for identifying common mechanism of response to therapeutic targets. *Bioinformatics*, 26(12):i97–i105, 2010a.

W. Zhang, F. Li, and L. Nie. Integrating multiple 'omics' analysis for microbial biology: application and methodologies. *Microbiology*, 156(2):287–301, 2010b.

Y. Zhang and Q. Yang. A survey on multi-task learning. *CoRR*, abs/1707.08114, 2017. URL `http://arxiv.org/abs/1707.08114`.

P. Zhao and B. Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2563, Dec. 2006.

J. Zhou, L. Yuan, J. Liu, and J. Ye. A multi-task learning formulation for predicting disease progression. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, page 814–822, New York, NY, USA, 2011. Association for Computing Machinery.