STATISTICAL METHODS FOR COMPLEX AND/OR HIGH DIMENSIONAL DATA

SHANSHAN QIN

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

April 2020

Abstract

This dissertation focuses on the development and implementation of statistical methods for high-dimensional and/or complex data, with an emphasis on p, the number of explanatory variables, larger than n, the number of observations, the ratio of p/n tending to a finite number, and data with outlier observations.

First, we propose a non-negative feature selection and/or feature grouping (nnFSG) method. It deals with a general series of sign-constrained high-dimensional regression problems, which allows the regression coefficients to carry a structure of disjoint homogeneity, including sparsity as a special case. To solve the resulting nonconvex optimization problem, we provide an algorithm that incorporates the difference of convex programming, augmented Lagrange and coordinate descent methods. Furthermore, we show that the aforementioned nnFSG method recovers the oracle estimate consistently, and yields a bound on the mean squared errors (MSE). Besides, we examine the performance of our method by using finite sample simulations and a real protein mass spectrum dataset.

Next, we consider a High-dimensional multivariate ridge regression model under the regime where both p and n are large enough with $p/n \to \kappa(0 < \kappa < \infty)$. On top of that, by using a double leave-one-out method, we develop a nonlinear system of two deterministic equations that characterize the behaviour of M-estimate. Meanwhile, the theoretical results have been confirmed by simulations.

Ultimately, we present matching quantiles M-estimation (MQME), a novel method establishing the relationship between the target response variable and the explanatory variables. MQME extends the matching quantiles estimation (MQE) method to a more general one by replacing the ordinary least-squares (OLS) estimation with an M-estimation, the latter being resistant to outlier observations of the target response. In addition, MQME is combined with an adaptive Lasso penalty so it can select informative variables. We also propose an iterative algorithm to compute the MQME estimate, the consistency of which has been proved, as is the MQE. Numerical experiments on simulated and real datasets demonstrate the efficient performance of our method.

Keywords: coordinate descent, cross-validation, difference convex programming, double leave-one-out, feature grouping, feature selection, high-dimensional, matching quantiles, M-estimation, multivariate regression, non-negative constraint, outlier observations, regularization

Acknowledgements

First of all, I would like to express my deepest gratitude towards my supervisor, Professor Yuehua Wu, for her kindness, patience, and constant encouragement. I have been greatly inspired by her brilliant insights and enthusiasm towards academic research, and her rigorous altitude to every problem. Professor Wu is also an excellent life mentor. She has taught me how to deal with the problems that I met in my life. I am grateful for learning so much from her for the past years, which will benefit me in my whole life.

My special thanks go to Professor Jianhong Wu and Professor Yuejiao Fu for taking the time to be my committee members. My sincere appreciation also goes to all faculty members, staffs and fellow graduate students in the Department of Mathematics and Statistics at York University. I am thankful to Professor Yuejiao Fu and Professor Xin Gao, with whom I was a teaching assistant. Their kindness, generosity, care for students and dedication to teaching inspire me greatly.

I would like to extend my special thanks to my collaborator, Dr. Hao Ding for

his significant contributions. My thanks also go to my academic sisters and brothers, Xiaoying Sun, Xiaoping Shi and Dongwei Wei for their help.

Life always has ups and downs. I am so grateful to have everyone around me in my life. Many thanks go to Pei Yuan, Jasmine Hou, Morgan Wang, and Hang Du, for being true friends, great listeners and making our time at York so enjoyable.

Finally, I deeply thank my parents and sisters for their endless love and continuous supports. The deepest appreciation goes to my husband for his patience, understanding, support, and love in the past few years. This dissertation is dedicated to my lovely son who brings so much happiness every day and motives me to grow up for being a great mother.

Table of Contents

List of Tables

List of Figures

1 Introduction

In the era of data explosion, the volume and the complexity of data are growing faster than ever. This creates opportunities to gain new insights but also demands novel techniques and statistical methods to analyze the data. High-dimensional data, p parallel to or exceeding n, can be found in many areas: genomics, neuroscience, finance, and among others. In data analysis, one of the most important and common questions is whether there is a statistical relationship between response variables and explanatory variables (also called predictors, covariates, regressors). Regression analysis is one of the classic tools to modelize this relationship. In regression models, the response may be univariate or multivariate among which multiple responses are correlated. Another method that one can use is MQE, which aims at finding a linear combination of explanatory variables such that its distribution matches that of the response variable.

Statistical modeling can entail many challenges stemming from the complexity of data. In high-dimensional regression models, there are some commonly stated constraints that should be imposed on the regression coefficients in order to avoid physically impossible or uninterpretable results. For instance, non-negativity is a common constraint especially when modeling non-negative data, say, time measurements, count data, chemical concentrations, intensity values of an image and economical quantities such as prices, incomes and growth rates (Slawski and Hein, 2013). In addition, the regression vector may be sparse in the sense that the majority of elements are zeros. One may also be interested in a situation, which is common in biology (El Karoui, 2018), wherein the regression parameter vector is not sparse but diffuse, i.e., all of the elements are small. Meanwhile, the rapid growth of data volume may bring up another issue, that is, the data sets may encounter outliers due to some uncontrollable factors. Ignoring the existence of outliers and directly applying statistical methods that are not resistant to outliers can lead to inaccurate results and unreasonable scientific conclusions.

This thesis thus incorporates some of these important statistical methods: high-dimensional regression analysis by imposing non-negative constraints on regression coefficients, M-estimation of the high-dimensional multivariate linear model associated with diffuse regression vectors, and MQME for selecting representative portfolios when response observations contain outliers.

1.1 High-dimensional regression analysis

Regression analysis aims at identifying the related explanatory variables of the response and achieving high prediction accuracy (Rekabdarkolaee et al., 2017). For high-dimensional regression problems, regularization methods are of critical importance in a broad sense, and great attention has been devoted to exploring sparseness of regression vectors, including the Bridge regression (Frank and Friedman, 1993), Lasso (Tibshirani, 1996), SCAD (Fan and Li, 2001), elastic net (Zou and Hastie, 2005), adaptive Lasso (Zou, 2006), and MCP (Zhang, 2010). Moreover, extracting one kind of lower-dimensional structure defined by groups has received increasing attention. Literatures can be found in Arnold and Tibshirani (2016); Huang et al. (2009); Jang et al. (2011); She (2010); Shen et al. (2012); Tibshirani and Taylor (2011); Tibshirani et al. (2005); Xiang et al. (2015); Yang et al. (2012); Yuan and Lin (2006); Zhu et al. (2013), among others. Methods introduced in the above articles intend to solve the problems where the regression vectors may carry a structure, which partitions those vectors into disjoint homogeneous subgroups.

The non-negativity constraint on the regression coefficients is an effective regularization technique for a certain class of high-dimensional regression problems. Slawski et al. (2012) proposed non-negative least squares (NNLS)/non-negative least absolute deviation (NNLAD) regression to extract patterns from a raw spectrum. Slawski and

Hein (2013) showed that the performance of NNLS is comparable to that of Lasso in terms of prediction and estimation. Similarly, Meinshausen (2013) confirmed the effectiveness of the sign constraint for sparse recovery if explanatory variables are strongly correlated, and provided an application on the link-level network topography. Koike and Tanoue (2019) extended the results of Slawski and Hein (2013) and Meinshausen (2013) to a more general setup, making them possible of general convex loss functions and non-linearity of responses with respect to explanatory variables. Wen et al. (2015) proposed a projection-based gradient descent method for solving NNLS problems, and then applied it to the inverse problem of constructing a probabilistic Boolean network. Shadmi et al. (2019) investigated NNLS for recovering sparse non-negative vectors from noisy linear and biased measurements, as good as l_1 regularized estimations but without tuning parameters.

Other methods for dealing with such non-negative and sparse structures of regression coefficients combine the regularization techniques with non-negativity constraints, like, non-negative Lasso (Itoh et al., 2016; Wu et al., 2014), non-negative elastic net (Wu and Yang, 2014) and non-negative adaptive Lasso (Yang and Wu, 2016). Esser et al. (2013) added sparsity penalties, which are related to the ratio of l_1 and l_2 norms, to the objective function in an NNLS-type model to solve linear unmixing problems. Hu et al. (2015) applied a non-negative Lasso-based variable selection to identify the important amino acid sites and to evaluate their importance. Mandal and Ma (2016) proposed an efficient regularization path algorithm for generalized linear models with non-negative regression coefficients.

Chapter 2 centers on high-dimensional linear regression problems by imposing non-negative constraints on the regression coefficients. A regularization scheme-based method is thus proposed. In addition to regression coefficients with sparsity and non-negativity, the method is applicable to the cases where those regression coefficients may carry homogeneous subgroups.

1.2 High-dimensional multivariate M-estimation

A well-known method for estimating the regression coefficients is OLS that is mathematically convenient and efficient for normally distributed errors. However, OLS is sensitive to outliers and unstable with respect to deviations from various assumptions. Huber (1964, 1973) thus introduced an M-estimation, which plays an important and complementary role in the development of robust methods. In the past fifty-five years, many procedures based on the M-estimation have been proposed in literature and their asymptotic properties have been investigated. Generally speaking, the asymptotic theories of M-estimation in linear regression models are under three main regimes: (i) the classic regime that allows n to go to infinity with fixed p (see, e.g., Bickel (1975); He and Shao (1996); Huber (1973); Yohai and Maronna (1979)); (ii) the second regime that permits both n and p to go to infinity but restricts $p/n \to 0$, which can be found in the following references, e.g., Bai and Wu (1994); He and Shao (2000); Li et al. (2011); Mammen (1989); Portnoy (1984, 1985); Welsh (1989); Yohai and Maronna (1979); (iii) the most recent regime covers two cases: (a) $p/n \to \kappa$ with $0 < \kappa < 1$ (El Karoui et al., 2013; Lei et al., 2018) and with $0 < \kappa < \infty$ (El Karoui, 2013, 2018); (b) $p >> n$ (Loh, 2017; Loh and Wainwright, 2015; Negahban, 2012). It is noted that El Karoui et al. (2013) proposed a nonlinear system of two deterministic equations to characterize the behavior of M-estimate under random design settings. This topic was also extended to the ridge-regularized M-estimation in El Karoui (2013). Recently, El Karoui (2018) presented rigorous proofs for a general situation, 'elliptical-like' distributed random covariates and heavy-tailed random errors. Lei et al. (2018) investigated asymptotic distributions of each coordinate of the regression M-estimate under the case where random errors are the only source of randomness.

Great work has been done to solve the M-estimation problems of univariate linear regression models. In many statistical applications, however, one may encounter multivariate cases encompassing more than one outcome variable. It seems that the M-estimation problem of multivariate responses has been rarely studied in the literature. In chapter 3, we consider a high-dimensional multivariate regression model under the regime (iii), say, both p and n are large with $p/n \to \kappa(0 < \kappa < \infty)$. We investigate theoretically some asymptotic properties on the ridge-regularized M-estimate.

1.3 Matching quantiles M-estimation

In recent decades, financial market data have become available with increasingly high frequency and dimension. For example, the number of the trades between two major banks could easily be in the magnitude of tens of thousands or more. Backtesting representative portfolios, a subset of all the trades, plays a key role in recalibrating simulations and/or pricing models. Sgouropoulos et al. (2015) argued that the representative portfolios should represent various characteristics of the total portfolios, i.e., risk exposures, sensitivity to the risk factors, etc. Instead of building a regression relationship between the total portfolio and the representative portfolio, Sgouropoulos et al. (2015) thus constructed their distributions matching relationship, that is, the representative portfolio is selected by matching the distribution of the target total portfolio (response) by that of a linear combination of a subset of all trades (covariates). MQE aims to minimize the mean-squared difference between the quantiles of the two distributions across all levels, rather than matching the two distributions directly.

Although MQE achieves high goodness in matching distributions, it is sensitive to outliers due to the fact that it is based on OLS. The existence of outliers deteriorates its performance. As we discussed in Section 1.2, M-estimation has received considerable attention in the literature and its applications in many fields gain great popularity as well. Some recent examples include (1) Lambert-Lacroix and Zwald (2011) proposed an M-estimation by combining Huber's discrepancy with a Lasso penalty, which is resistant to heavy-tailed errors or outliers in observations of the response variable; (2) Zhang et al. (2016) applied an adaptive Huber's M-estimation to the cubature Kalman filter to handle abnormal measurement noise, whose advantages in terms of estimation accuracy, outlier-resistance, and reliability were demonstrated by simulation studies; (3) Ollila et al. (2016) introduced two penalized M-estimation methods for the problem of joint estimation of group covariance matrices.

Since a proper choice of discrepancy function can result in robustness against outliers, one learns that there are statistical procedures, say, M-estimation, one can be used to modify MQE with the purpose of minimizing sensitivity to outliers. We thus propose an enhancement of MQE by replacing OLS with M-estimation in Chapter 4.

1.4 Notations

The following general notations will be used in subsequent chapters. More specialized notations are introduced in context.

- For any $a, b \in \mathbb{R}$, $\min\{a, b\}$ and $\max\{a, b\}$ return the minimum and maximum of a and b, respectively. $\mathrm{sign}(a)$ is the sign of a. $a_+ = a$ if $a \geq 0$, otherwise $a_+ = 0$.

- We denote the l_2-norm, l_1-norm and l_∞-norm of a vector $\boldsymbol{a}$ by $\|\boldsymbol{a}\|$, $\|\boldsymbol{a}\|_1$, $\|\boldsymbol{a}\|_\infty$, respectively. We use $I_{\{\cdot\}}$ and I_p to denote an indicator function and a $p \times p$ identity matrix, respectively. Let $\mathbf{1}_d$ be a $d \times 1$ vector with all elements 1.

- For a square matrix A, we define its smallest and largest eigenvalues by $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$, respectively, and let $\mathrm{tr}(A)$ be the trace of A. $A \succeq (\succ) 0$ means that A is a positive semi-definite (definite) matrix. We write $A_1 \succeq A_2$ if $A_1 - A_2$ is positive semi-definite. If A is an $m \times n$ matrix, we denote $\sqrt{\lambda_{\max}(A^\top A)}$ by $\|A\|_{\max}$ and its vectorization by $\mathrm{vec}(A)$. Define P_A as the projection matrix onto the columns of A. For any two matrices A_1 and A_2, we write their Kronecker product by $A_1 \otimes A_2$.

- For any set $\mathcal{A}$, $|\mathcal{A}|$ and $\mathcal{A}^c$ denote the size and the complement of $\mathcal{A}$, respectively. For any $\mathcal{B} \subset \mathcal{A}$, $\mathcal{A} \backslash \mathcal{B} = \mathcal{A} \cap \mathcal{B}^c$.

- For some generic random variable ξ, we use $\mathcal{L}(\xi)$ to denote its distribution and $F_\xi(\cdot)$ and $f_\xi(\cdot)$ to denote its distribution function and probability density function, respectively. We use $\Phi(\cdot)$ to denote the cumulative distribution function of the standard normal distribution. $Q_\xi(\alpha)$ is the αth quantile of random variable ξ, i.e., $P\{\xi \leq Q_\xi(\alpha)\} = \alpha, \text{for } \alpha \in [0, 1]$. Similarly, we write $Q_{n,\xi}(i/n)$ to denote the i/nth sample quantile of random variable ξ.

- For any random vector $\boldsymbol{\xi}$, we write $\boldsymbol{\xi} \sim (\mathbf{0}, \Sigma)$ when $\boldsymbol{\xi}$ is distributed according to a distribution with mean vector $\mathbf{0}$ and covariance matrix Σ.

- For any vector-valued function $\boldsymbol{g} : \mathbb{R}^m \to \mathbb{R}^m$, we denote the derivative of $\boldsymbol{g}(\boldsymbol{x})$ by $\nabla \boldsymbol{g}(\boldsymbol{x})$, an $m \times m$ matrix, for $\boldsymbol{x} \in \mathbb{R}^m$. If $\boldsymbol{g}(\boldsymbol{x})$ is invertible, we write the inverse function of $\boldsymbol{g}(\boldsymbol{x})$ by $\boldsymbol{g}^{-1}(\boldsymbol{x})$.

- For any twice differentiable function $\rho : \mathbb{R}^m \to \mathbb{R}$, we denote the first and second derivative of $\rho(\boldsymbol{x})$ by $\boldsymbol{\psi}(\boldsymbol{x})$, and $\nabla \boldsymbol{\psi}(\boldsymbol{x})$, for $\boldsymbol{x} \in \mathbb{R}^m$, respectively. If $\nabla \boldsymbol{\psi}(\boldsymbol{x})$ is positive semi-definite, we write $\nabla \boldsymbol{\psi}^{1/2}(\boldsymbol{x})$ as its square root.

- Denote the convergence in probability by '$\xrightarrow{p}$', the convergence almost surely by '$\xrightarrow{a.s.}$', and the convergence in distribution by '$\xrightarrow{\mathcal{D}}$',

2 Sign constrained feature selection and/or grouping via regularization method

2.1 Introduction

In recent decades, high-dimensional problems appear in many fields due to the explosion of massive data. One standard tool to perform data analysis statistically is through a linear regression model, i.e.,

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \epsilon_i \quad (\ i = 1, \dots, n), \tag{2.1}$$

where y_i are response observations, $\boldsymbol{x}_i = (x_{i1}, \dots, x_{ip})^\top$ are p-dimensional vectors of predictors, $\boldsymbol{\beta} \in \mathbb{R}^p$ is a vector of unknown regression coefficients, ϵ_i are random errors, and $\boldsymbol{x}_i$ are independent of ϵ_i. In the high-dimensional setting, p is at least of the same order of magnitude as n, say $p = O(n)$ (p is not fixed), or even $p >> n$, in which case $\boldsymbol{\beta}$ may be sparse (Slawski and Hein, 2013).

From a practical viewpoint, due to the inherent physical characteristics of systems under investigation, there are some commonly stated constraints, say, non-negativity, that should be imposed on the regression parameters to avoid physically impossible and uninterpretable results. We propose in this chapter a regularization scheme-based method to deal with the non-negative feature selection problem. Additionally, our method is applicable to the feature grouping cases where those regression coefficients may carry homogeneous subgroups within which the elements are similar or identical. Note that, throughout this study, we only consider non-negative constraints since one can replace the covariates that are imposed to be negative coefficients by their negative counterparts (Meinshausen, 2013).

In light of Shen et al. (2012), we initially introduce the nnFSG in its constrained form, followed by the regularized one. A hybrid algorithm, combing with the difference convex programming, augmented Lagrange and coordinate descent, is provided to solve the non-convex optimization problem. We investigate some theoretical properties of our nnFSG estimates in terms of grouping consistency and bounds on MSE. We stress that feature selection can be regarded as a special case where only a group of zeros is included. Our finite sample simulations show that the proposed method is superior to some other methods that are available for computing non-negative estimates. We also examine the performance of our proposed method using a real protein mass spectrum dataset.

The rest of this chapter is organized as follows. In Section 2.2, we introduce the constrained form of nnFSG as well as the theoretical properties of the constrained estimate. We describe the regularized form of the nnFSG in Section 2.3, along with an algorithm to solve the resulting non-convex optimization problem and theoretical properties of the estimate. We present the numerical studies in Section 2.4. The proofs of these lemmas and theorems are relegated to appendix A.

2.2 Constrained nnFSG

2.2.1 The formulation of constrained nnFSG

Consider the linear regression model (2.1), where $\boldsymbol{\beta}$ might be sparsity with non-negative constraints. Suppose that $\boldsymbol{\beta}^0$ is the true regression vector. The non-negative feature selection (nnFS) is formulated by the constrained least squares criterion

$$\min_{\boldsymbol{\beta} \geq \mathbf{0}} \frac{1}{2n} \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2, \tag{2.2}$$

subject to

$$\sum_{j=1}^{p} \min\left\{ \frac{|\beta_j|}{\tau}, 1 \right\} \leq s_1, \tag{2.3}$$

where $s_1 (> 0)$ is a tuning parameter that controls feature selection. $\tau > 0$, a threshold parameter, determines when a small regression coefficient should be penalized.

In particular, the unknown vector $\boldsymbol{\beta}$ may carry a structure with disjoint homogeneous subgroups within which the coordinates are identical or similar. Let the number of disjoint subgroups be $K+1$ ($K \leq p-1$), and denote the coefficients index of k-th group by $\mathcal{G}_k$ satisfying $\cup_k \mathcal{G}_k = \{1, 2, \dots, p\}$ and $\cap_k \mathcal{G}_k = \emptyset$. Denote $\boldsymbol{\beta} = (\boldsymbol{\beta}_0^\top, \boldsymbol{\beta}_1^\top, \dots, \boldsymbol{\beta}_K^\top)^\top$, where $\boldsymbol{\beta}_k = \alpha_k \mathbf{1}_{|\mathcal{G}_k|}$, $\alpha_0 = 0$ and $\alpha_k > 0$ for $k = 1, \dots, K$. In light of Shen et al. (2012), the constrained nnFSG is formulated by solving the problem (2.2) subjecting to

$$\sum_{j=1}^{p} \min\left\{\frac{|\beta_j|}{\tau}, 1\right\} \leq s_1, \text{and} \sum_{(j,j')\in\varepsilon} \min\left\{\frac{|\beta_j - \beta_{j'}|}{\tau}, 1\right\} \leq s_2, \tag{2.4}$$

where $\varepsilon = \{(j, j') : j < j', j, j' = 1, \dots, p\}$, an arbitrary undirected graph. The tuning parameter, $s_2 (> 0)$, controls feature grouping. $\tau (> 0)$ also determines when a small difference between two coefficients should be penalized. More details on the constraints of (2.4) can be referred to Shen et al. (2012, 2013). Note that nnFSG is reduced to nnFS if $K = p - 1$. Throughout this chapter, we thus only consider the nnFSG problem. Our goal is to estimate $\boldsymbol{\beta}$ or equivalently, $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_K)^\top$ and $\mathcal{G} = (\mathcal{G}_0, \mathcal{G}_1, \dots, \mathcal{G}_K)^\top$. A solution to (2.2) subjecting to (2.4) is referred to as a constrained nnFSG estimate, denoted by $\hat{\boldsymbol{\beta}}^{cons}$.

2.2.2 Theoretical properties of the constrained nnFSG estimate

We denote the true grouping by $\mathcal{G}^0 = (\mathcal{G}_0^0, \mathcal{G}_1^0, \ldots, \mathcal{G}_{K^0}^0) = (\mathcal{G}_0^0, \mathcal{G}_0^{0c})$, and the true regression parameter for the group k by α_k^0 for $k = 1, \ldots, K^0$, where K^0+1 is the true grouping number. Then $\boldsymbol{\beta}^0$ can be written as $\boldsymbol{\beta}^0 = (0\mathbf{1}_{|\mathcal{G}_0^0|}^\top, \alpha_1^0\mathbf{1}_{|\mathcal{G}_1^0|}^\top, \ldots, \alpha_{K^0}^0\mathbf{1}_{|\mathcal{G}_{K^0}^0|}^\top)^\top$. Denote $\boldsymbol{\alpha}^0 = (\alpha_1^0, \ldots, \alpha_{K^0}^0)^\top$, and $Z_{\mathcal{G}_0^{0c}} = (X_{\mathcal{G}_1^0}\mathbf{1}_{|\mathcal{G}_1^0|}, \ldots, X_{\mathcal{G}_{K^0}^0}\mathbf{1}_{|\mathcal{G}_{K^0}^0|})$. Now, we define the oracle estimate,

$$\hat{\boldsymbol{\beta}}^{ora} = (\hat{\beta}_1^{ora}, \ldots, \hat{\beta}_p^{ora})^\top = (0\mathbf{1}_{|\mathcal{G}_0^0|}^\top, \hat{\alpha}_1^{ora}\mathbf{1}_{|\mathcal{G}_1^0|}^\top, \ldots, \hat{\alpha}_{K^0}^{ora}\mathbf{1}_{|\mathcal{G}_{K^0}^0|}^\top)^\top,$$

where $\hat{\boldsymbol{\alpha}}^{ora} = (\hat{\alpha}_1^{ora}, \ldots, \hat{\alpha}_{K^0}^{ora})^\top$, satisfying that

$$\hat{\boldsymbol{\alpha}}^{ora} = \arg\min_{\boldsymbol{\alpha}} \frac{1}{2n}\|\boldsymbol{y} - Z_{\mathcal{G}_0^{0c}}^\top\boldsymbol{\alpha}\|^2, \;\; \alpha_k > 0, \; k = 1, \ldots, K^0.$$

We denote the OLS estimate by

$$\hat{\boldsymbol{\beta}}^{ols} = (\hat{\beta}_1^{ols}, \ldots, \hat{\beta}_p^{ols})^\top = (0\mathbf{1}_{|\mathcal{G}_0^0|}^\top, \hat{\alpha}_1^{ols}\mathbf{1}_{|\mathcal{G}_1^0|}^\top, \ldots, \hat{\alpha}_{K^0}^{ols}\mathbf{1}_{|\mathcal{G}_{K^0}^0|}^\top)^\top,$$

where $\hat{\boldsymbol{\alpha}}^{ols} = (\hat{\alpha}_1^{ols}, \ldots, \hat{\alpha}_{K^0}^{ols})^\top = (Z_{\mathcal{G}_0^{0c}}^\top Z_{\mathcal{G}_0^{0c}})^{-1}Z_{\mathcal{G}_0^{0c}}^\top\boldsymbol{y}$ with $Z_{\mathcal{G}_0^{0c}}^\top Z_{\mathcal{G}_0^{0c}}$ invertible. Note that both $\hat{\boldsymbol{\beta}}^{ora}$ and $\hat{\boldsymbol{\beta}}^{ols}$ are defined based on the true grouping $\mathcal{G}^0$.

Before proceeding, we provide two metrics proposed by Shen et al. (2012) and Zhu et al. (2013), which reflect the model's difficulty. One level of the difficulty is given by,

$$C_{\min} = \min_{\mathcal{G}\in\mathcal{T}} \frac{\|(I - P_{Z_{\mathcal{G}_0^c}})X\boldsymbol{\beta}^0\|^2}{n\max\{|\mathcal{G}_0\backslash\mathcal{G}_0^0|, 1\}},$$

where $\mathcal{G} = (\mathcal{G}_0, \mathcal{G}_0^c) = (\mathcal{G}_0, \mathcal{G}_1, \ldots, \mathcal{G}_K)$, $P_{Z_{\mathcal{G}_0^c}} = Z_{\mathcal{G}_0^c}(Z_{\mathcal{G}_0^c}^\top Z_{\mathcal{G}_0^c})^{-1} Z_{\mathcal{G}_0^c}^\top$ with $Z_{\mathcal{G}_0^c} = (X_{\mathcal{G}_1}\mathbf{1}_{|\mathcal{G}_1|}, \ldots, X_{\mathcal{G}_K}\mathbf{1}_{|\mathcal{G}_K|})$, $\mathcal{T} = \{\mathcal{G} \neq \mathcal{G}^0 : \sum_{j=1}^p I_{\{\beta_j > 0\}} \leq s_1^0, \sum_{(j,j')\in\varepsilon} I_{\{\beta_j \neq \beta_{j'}\}} \leq s_2^0\}$, a constrained set corresponding to (2.4) with $s_1^0 = |\mathcal{S}| = p - |\mathcal{G}_0^0|$ and $s_2^0 = \sum_{(j,j')\in\varepsilon} I_{\{\beta_j^0 \neq \beta_{j'}^0\}}$. We remark that $C_{\min}$ defines the degree of separation between $\mathcal{G}_0^0$ and $\mathcal{G}_0$ of a least favorable candidate model in the l_2-norm. Another one is the resolution level of the true regression coefficients,

$$\gamma_{\min} = \min_{\{j,j' \in \mathcal{G}_0^{0^c}, (j,j')\in\varepsilon\}} \{\beta_j^0, |\beta_j^0 - \beta_{j'}^0|\}.$$

The smaller the values of $C_{\min}$ and $\gamma_{\min}$, the more difficult the situation. Denote $\bar{K} = \max_{1\leq i\leq s_1^0} K_i^*/i$, where $K_i^* = \max_{\{\mathcal{G}\in\mathcal{T}, |\mathcal{G}_0 \backslash \mathcal{G}_0^0| = i\}} K(\mathcal{G}_0^c)$, and $K(\mathcal{G}_0^c)$ is the grouping number of $\mathcal{G}_0^c$. Let $\bar{T} = \max_{1\leq i\leq s_1^0} \log T_i/i$, where $T_i = \max_{\{\mathcal{G}\in\mathcal{T}, |\mathcal{G}_0\backslash\mathcal{G}_0^0|=i\}} |T_{\mathcal{G}_0^c}|$, and $T_{\mathcal{G}_0^c} = \{\mathcal{G} = (\mathcal{G}_0^*, \mathcal{G}_1, \ldots, \mathcal{G}_K) \in \mathcal{T} : \mathcal{G}_0^* = \mathcal{G}_0^0\}$, a set of groupings indexed by the sets of positive coefficients. More details on $C_{\min}$, $\gamma_{\min}$, $\bar{T}$ and $\bar{K}$ can be referred to Shen et al. (2012) and Zhu et al. (2013).

Now, we make the following assumptions.

(A1) $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$, $i = 1, \ldots, n$.

(A2) There exists a constant c_0 such that $\lambda_{\min}\left(n^{-1} Z_{\mathcal{G}_0^{0c}}^\top Z_{\mathcal{G}_0^{0c}}\right) \geq c_0 > 0$.

(A3) For the same constant c_0 as in (A2), $\gamma_{\min} > [2\sigma^2 \log\{2nK^0/(2\pi)^{1/2}\}/(nc_0)]^{1/2}$.

Lemma 2.1 *Under the assumptions (A1)-(A3), it holds that*

$$P(\hat{\boldsymbol{\beta}}^{ora} \neq \hat{\boldsymbol{\beta}}^{ols}) = O\left(\frac{1}{n(\log n)^{1/2}}\right).$$

In Lemma 2.1, we show that $\min_{1\leq k\leq K^0} \hat{\alpha}_k^{ols} > 0$ with probability at least $1 - 2K^0\left\{1 - \Phi\left([2\log\{2nK^0/(2\pi)^{1/2}\}]^{1/2}\right)\right\}$, which implies that with the same probability, $\hat{\boldsymbol{\beta}}^{ora} = \hat{\boldsymbol{\beta}}^{ols}$.

Theorem 2.1 *Under the assumptions (A1)-(A3), it follows that, for any* $0 < \tau \leq \sigma\{\log p/[2np\lambda_{\max}(X^\top X)]\}^{1/2}$,

$$P\left(\hat{\boldsymbol{\beta}}^{cons} \neq \hat{\boldsymbol{\beta}}^{ora}\right) \leq (\exp(1)+1)\exp\left(-\frac{n}{10\sigma^2}\left\{C_{\min} - \frac{10\sigma^2}{n}\left(3\log p + \bar{T} + \frac{\bar{K}}{2}\right)\right\}\right) + \frac{c}{n(\log n)^{1/2}}.$$

If, additionally, $C_{\min} \geq 10\sigma^2 n^{-1}\left(\log n + \log\log n/2 + 3\log p + \bar{T} + \bar{K}/2\right)$, *then*

(1) $P\left(\hat{\boldsymbol{\beta}}^{cons} \neq \hat{\boldsymbol{\beta}}^{ora}\right) = O\left(n^{-1}(\log n)^{-1/2}\right)$;

(2) $n^{-1}E\left\|X\hat{\boldsymbol{\beta}}^{cons} - X\boldsymbol{\beta}^0\right\|^2 = n^{-1}K^0\sigma^2(1+o(1))$.

$\hat{\boldsymbol{\beta}}^{cons}$ yields a consistent recovery of $\hat{\boldsymbol{\beta}}^{ora}$, and also generates a bounded MSE. Since $\hat{\boldsymbol{\beta}}^{ora}$ is defined based on the true grouping, Theorem 2.1 implies that $\hat{\boldsymbol{\beta}}^{cons}$ identifies the true grouping consistently.

2.3 Regularized nnFSG

2.3.1 The formulation of the regularized nnFSG

By Lemma 1 of Shen et al. (2012), the minimizer of

$$\frac{1}{2n}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2 \text{ subject to } \sum_{j=1}^{p} \min\left\{\frac{|\beta_j|}{\tau}, 1\right\} \leq s_1, \sum_{(j,j')\in\varepsilon} \min\left\{\frac{|\beta_j - \beta_{j'}|}{\tau}, 1\right\} \leq s_2,$$

is a local minimizer of

$$f(\boldsymbol{\beta}) = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2 + \lambda_1 p_1(\boldsymbol{\beta}) + \lambda_2 p_2(\boldsymbol{\beta}),$$

where $p_1(\boldsymbol{\beta}) = \sum_{j=1}^{p} \min\{|\beta_j|/\tau, 1\}$, and $p_2(\boldsymbol{\beta}) = \sum_{(j,j')\in\varepsilon} \min\{|\beta_j - \beta_{j'}|/\tau, 1\}$. We impose non-negative constraints on $\boldsymbol{\beta}$, i.e.,

$$\min_{\boldsymbol{\beta}} f(\boldsymbol{\beta}) \text{ subject to } \beta_j \geq 0, j = 1, \ldots, p. \tag{2.5}$$

Using the penalty method in Chapter 13 of Luenberger and Ye (2015), the regularized version of (2.5) is thus given by

$$\min_{\boldsymbol{\beta}} \frac{1}{2n}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2 + \lambda_1 p_1(\boldsymbol{\beta}) + \lambda_2 p_2(\boldsymbol{\beta}) + \lambda_3 p_3(\boldsymbol{\beta}), \tag{2.6}$$

where $p_3(\boldsymbol{\beta}) = \sum_{j=1}^{p}(\min\{\beta_j, 0\})^2$. $\lambda_1(> 0), \lambda_2(\geq 0)$ correspond to s_1, s_2 in (2.4), respectively. $\lambda_3(> 0)$ controls the shrinkage speed of negative regression coefficients. Obviously, by setting $\lambda_2 = 0$, (2.6) reduces to the regularized nnFS, which solves the feature selection problems with non-negative constraints on the regression coefficients. A solution to (2.6), denoted by $\hat{\boldsymbol{\beta}}$, is referred to as a nnFSG estimate.

Denote $S(\boldsymbol{\beta}) = (2n)^{-1} \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2 + \lambda_1 p_1(\boldsymbol{\beta}) + \lambda_2 p_2(\boldsymbol{\beta}) + \lambda_3 p_3(\boldsymbol{\beta})$. Since $S(\boldsymbol{\beta})$ is non-convex, the difference of convex programming is thus applied to solve (2.6). Our main technical contribution is to extend the algorithm in Shen et al. (2012) to a more general one by adding another penalty term $p_3(\boldsymbol{\beta})$, which, together with $p_1(\boldsymbol{\beta})$, controls the non-negativity of the regression coefficients.

Firstly, decompose $S(\boldsymbol{\beta})$ into the difference of two convex functions as follows,

$$S(\boldsymbol{\beta}) = S_1(\boldsymbol{\beta}) - S_2(\boldsymbol{\beta}), \tag{2.7}$$

where the convex functions $S_1(\boldsymbol{\beta})$ and $S_2(\boldsymbol{\beta})$ are given respectively by

$$\begin{aligned} S_1(\boldsymbol{\beta}) &= \frac{1}{2n} \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2 + \frac{\lambda_1}{\tau} \sum_{j=1}^{p} |\beta_j| + \frac{\lambda_2}{\tau} \sum_{(j,j')\in\varepsilon} |\beta_j - \beta_{j'}| + \lambda_3 \sum_{j=1}^{p} \beta_j^2, \\ S_2(\boldsymbol{\beta}) &= \frac{\lambda_1}{\tau} \sum_{j=1}^{p} (|\beta_j| - \tau)_+ + \frac{\lambda_2}{\tau} \sum_{(j,j')\in\varepsilon} (|\beta_j - \beta_{j'}| - \tau)_+ + \lambda_3 \sum_{j=1}^{p} ((\beta_j)_+)^2. \end{aligned}$$

Define $\boldsymbol{\eta} = (|\beta_1|, \ldots, |\beta_p|, |\beta_{12}|, \ldots, |\beta_{1p}|, \ldots, |\beta_{(p-1)p}|, \beta_1^2, \ldots, \beta_p^2)^\top$, where $\beta_{jj'} = \beta_j - \beta_j'$, $(j, j') \in \varepsilon$. Then, $S_2(\boldsymbol{\beta})$ can be expressed to

$$\tilde{S}_2(\boldsymbol{\eta}) = \frac{\lambda_1}{\tau} \sum_{j=1}^{p} (|\beta_j| - \tau)_+ + \frac{\lambda_2}{\tau} \sum_{(j,j')\in\varepsilon} (|\beta_{jj'}| - \tau)_+ + \lambda_3 \sum_{j=1}^{p} \beta_j^2 I_{\{\beta_j \geq 0\}}.$$

Approximate $\tilde{S}_2(\boldsymbol{\eta})$ by its affine minorization $\tilde{S}_2(\boldsymbol{\eta}^*) + \langle \boldsymbol{\eta} - \boldsymbol{\eta}^*, \partial \tilde{S}_2(\boldsymbol{\eta}^*) \rangle$ at a neighbourhood of $\boldsymbol{\eta}^* \in \mathbb{R}^{(p^2+3p)/2}$, where $\partial \tilde{S}_2(\boldsymbol{\eta})$ is the first derivative of $\tilde{S}_2(\boldsymbol{\eta})$ with respect to $\boldsymbol{\eta}$; $\langle \cdot, \cdot \rangle$ is the inner product. Now we construct a sequence of approximations of $S_2(\boldsymbol{\beta})$ iteratively. At the m-th iteration, we replace $S_2(\boldsymbol{\beta})$ by $S_2^{(m)}(\boldsymbol{\beta}) = \tilde{S}_2^{(m)}(\boldsymbol{\eta}) = \tilde{S}_2(\hat{\boldsymbol{\eta}}^{(m-1)}) + \langle \boldsymbol{\eta} - \hat{\boldsymbol{\eta}}^{(m-1)}, \partial \tilde{S}_2(\hat{\boldsymbol{\eta}}^{(m-1)}) \rangle$. Specifically,

$$\begin{aligned} S_2^{(m)}(\boldsymbol{\beta}) &= S_2(\hat{\boldsymbol{\beta}}^{(m-1)}) + \frac{\lambda_1}{\tau} \sum_{j=1}^{p} I_{\{|\hat{\beta}_j^{(m-1)}| \geq \tau\}} \left(|\beta_j| - |\hat{\beta}_j^{(m-1)}| \right) \\ &+ \frac{\lambda_2}{\tau} \sum_{(j,j') \in \varepsilon} I_{\{|\hat{\beta}_j^{(m-1)} - \hat{\beta}_{j'}^{(m-1)}| \geq \tau\}} \left(|\beta_j - \beta_{j'}| - |\hat{\beta}_j^{(m-1)} - \hat{\beta}_{j'}^{(m-1)}| \right) \\ &+ \lambda_3 \sum_{j=1}^{p} I_{\{\hat{\beta}_j^{(m-1)} \geq 0\}} \left(\beta_j^2 - (\hat{\beta}_j^{(m-1)})^2 \right). \end{aligned}$$

Finally, an approximation to $S(\boldsymbol{\beta})$ in (2.7) at the m-th iteration can be obtained by $S^{(m)}(\boldsymbol{\beta}) = S_1(\boldsymbol{\beta}) - S_2^{(m)}(\boldsymbol{\beta})$, which formulates the following subproblem,

$$\begin{aligned} \min_{\boldsymbol{\beta}} \frac{1}{2n} \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2 + \frac{\lambda_1}{\tau} \sum_{j \in \mathcal{F}^{(m-1)}} |\beta_j| + \frac{\lambda_2}{\tau} \sum_{(j,j') \in \varepsilon^{(m-1)}} |\beta_j - \beta_{j'}| \\ + \lambda_3 \sum_{j \in \mathcal{N}^{(m-1)}} \beta_j^2, \end{aligned} \tag{2.8}$$

where

$$\begin{aligned} \mathcal{F}^{(m-1)} &= \{j : |\hat{\beta}_j^{(m-1)}| < \tau\}, \\ \varepsilon^{(m-1)} &= \{(j,j') : j < j', |\hat{\beta}_j^{(m-1)} - \hat{\beta}_{j'}^{(m-1)}| < \tau\}, \\ \mathcal{N}^{(m-1)} &= \{j : \hat{\beta}_j^{(m-1)} < 0\}. \end{aligned} \tag{2.9}$$

How to efficiently solve the subproblem (2.8) plays a key role in solving the problem (2.6). Though we can apply quadratic programming to solve the subproblem (2.8), it is inefficient for large-scale problems.

2.3.2 Algorithm

In light of Shen et al. (2012), an algorithm integrated with augmented Lagrange and coordinate descent methods is developed to solve the subproblem (2.8). We convert the subproblem (2.8) with linear constraints to its unconstrained version through slack variables $\beta_{jj'} = \beta_j - \beta_{j'}$. Define $\boldsymbol{\xi} = (\beta_1, \dots, \beta_p, \beta_{12}, \dots, \beta_{1p}, \dots, \beta_{(p-1)p})^\top$. Then an augmented equivalent problem of (2.8) is given, i.e.,

$$\min_{\boldsymbol{\xi}} \frac{1}{2n} \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2 + \frac{\lambda_1}{\tau} \sum_{j \in \mathcal{F}^{(m-1)}} |\beta_j| + \frac{\lambda_2}{\tau} \sum_{(j,j') \in \varepsilon^{(m-1)}} |\beta_{jj'}| \\ + \lambda_3 \sum_{j \in \mathcal{N}^{(m-1)}} \beta_j^2. \tag{2.10}$$

For (2.10), the augmented Lagrange is employed to solve its equivalent unconstrained problem iteratively with respect to t at the m-th iteration. Denote $\tilde{S}^{(m)}(\boldsymbol{\xi}) = (2n)^{-1} \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2 + \lambda_1 \tau^{-1} \sum_{j \in \mathcal{F}^{(m-1)}} |\beta_j| + \lambda_2 \tau^{-1} \sum_{(j,j') \in \varepsilon^{(m-1)}} |\beta_{jj'}| + \lambda_3 \sum_{j \in \mathcal{N}^{(m-1)}} \beta_j^2$. In the t-th iteration, we minimize

$$\bar{S}^{(m)}(\boldsymbol{\xi}) = \tilde{S}^{(m)}(\boldsymbol{\xi}) + \sum_{(j,j') \in \varepsilon^{(m-1)}} \tau_{jj'}^{(t)} (\beta_j - \beta_{j'} - \beta_{jj'}) \\ + \frac{1}{2} \nu^{(t)} \sum_{(j,j') \in \varepsilon^{(m-1)}} (\beta_j - \beta_{j'} - \beta_{jj'})^2, \tag{2.11}$$

where $\tau_{jj'}^{(t)}$, $\nu^{(t)}$ are Lagrange multipliers. Update $\tau_{jj'}$ and ν by

$$\tau_{jj'}^{(t+1)} = \tau_{jj'}^{(t)} + \nu^{(t)} (\hat{\beta}_j^{(m,t)} - \hat{\beta}_{j'}^{(m,t)} - \hat{\beta}_{jj'}^{(m,t)}) \quad \text{and} \quad \nu^{(t+1)} = \rho \nu^{(t)}, \tag{2.12}$$

where ρ controls the speed of convergence. To speed up convergence, ρ is chosen larger than 1.

We use the coordinate descent method to compute $\hat{\boldsymbol{\xi}}^{(m,t)}$ in terms of (2.11). For each component of $\boldsymbol{\xi}$, we fix the other components at their current values. Set an initial value $\hat{\boldsymbol{\xi}}^{(m,0)} = \hat{\boldsymbol{\xi}}^{(m-1)}$, where $\hat{\boldsymbol{\xi}}^{(m-1)}$ is the solution of the subproblem (2.8). Then update $\hat{\boldsymbol{\xi}}^{(m,t)}$ by the following formulas, $t = 1, 2, \ldots$.

(I) Given $\hat{\beta}_l^{(m,t-1)}$, updating $\hat{\beta}_l^{(m,t)}, (l = 1, 2, \ldots, p)$ by:

$$\hat{\beta}_l^{(m,t)} = \alpha^{-1}\gamma, \tag{2.13}$$

where

$$\alpha = \frac{1}{n}\sum_{i=1}^{n} x_{il}^2 + 2\lambda_3 I_{\{\hat{\beta}_l^{(m-1)}<0\}} + \nu^{(t)} \left| j' : (l, j') \in \varepsilon^{(m-1)} \text{ or } (j', l) \in \varepsilon^{(m-1)} \right|,$$

and

$$\gamma = n^{-1}\sum_{i=1}^{n} x_{il} b_{i,l}^{(m,t)} - \sum_{(l,j')\in\varepsilon^{(m-1)}} \tau_{lj'}^{(t)} + \nu^{(t)} \sum_{(l,j')\in\varepsilon^{(m-1)}} \left(\hat{\beta}_{j'}^{(m,t)} + \hat{\beta}_{lj'}^{(m,t)}\right),$$

if $|\hat{\beta}_l^{(m-1)}| > \tau$, i.e., $l \in F^{(m-1)^c}$; and

$$\gamma = ST\left(\frac{1}{n}\sum_{i=1}^{n} x_{il} b_{i,l}^{(m,t)} - \sum_{(l,j')\in\varepsilon^{(m-1)}} \tau_{lj'}^{(t)} + \nu^{(t)} \sum_{(l,j')\in\varepsilon^{(m-1)}} \left(\hat{\beta}_{j'}^{(m,t)} + \hat{\beta}_{lj'}^{(m,t)}\right), \frac{\lambda_1}{\tau}\right),$$

if $0 < |\hat{\beta}_l^{(m-1)}| < \tau$, i.e., $l \in F^{(m-1)}$. Herein, $b_{i,l}^{(m,t)} = y_i - \boldsymbol{x}_{i(l)}^\top \hat{\boldsymbol{\beta}}_{(l)}^{(m,t)}$; $\boldsymbol{x}_{i(l)}$ is the vector $\boldsymbol{x}_i$ after deleting the l-th element; x_{il} is the l-th element of vector $\boldsymbol{x}_i$;

$\tau_{lj'} = -\tau_{j'l}$ if $l > j'$; $\beta_{lj'} = -\beta_{j'l}$ if $l > j'$. $ST(b, \delta) = \text{sign}(b)(|b| - \delta)_+$ is the soft-thresholding operator.

(II) Given $\hat{\beta}_{jj'}^{(m,t-1)}$, updating $\hat{\beta}_{jj'}^{(m,t)}$, $(j, j') \in \varepsilon$ by:

$$\hat{\beta}_{jj'}^{(m,t)} = \begin{cases} (\nu^{(t)})^{-1} ST\left(\tau_{jj'}^{(t)} + \nu^{(t)}(\hat{\beta}_j^{(m,t)} - \hat{\beta}_{j'}^{(m,t)}), \frac{\lambda_2}{\tau}\right) & (j, j') \in \varepsilon^{(m-1)}, \\ \hat{\beta}_{jj'}^{(m-1)} & (j, j') \in \varepsilon^{(m-1)^c}. \end{cases} \tag{2.14}$$

The process of coordinate descent iterates until convergence, which satisfies the terminate condition $\|\hat{\boldsymbol{\beta}}^{(m,t)} - \hat{\boldsymbol{\beta}}^{(m,t-1)}\|_\infty \leq \delta^*$, where δ^* is a given small positive value. Hence, $\hat{\boldsymbol{\beta}}^{(m)} = \hat{\boldsymbol{\beta}}^{(m,t^*)}$, where t^* denotes the iteration at termination. The pseudo codes of the developed algorithm are summarized in Algorithm 1, the convergence of which is given in Theorem 2.2.

When solving the problem (2.6), the proposed method could potentially lead to a local optimum as the objective function in (2.6) is non-convex. Hence it is critical to assign a suitable initial value of $\boldsymbol{\beta}$, which controls the initial values of $\mathcal{F}^{(0)}, \varepsilon^{(0)}, \mathcal{N}^{(0)}$. A candidate initial value is adopted by the estimate of ncTLF in the R package `FGSG` (Yang et al., 2012).

Theorem 2.2 *The proposed Algorithm 1 converges, that is,*

$$S(\hat{\boldsymbol{\beta}}^{(m)}) \to c, \text{ as } m \to +\infty, \tag{2.15}$$

where c is a non-negative constant.

Algorithm 1 A hybrid algorithm integrated with augmented Lagrange and coordinate descent

```
Input: design matrix $X \in R^{n\times p}$, response vector $\boldsymbol{y} \in R^{n\times 1}$, parameters $\tau, \lambda_1, \lambda_2, \lambda_3, \rho, v, \delta^*$.
Output: $\hat{\boldsymbol{\beta}}^{(m)}$
 Initialization: $\hat{\boldsymbol{\beta}}^{(0)}, m = 0$
 do
     $m \leftarrow m + 1$.
     Update $\mathcal{F}^{(m)}, \varepsilon^{(m)}, \mathcal{N}^{(m)}$ according to (2.9).
     Initialization: $\hat{\boldsymbol{\beta}}^{(m,0)} \leftarrow \hat{\boldsymbol{\beta}}^{(m-1)}, t = 0$
     do
         $t \leftarrow t + 1$.
         Update $\hat{\beta_l}^{(m,t)}$ according to updating formulas (2.13).
         Update $\hat{\beta}_{jj'}^{(m,t)}$ according to updating formula (2.14).
     while $\|\hat{\boldsymbol{\beta}}^{(m,t)} - \hat{\boldsymbol{\beta}}^{(m,t-1)}\|_\infty \geq \delta^*$
 while $S(\hat{\boldsymbol{\beta}}^{(m)}) - S(\hat{\boldsymbol{\beta}}^{(m+1)}) > 0$
```

We derive the convergence of the proposed algorithm that is analogous to Shen et al. (2012).

2.3.3 Theoretical properties of the regularized nnFSG estimate

We show some properties of the proposed nnFSG estimate $\hat{\boldsymbol{\beta}}$. Before proceeding, we make the following assumption.

(A4) $4\tau^{-2}(\lambda_1 s^* + \lambda_2 |\mathcal{N}|) < \min_{K(\mathcal{G}_0^c) \leq K^*} \lambda_{\min}(n^{-1} Z^{\top}_{\mathcal{G}_0^c} Z_{\mathcal{G}_0^c})$, where s^* and K^* are upper bounds of the maximal number of non-zero coefficients and of the non-zero groupings, respectively. $|\mathcal{N}|$ is the maximal number of direct connections of variable x_j to variable $x_{j'}$, where $(j, j') \in \varepsilon$ and $j, j' \in \mathcal{G}_k, k = 1, \ldots, K$.

We remark that for a full connection $\varepsilon = \{(j, j') : j < j', j, j' = 1, \ldots, p\}$, $|\mathcal{N}| = s^*(s^* - 1)/2$. Specifically, $s_1^0 \leq s^* \leq p$, $K^0 \leq K^* \leq s^*$.

Theorem 2.3 *Under the assumptions (A1)-(A4), if* $\gamma_{\min} > 2\tau$,

$$\left\{(\gamma_{\min} - 2\tau) n^{1/2} \lambda_{\min}^{1/2}(n^{-1} Z^{\top}_{\mathcal{G}_0^{0c}} Z_{\mathcal{G}_0^{0c}}) \sigma^{-1}\right\}^2 \geq \max\left\{8 \log \frac{nK^0(K^0 - 1)}{(2\pi)^{1/2}}, 2 \log \frac{2n(p - |\mathcal{G}_0^0|)}{(2\pi)^{1/2}}\right\},$$

$$\left(\frac{n\lambda_1/\tau}{\sigma \max\limits_{1 \leq i \leq p} \|\boldsymbol{x}_{(j)}\|}\right)^2 \geq 2 \log \frac{2n|\mathcal{G}_0^0|}{(2\pi)^{1/2}}, \quad \left(\frac{n\lambda_2/\tau}{2\sigma\mathcal{D}}\right)^2 \geq 2 \log \frac{2n|\mathcal{N}|}{(2\pi)^{1/2}},$$

where $\mathcal{D} = \max\limits_{k, A \subset \mathcal{G}_k^0} \|X_A \mathbf{1}\| / |\varepsilon \cap \{A \times (\mathcal{G}_k^0 \setminus A)\}|$, *and* $'\times'$ *denotes the Cartesian product, then*

$$P(\hat{\boldsymbol{\beta}} \neq \hat{\boldsymbol{\beta}}^{ora}) = O\left(\frac{1}{n(\log n)^{1/2}}\right).$$

Furthermore, if

$$\frac{1}{n}\|X\boldsymbol{\beta}^0\|^2 + \frac{\tau^2}{16} \min_{K(\mathcal{G}_0^c) \leq K^*} \lambda_{\min}\left(\frac{1}{n} Z^{\top}_{\mathcal{G}_0^c} Z_{\mathcal{G}_0^c}\right) = o(K^0 (\log n)^{1/2}),$$

then we have

$$\frac{1}{n}E\left\|X\hat{\boldsymbol{\beta}} - X\boldsymbol{\beta}^0\right\|^2 = \frac{K^0\sigma^2}{n}(1+o(1)).$$

Note that the results of Theorem 2.3 are parallel to that of Theorem 2.1. The proposed non-negative estimate $\hat{\boldsymbol{\beta}}$ identifies consistently the true grouping, and also yields a mean squared error bounded by $n^{-1}K^0\sigma^2(1+o(1))$.

2.4 Numerical studies

2.4.1 Evaluation measures

The criteria used for measuring the prediction accuracy of the estimate $\hat{\boldsymbol{\beta}}$ are the mean squared error (MSE), MSE $= n^{-1}\|X(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|^2$, and mean absolute error (MAE), MAE $= n^{-1}\|X(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|_1$. Since the regression vetctor is sparse, and may also carry a structure with disjoint subgroups, in light of Yang et al. (2012), we thus provide another two metrics, feature true positive rate (FTP),

$$\text{FTP} = \frac{\sum_{j\in\mathcal{G}_0^0} I_{\{\hat{\beta}_j=0\}} + \sum_{j\notin\mathcal{G}_0^0} I_{\{\hat{\beta}_j\neq 0\}}}{p},$$

and group true positive rate (GTP),

$$\text{GTP} = \frac{\sum_{k=1}^{K^0}\text{GTP}_k + \text{FTP}}{K^0+1},$$

where

$$\mathrm{GTP}_k = \frac{\sum_{j'\neq j, j', j\in\mathcal{G}_k^0} I_{\{\hat{\beta}_{j'}=\hat{\beta}_j\}} + \sum_{j'\neq j, j'\in\mathcal{G}_k^0, j\notin\mathcal{G}_k^0} I_{\{\hat{\beta}_{j'}\neq\hat{\beta}_j\}}}{|\mathcal{G}_k^0|(p-1)}, k = 1, \ldots, K^0.$$

FTP and GTP measure respectively the accuracy of model's performance in terms of feature selection and feature grouping. It is clear that FTP, $\mathrm{GTP}_k(k = 1, \ldots, K^0)$ and GTP $\in [0, 1]$. Ideally, they should be close to 1.

2.4.2 Tuning free parameter: λ_3

Although the regularization problem (2.6) contains four parameters, say, $\tau, \lambda_1, \lambda_2$ and λ_3, the amount of work to select tuning parameters is parallel to that of Shen et al. (2012). Among those, $\tau, \lambda_1, \lambda_2$ are selected by five-fold cross-validation. λ_3 shrinks the negative coordinates of $\boldsymbol{\beta}$, which, together with λ_1, controls the non-negativity. It is easy to see that λ_3 is not required to be tuned precisely. Indeed, a large positive value is enough. Now, we perform finite sample simulations to illustrate the effects of λ_3 by fixing $\tau, \lambda_1, \lambda_2$.

We generate the samples $(\boldsymbol{x}_i, y_i), i = 1, \ldots, n$, via a linear model $y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \epsilon_i$, where $\boldsymbol{x}_i \overset{iid}{\sim} N_p(\mathbf{0}, \Sigma)$ with $\Sigma = (\sigma_{\ell j})$ and $\sigma_{\ell j} = 0.5^{|\ell - j|}, \ell, j = 1, \ldots, p$; the random error $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$. We set the true regression coefficient

$$\boldsymbol{\beta}^0 = (\underbrace{1, \ldots, 1}_{4}, \underbrace{2, \ldots, 2}_{4}, \underbrace{3, \ldots, 3}_{4}, \underbrace{4, \ldots, 4}_{4}, \underbrace{0, \ldots, 0}_{p-16})^\top \in \mathbb{R}^p.$$

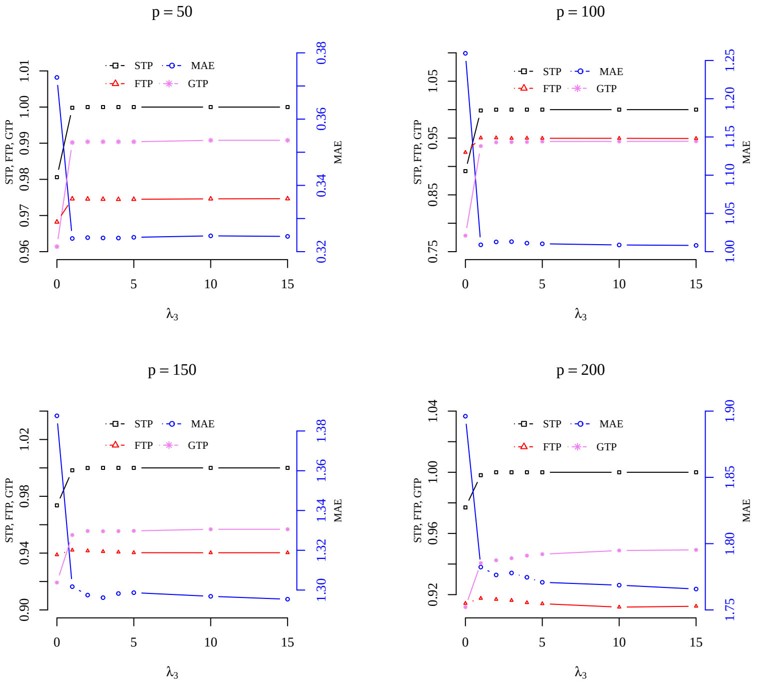

Figure 2.1: The values of STP, MAE, FTP and GTP with different λ_3 and fixed $\lambda_1, \lambda_2, \tau$, averaged over 100 simulations.

Let $\tau = 0.1, \lambda_1 = \lambda_2 = 10^{-3}\bar{\lambda}$, where $\bar{\lambda} = \|X^\top \boldsymbol{y}\|_\infty$, and $\lambda_3 \in \{0, 1, 2, 3, 4, 5, 10, 15\}$. Herein, we take $\sigma = 1, n = 100, p = 50, 100, 150, 200$.

Figure 2.1 displays the values of MAE, FTP and GTP, averaged over 100 simulations for the post samples. In addition, we report the sign true positive rate (STP) in the same figure. STP is defined as the proportion of non-negative coordinates of $\hat{\boldsymbol{\beta}}$ that is, STP $= \sum_{j=1}^{p} I_{\{\hat{\beta}_j \geq 0\}}/p$. In an instance where p is fixed, all the criteria values appear commensurate as λ_3 exceeds a critical value. For example, when $p = 100$, the values of FTP, GTP and MAE tend to be stable and the STP values are exactly 1 as $\lambda_3 \geq 1$, which completely captures the non-negativity of the underlying true coefficients. It indicates that the method's performance in achieving non-negative estimate is independent of λ_3 as it exceeds a critical value. Indeed, we arrive at the same results for the other instances, say, $p = 50, 150, 200$. We remark that the critical value may be different under different model's settings. In a real application, we thus take a large value of λ_3, say, 10 or even larger.

2.4.3 Model comparisons of non-negative feature selection

In our simulation study, we are interested in the performance of our proposed method in feature selection. We carry out finite sample simulations via the linear model $y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \epsilon_i,\ i = 1, \ldots, n$, where the settings of $\boldsymbol{x}_i$ and ϵ_i are the same as in Section 2.4.2. The positive elements of the true coefficient vector $\boldsymbol{\beta}^0$ are randomly generated from a uniform distribution $[0.5, 5]$, $s_1^0 = 10$, and $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$. Put $\lambda_2 = 0, \lambda_3 = 10$. λ_1 and τ are selected from candidate sets using five-fold cross-validation. We take $n = 100,\ p = 50, 100, 200,\ \sigma = 0.5, 1, 2$.

We compare our method with others that also achieve non-negative estimates by using the R packages, say, `nnls` (Mullen and van Stokkum, 2012), `glmnet` (Friedman et al., 2016), `penalized` (Goeman, 2010), `CVXR` (Fu et al., 2017). The comparisons

Table 2.1: Comparison of glmnet, penalized, CVXR, nnls, and nnFSG under the settings with $n = 100, p = 50, 100, 200, \sigma = 0.5, 1, 2$. The average values of MSE, MAE, FTP as well as their standard deviations (in parenthesis) are based on 500 simulations.

p	Method	$\sigma = 0.5$			$\sigma = 1$			$\sigma = 2$		
		MSE	MAE	FTP	MSE	MAE	FTP	MSE	MAE	FTP
50	glmnet	0.1201(0.0585)	0.2683(0.0672)	0.9817(0.0304)	0.2033(0.0928)	0.3515(0.0788)	0.8920(0.0569)	0.8118(0.3694)	0.7023(0.1572)	0.8914(0.0573)
	penalized	0.0523(0.0239)	0.1783(0.0401)	0.9120(0.0455)	0.2091(0.0956)	0.3565(0.0803)	0.9120(0.0456)	0.8349(0.3795)	0.7125(0.1598)	0.9120(0.0454)
	CVXR	0.0513(0.0233)	0.1765(0.0396)	0.9112(0.0455)	0.2051(0.0933)	0.3531(0.0792)	0.9112(0.0455)	0.8190(0.3702)	0.7056(0.1575)	0.9111(0.0454)
	nnls	0.0719(0.0297)	0.2104(0.0425)	0.7540(0.0582)	0.2874(0.1187)	0.4208(0.0850)	0.7539(0.0581)	1.1486(0.4733)	0.8413(0.1696)	0.7535 (0.0583)
	nnFSG	0.0303(0.0176)	0.1340(0.0377)	1.0000(0.0000)	0.1447(0.0997)	0.2891(0.0943)	0.9988(0.0049)	0.8012(0.4774)	0.6877(0.2003)	0.9845(0.0163)
100	glmnet	0.1373(0.0579)	0.2905(0.0610)	0.9910(0.0129)	0.2502(0.1045)	0.3915(0.0832)	0.9337(0.0295)	1.0188(0.4117)	0.7910(0.1630)	0.9131 (0.0428)
	penalized	0.0669(0.0294)	0.2022(0.0445)	0.9409(0.0252)	0.2676(0.1179)	0.4043(0.0892)	0.9407(0.0253)	1.0705(0.4715)	0.8085(0.1783)	0.9411 (0.0252)
	CVXR	0.066(0.0293)	0.2007(0.0446)	0.9401(0.0253)	0.2641(0.1175)	0.4014(0.0894)	0.9398(0.0253)	1.0562(0.4693)	0.8028(0.1786)	0.9400 (0.0255)
	nnls	0.1563(0.0543)	0.3115(0.0544)	0.7238(0.0421)	0.6252(0.2174)	0.6230(0.1088)	0.7237(0.0420)	2.4963(0.8663)	1.2449(0.2169)	0.7235(0.0421)
	nnFSG	0.0283(0.0141)	0.1305(0.0329)	1.0000(0.0000)	0.1377(0.1019)	0.2814(0.0936)	0.9991(0.0029)	0.7867(0.6329)	0.6728(0.2302)	0.9902(0.0124)
200	glmnet	0.1319(0.0519)	0.2854(0.0560)	0.9917(0.0098)	0.2941(0.1163)	0.4268(0.0841)	0.9514(0.0228)	1.2192(0.4997)	0.8685 (0.1755)	0.9442(0.0315)
	penalized	0.0811(0.0341)	0.2235(0.0467)	0.9665(0.0127)	0.3244(0.1364)	0.4470(0.0934)	0.9664(0.0128)	1.2928(0.5430)	0.8924(0.1861)	0.9664(0.0127)
	CVXR	0.0798(0.0332)	0.2216(0.0462)	0.9660(0.0127)	0.3191(0.1329)	0.4433(0.0924)	0.9659(0.0127)	1.2714(0.5293)	0.8850(0.1841)	0.9659(0.0127)
	nnls	0.5842(0.3222)	0.5923(0.1485)	0.7109(0.0333)	2.3369(1.2887)	1.1847(0.2970)	0.7109(0.0334)	9.3089(5.1095)	2.3648(0.5911)	0.7105 (0.0336)
	nnFSG	0.0277(0.0141)	0.1295(0.0329)	1.0000(0.0000)	0.1168(0.0660)	0.2643(0.0718)	0.9999(0.0009)	0.7421(0.5810)	0.6498(0.2334)	0.9953(0.0074)

among those methods are based on how well they estimate the underlying true regression vector, measured using MSE and MAE; and how well they perform in terms of feature selection, measured using FTP. The larger the values of FTP, the better the performance of feature selection. Table 2.1 reports the averages and standard deviations of MSE, MAE and FTP, which are obtained based on 500 simulations. As Table 2.1 illustrates, nnFSG outperforms the other methods in terms of MSE, MAE and FTP uniformly.

2.4.4 Method performance in non-negative feature selection and grouping

nnFSG also permits non-negative disjoint smoothness of regression coefficients whose values are similar or identical within a subgroup. We now conduct simulation experiments to demonstrate the numerical performance of our proposed method.

We carry out simulations via the linear model $y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \epsilon_i,\ i = 1, \ldots, n$. The true coefficients $\boldsymbol{\beta}^0 = (\underbrace{1, \ldots, 1}_{4}, \underbrace{2, \ldots, 2}_{4}, \underbrace{3, \ldots, 3}_{4}, \underbrace{4, \ldots, 4}_{4}, 0, \ldots, 0)^\top \in \mathbb{R}^p$, a p-dimensional regression coefficients with 5 groups, among which $s_1^0 = 16$. Note that the order of the elements of $\boldsymbol{\beta}^0$ is randomly given. The settings of $\boldsymbol{x}_i$ and n are same as in Section 2.4.3. Let $p = 50, 100, 200, 500$ and $\sigma = 0.5, 1, 1.5$. Put $\lambda_3 = 10$. For simplicity, we further set $\lambda_1 = \lambda_2$. The optimal regularization parameters τ, λ_1 or λ_2

Table 2.2: The average values of MSE, MAE, FTP, GTP as well as their standard deviations (in parenthesis) based on 500 simulations.

p	σ	MSE	MAE	FTP	GTP
50	0.5	0.0115(0.0113)	0.0786(0.0341)	1.0000(0.0000)	0.9989(0.0072)
	1.0	0.0701(0.0982)	0.1829(0.1073)	0.9994(0.0041)	0.9955(0.0137)
	1.5	0.4815(0.4725)	0.4857(0.2672)	0.9914(0.0172)	0.9760(0.0253)
100	0.5	0.0128(0.0159)	0.0820(0.0383)	1.0000(0.0000)	0.9995(0.0036)
	1.0	0.0883(0.1341)	0.2024(0.1246)	0.9995(0.0028)	0.9962(0.0092)
	1.5	0.6985(0.5810)	0.6028(0.2878)	0.9886(0.0183)	0.9793(0.0161)
200	0.5	0.0277(0.1190)	0.0910(0.0792)	0.9999(0.0014)	0.9983(0.0066)
	1.0	0.2904(0.5038)	0.3650(0.2284)	0.9981(0.0068)	0.9869(0.0153)
	1.5	2.2393(3.8810)	1.0082(0.6470)	0.9766(0.0358)	0.9688(0.0297)
500	0.5	0.0469(0.1544)	0.1312(0.1140)	0.9999(0.0005)	0.9969(0.0072)
	1.0	0.6388(6.5795)	0.4327(0.5002)	0.9988(0.0048)	0.9915(0.0143)
	1.5	3.1120(7.9409)	1.2318(0.7128)	0.9849(0.0140)	0.9794(0.0252)

are chosen by five-fold cross-validation. When the number of independent variables is very large, in order to promote the computational efficiency, we first screen out those variables whose coefficients are identified as zeros through non-negative least squares method by using R package `nnls`.

Table 2.2 reports the MSE, MAE, FTP and GTP, averaged over 500 simulations. As the variance of random errors increases, both MSE and MAE increase, implying that the method's prediction accuracy decreases. Meanwhile, the values of FTP and GTP decrease, which indicates that the method's performance of feature selection and grouping degrades slightly. As the dimension p increases, the prediction accuracy decreases as well. However, the values of GTP and FTP are both above 0.97 over the whole scenarios, which reflects that nnFSG performs well in feature grouping and selection.

2.4.5 Synthetic malaria vaccine data

Table 2.3: Comparison of glmnet, penalized, CVXR, nnls, and nnFSG assessed on synthetic malaria vaccine data under the settings with $n = 100, p = 90, \sigma = 0.2, 0.3$. The average values of MSE and MAE as well as their standard deviations (in parenthesis) are based on 500 simulations.

Case	Method	$\sigma = 0.2$		$\sigma = 0.3$	
		MSE	MAE	MSE	MAE
I	glmnet	0.3833(0.1834)	0.4787(0.1083)	0.3956(0.1808)	0.4872(0.1072)
	penalized	571.16(1692.30)	8.3371(21.4788)	589.02(1702.16)	8.8518(21.6545)
	CVXR	26.4783(81.8077)	1.5315(3.7746)	26.9999(81.7741)	1.6466(3.7712)
	nnls	0.0336(0.0119)	0.1446(0.0253)	0.0756(0.0267)	0.2169(0.0379)
	nnFSG	0.0004(0.0006)	0.0155(0.0115)	0.0009(0.0013)	0.0234(0.0175)
II	glmnet	0.0316(0.0112)	0.1403(0.0243)	0.0706(0.0248)	0.2097(0.0362)
	penalized	1.7799(10.0871)	0.6020(1.1177)	2.2497(10.2750)	0.8325(1.1499)
	CVXR	0.1082(0.4842)	0.1820(0.1851)	0.1576(0.4739)	0.2593(0.1784)
	nnls	0.0336(0.0119)	0.1446(0.0253)	0.0756(0.0267)	0.2169(0.0379)
	nnFSG	0.0024(0.0046)	0.0306(0.0256)	0.0105(0.0167)	0.0630(0.0530)
III	glmnet	0.0313(0.0110)	0.1396(0.0241)	0.0701(0.0244)	0.2090(0.0359)
	penalized	0.5864(1.0868)	0.5718(0.4335)	1.4806(1.5629)	1.0077(0.5419)
	CVXR	0.0512(0.0486)	0.1697(0.0611)	0.1204(0.0775)	0.2656(0.0768)
	nnls	0.0366(0.0119)	0.1447(0.0253)	0.0755(0.0266)	0.2170(0.0378)
	nnFSG	0.0250(0.0205)	0.1187(0.0444)	0.0876(0.0556)	0.2272(0.0701)

In vaccine design study, it is very crucial to locate the important amino acid sites and their associated importance (regression coefficients). A vaccine that is designed

Table 2.4: Comparison of glmnet, penalized, CVXR, nnls, and nnFSG assessed on synthetic malaria vaccine data under the settings with $n = 100, p = 90, \sigma = 0.2$. The average values of Sen, Spe, Info, FTP and GTP as well as their standard deviations (in parenthesis) are based on 500 simulations.

Case	Method	Sen	Spe	Info	FTP	GTP
I	glmnet	1.0000(0.0000)	0.8033(0.0527)	0.8033(0.0527)	0.8558(0.0387)	0.7987(0.0193)
	penalized	0.9503(0.1635)	0.8608(0.0585)	0.8112(0.1342)	0.8847(0.0370)	0.7988(0.0550)
	CVXR	0.9824(0.0658)	0.7464(0.0557)	0.7288(0.0738)	0.8094(0.0398)	0.7709(0.0281)
	nnls	1.0000(0.0000)	0.6735(0.0609)	0.6735(0.0609)	0.7606(0.0446)	0.7512(0.0223)
	nnFSG	1.0000(0.0000)	0.9980(0.0447)	0.9980(0.0447)	0.9985(0.0328)	0.9993(0.0164)
II	glmnet	0.9996(0.0041)	0.6952(0.0601)	0.6948(0.0604)	0.7766(0.0441)	0.8433(0.0147)
	penalized	0.9930(0.0473)	0.8459(0.0439)	0.8389(0.0528)	0.8854(0.0303)	0.8765(0.0247)
	CVXR	0.9972(0.0247)	0.7400(0.0524)	0.7372(0.0567)	0.8088(0.0385)	0.8531(0.0161)
	nnls	0.9997(0.0037)	0.6734(0.0608)	0.6731(0.0609)	0.7606(0.0446)	0.8379(0.0149)
	nnFSG	0.9992(0.0058)	0.9997(0.0021)	0.9989(0.0080)	1.0000(0.0000)	0.9916(0.0249)
III	glmnet	0.9921(0.0187)	0.6969(0.0586)	0.6890(0.0623)	0.7798(0.0431)	0.8860(0.0109)
	penalized	0.9617(0.0487)	0.8505(0.0465)	0.8122(0.0585)	0.8948(0.0315)	0.9092(0.0186)
	CVXR	0.9789(0.0324)	0.7396(0.0521)	0.7185(0.0635)	0.8126(0.0379)	0.8926(0.0126)
	nnls	0.9934(0.0173)	0.6710(0.0608)	0.6644(0.0638)	0.7604(0.0447)	0.8812(0.0112)
	nnFSG	0.9566(0.0353)	0.9794(0.0183)	0.9360(0.0447)	0.9881(0.0183)	0.9301(0.0180)

to match those important sties can improve induced immunity. Furthermore, the sites associated with immune response with negative coefficients should be excluded in the model (Hu et al., 2015), in which a non-negative lasso method was applied to identify the important amino acid and estimate their relative importance. For some confidential reasons, we are not allowed to access the original data. We thus assess the performance of our proposed method using the synthetic, but realistic, data

Table 2.5: Comparison of glmnet, penalized, CVXR, nnls, and nnFSG assessed on synthetic malaria vaccine data under the settings with $n = 100, p = 90, \sigma = 0.3$. The average values of Sen, Spe, Info, FTP and GTP as well as their standard deviations (in parenthesis) are based on 500 simulations.

Case	Method	Sen	Spe	Info	FTP	GTP
I	glmnet	1.0000(0.0000)	0.7935(0.0526)	0.7935(0.0526)	0.8486(0.0386)	0.7951(0.0193)
	penalized	0.9500(0.1643)	0.8618(0.0588)	0.8118(0.1344)	0.8853(0.0373)	0.7990(0.0550)
	CVXR	0.9823(0.0656)	0.7468(0.0561)	0.7291(0.0736)	0.8096(0.0400)	0.7709(0.0280)
	nnls	1.0000(0.0000)	0.6735(0.0608)	0.6735(0.0608)	0.7606(0.0446)	0.7512(0.0223)
	nnFSG	1.0000(0.0000)	0.9900(0.0996)	0.9900(0.0996)	0.9927(0.0730)	0.9961(0.0367)
II	glmnet	0.9993(0.0052)	0.6962(0.0597)	0.6955(0.0601)	0.7774(0.0438)	0.8435(0.0146)
	penalized	0.9912(0.0487)	0.8461(0.0443)	0.8373(0.0542)	0.8857(0.0304)	0.8764(0.0251)
	CVXR	0.9962(0.0253)	0.7405(0.0522)	0.7367(0.0569)	0.8094(0.0383)	0.8532(0.0161)
	nnls	0.9994(0.0049)	0.6733(0.0609)	0.6727(0.0611)	0.7606(0.0446)	0.8378(0.0149)
	nnFSG	0.9972(0.0111)	0.9984(0.0055)	0.9956(0.0157)	0.9996(0.0026)	0.9773(0.0368)
III	glmnet	0.9772(0.0306)	0.6927(0.0580)	0.6699(0.0677)	0.7803(0.0431)	0.8858(0.0113)
	penalized	0.8957(0.0768)	0.8648(0.0464)	0.7605(0.0785)	0.9042(0.0310)	0.8940(0.0352)
	CVXR	0.9423(0.0514)	0.7394(0.0508)	0.6817(0.0768)	0.8152(0.0380)	0.8878(0.0148)
	nnls	0.9821(0.0265)	0.6674(0.0609)	0.6494(0.0678)	0.7605(0.0450)	0.8809(0.0116)
	nnFSG	0.9439(0.0456)	0.9538(0.0326)	0.8977(0.0516)	0.9515(0.0338)	0.9005(0.0303)

under the simulation benchmarks that are similar to Hu et al. (2015). In that article, three cases were considered under the settings of $n = 100, p = 90, s_1^0 = 24$. Suppose that the explanatory variables are all independent. We randomly generate the i-th sample $(\boldsymbol{x}_i, y_i)$ via the linear model $y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \epsilon_i$. Denote $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^\top$, where $x_{ij} \overset{iid}{\sim} \text{Bernoulli}(p_j)$, $p_j \overset{iid}{\sim} \text{Beta}(2,5)$ for $j = 1, \ldots, p$. And $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ for $i = 1, \ldots, n$. Consider three cases for the true $\boldsymbol{\beta}^0$,

- Case I: $\boldsymbol{\beta}^0 = (\underbrace{10, \ldots, 10}_{s_1^0}, \underbrace{0, \ldots, 0}_{p-s_1^0})^\top$.
- Case II: $\boldsymbol{\beta}^0 = (\underbrace{2, \ldots, 2}_{s_1^0/2}, \underbrace{1, \ldots, 1}_{s_1^0/2}, \underbrace{0, \ldots, 0}_{p-s_1^0})^\top$.
- Case III: $\boldsymbol{\beta}^0 = (\underbrace{1, \ldots, 1}_{s_1^0/3}, \underbrace{0.5, \ldots, 0.5}_{s_1^0/3}, \underbrace{0.3, \ldots, 0.3}_{s_1^0/3}, \underbrace{0, \ldots, 0}_{p-s_1^0})^\top$.

We remark that the order of the components of $\boldsymbol{\beta}^0$ is randomly given. Note that y_i are the immune response observations that are usually measured by the growth inhibition assay. The source of measurement error may result from systematic error. It is sensible to assume that the variability of measurement error is small. Let $\sigma = 0.2, 0.3$. Again, we set $\lambda_3 = 10$, and the rest tuning parameters $\tau, \lambda_1, \lambda_2$ are selected by five-fold cross-validation. We also compare our proposed method with the others that are estimable by using the R package `glmnet`, `penalized`, `CVXR`, `nnls`. Hu et al. (2015) adopted the evaluation criteria sensitivity (Sen), measuring the probability of an important variable associated with a non-zero coefficient being selected, and specificity (Spe), measuring the probability of an unimportant variable associated with a zero coefficient not being selected. The larger the values of both Spe and Sen, the better the performance of the method. A perfect situation would be described as 100% sensitivity, meaning all important sites are correctly identified, and 100% specificity, meaning all unimportant sites are excluded. In reality, there is usually a trade-off between these two measures. We thus apply the third criteria informedness (Info), Info = Spe + Sen - 1, the magnitude of which measures the probability of an informed decision. The simulation results are shown in Tables 2.3-2.5. From Table 2.3, we observe that our proposed method performs best for almost all scenarios in terms of MSE and MAE except the case III when $\sigma = 0.3$. Since the objective of that work is to find the important site where the associated coefficient is non-zero, we are more interested in Spe, Sen, Info and FTP (see Tables 2.4-2.5), measuring effectiveness in identifying main sites or features. Moreover, one might be interested in identifying the subgroups within which the important sites are similar or identical. Considering the simulation settings of $\boldsymbol{\beta}^0$, we thus provide the GTP values in Tables 2.4-2.5 as well. In terms of Spe and Sen, nnFSG achieves the largest values in case I when $\sigma = 0.2, 0.3$. In case II and III, the largest values of Spe are obtained by our method, albeit the Sen values are slightly smaller than those obtained by `glmnet` and `nnls`. Given the trade-off between Sen and Spe, our method dominates all other methods uniformly, which is reflected by the largest values of Info. Our method's outperformance in feature selection and grouping is also demonstrated by the largest values of FTP and GTP among those scenarios.

2.4.6 Protein mass spectrometry data

Mass spectrometry (MS) analysis has become a key role in extracting reliable proteomic features (peptides) from complex biological mixtures (Renard et al., 2008), a fundamental step in the automated analysis of proteomic MS experiments. A peptide produces a signal at multiple mass positions, which becomes manifest in a series of regularly spaced peaks. For more details on the backgrounds, one can refer to Renard et al. (2008); Slawski et al. (2010, 2012). Figure 2.2 shows a protein mass spectrum of Myoglobine in the m/z 800-2500 range, $118,464$ (m/z, intensity) pairs in total. An initial part at the m/z range of 800-834 is zoomed. The peptides whose intensities differ drastically occur in different m/z-regions. The data set is kindly provided by B. Gregorius and A. Tholey, Department of Experimental Medicine, Working Group for Systematic Proteomics, Christian-Albrechts-Universitaet zu Kiel, which is avaiable in the R package `IPPD` (Slawski et al., 2012).

The peptides extraction problem is to identify those m/z-positions where peptides are located, which can be recast as a sparse recovery problem. Renard et al. (2008) and Slawski et al. (2010, 2012) proposed template matching-based methods to solve the problem. Motivated by Tibshirani and Wang (2007), we can also regard the peptides extraction as a 'hot spot' detection problem. The model's setup for the protein MS data is $p = n = 118,464$, and $X = I_p$, that is, $y_i = \beta_i + \epsilon_i, i = 1, \ldots, n$.

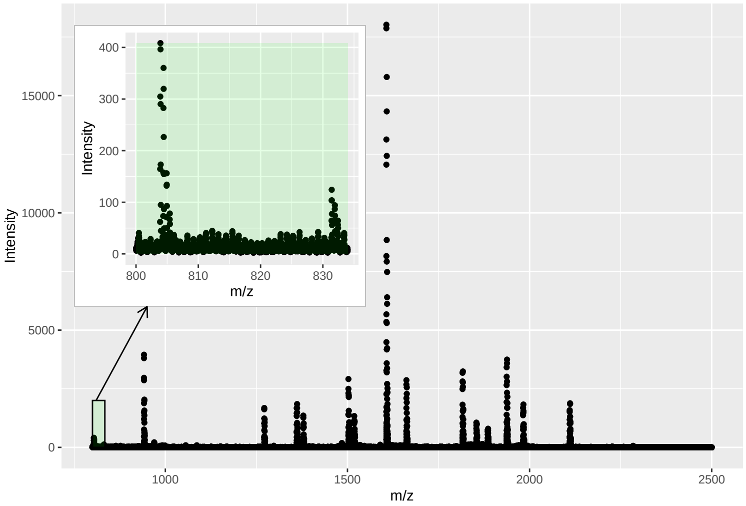

Figure 2.2: Raw protein mass spectrum of Myoglobine in the m/z 800-2500 range. The left upper panel zooms at the m/z range of 800-834.

Given the non-negativity of y_i (intensity), it is reasonable to impose non-negative constraints on β_i, the weight of the i-th m/z-site.

Simultaneously estimating the weights of all m/z sites for the MS data in Figure 2.2 is difficult since it is computationally unmanageable when p is ultrahigh, say, 118,464. We thus cut the data into consecutive blocks, which doesn't affect the estimating results. Herein, we choose m/z-sites in the range of 800-834 for analysis, giving a total of 2,009 points. To make the computation efficient, we further cut the 2,009 data points into four consecutive blocks, giving 500 data points in the first three blocks and 509 data in the last block. The performance of our proposed method on the MS data is illustrated in Figure 2.3. We can see that, on one hand,

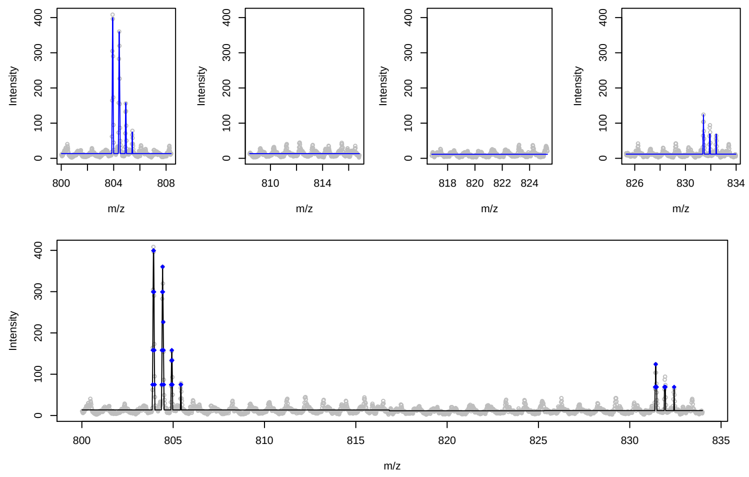

Figure 2.3: The upper panels show the four consecutive blocks from left to right. The lower panel shows the m/z range of 800-834. The grey points represent the MS data, and the solid blue line represents the estimated weights $\hat{\boldsymbol{\beta}}$ from the proposed method. The blue points describes the weights of these m/z sites that are extracted.

the proposed method successfully identifies the amplification. On the other hand, the method put same weights, rather than zeros, at those sites where the amplifications are not significant. We consider the sites with identical weights as one base group (see the horizontal blue solid line). Clearly, the second and third upper panels in Figure 2.3 show that there is no peptide in the m/z region. These sites that are not in the base group can be regarded as peptides, which are marked with blue points in the lower panel of Figure 2.3. We remark that if we increase λ_1, the weights of the base groups will be shrunken to zeros.

3 Asymptotic properties on high-dimensional multivariate regression M-estimation

3.1 Introduction

In many statistical applications, there are more than one correlated response variables. A general multivariate linear regression model is considered,

$$\mathbf{y}_i = \mathtt{X}_i \boldsymbol{\vartheta} + \mathbf{e}_i, \quad i = 1, \ldots, n, \tag{3.1}$$

where $\boldsymbol{\vartheta}$ is a p-vector of unknown parameters; $\mathtt{X}_i$ are $m \times p$ random matrices (m is fixed); $\mathbf{e}_i$ are iid distributed m-vectors, which encompasses the traditional linear regression model ($m = 1$). As pointed out by Bai et al. (1992), the model (3.1) is more general than the classical multivariate linear regression model,

$$\mathbf{y}_i = B\mathbf{x}_i + \mathbf{e}_i, \quad i = 1, \ldots, n, \tag{3.2}$$

where B is an $m \times p$ matrix of unknown parameters; $\mathbf{x}_i$ are $p \times 1$ vectors; $\mathbf{y}_i$ are $m \times 1$ response vectors; and $\mathbf{e}_i$ are iid m-vectors. More details on multivariate regression models can be found in Zellner (1962), and Koenker and Portnoy (1990).

A well-known method for estimating the regression coefficient vector $\boldsymbol{\vartheta}$ in (3.1) is the OLS estimation that is mathematically convenient and efficient for normal distributed errors. However, OLS is not resistant to outliers and not stable in respect to deviations from various assumptions. Huber (1964, 1973) thus introduced an M-estimation of $\boldsymbol{\vartheta}$ by minimizing

$$\sum_{i=1}^{n} \rho(\mathrm{y}_i - \mathbf{x}_i^\top \boldsymbol{\vartheta})$$

for a discrepancy function ρ. It is noted that Bai et al. (1992) developed asymptotic theories of M-estimate of $\boldsymbol{\vartheta}$ for model (3.1) under the classical regime that allows n to go to infinity but fixes p. In this chapter, we also consider this model but we study it under the regime that $p, n \to \infty, p/n \to \kappa$ with $0 < \kappa < \infty$.

Motivated by El Karoui (2013, 2018), we assume that the true regression parameter vector $\boldsymbol{\vartheta}_0$ is not sparse but diffuse, i.e., the elements of $\boldsymbol{\vartheta}_0$ are small and $\boldsymbol{\vartheta}_0$ cannot be well approximated by a sparse vector whose elements have mostly zeros. Inspired by El Karoui (2013, 2018), we characterize the behavior of the ridge-regularized high-dimensional regression M-estimate of $\boldsymbol{\vartheta}$ defined by (3.3) through a nonlinear system by using the double leave-one-out method. An analogous result of the unregularized high-dimensional regression M-estimate of $\boldsymbol{\vartheta}$ can be derived by setting $\tau = 0$. One should note that when $m = 1$, the model (3.1) reduces to the one studied by El

Karoui (2013).

The remainder of this chapter is organized as follows. In Section 3.2, we state the assumptions, main results and approximations. The derivation of double leave-one-out method are given in Section 3.3. Examples to validate our developed system are discussed in Section 3.4. We detail the propositions, lemmas and proofs of these theorems in Appendix B.

3.2 Methodology

Consider the multivariate linear model (3.1), assuming that $\mathbf{e}_i \sim (\mathbf{0}, \Sigma_{\mathbf{e}})$, the random matrix $\mathtt{X}_i \sim \mathcal{MN}_{m\times p}(\mathbf{0}, \Sigma_m, \Sigma_p)$, a matrix normal distribution, having mean matrix $\mathbf{0}$ $(m \times p)$, covariance matrix Σ_m (among-row) and Σ_p (among-column). Equivalently, $\mathrm{vec}(\mathtt{X}_i) \overset{iid}{\sim} \mathcal{N}_{mp}(\mathbf{0}, \Sigma_p \otimes \Sigma_m)$ with mean vector $\mathbf{0}$ $(mp \times 1)$ and covariance matrix $\Sigma_p \otimes \Sigma_m$ (Dawid, 1981). If $\Sigma_m = I_m$, the rows in $\mathtt{X}_i$ are independently distributed. Similarly, if $\Sigma_p = I_p$, the columns in $\mathtt{X}_i$ are independently distributed. Moreover, assume that $\mathbf{e}_i$ are independent of $\mathtt{X}_i$.

As we mainly consider the high-dimensional case, in light of El Karoui (2013, 2018), we estimate the high-dimensional regression parameter vector $\boldsymbol{\vartheta}$ by the following ridge-regularized high-dimensional regression M-estimate

$$\hat{\boldsymbol{\vartheta}} = \arg\min_{\boldsymbol{\vartheta}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} \rho\left(\Sigma_m^{-1/2}(\mathbf{y}_i - \mathtt{X}_i\boldsymbol{\vartheta})\right) + \frac{\tau}{2}\|\Sigma_p^{1/2}\boldsymbol{\vartheta}\|^2. \tag{3.3}$$

Here we assume that Σ_m and Σ_p are known, ρ is a continuously differentiable convex function from $\mathbb{R}^m$ to $\mathbb{R}$, and $\tau > 0$ is given. For simple presentation, we make the following transformation. By putting

$$X_i = \Sigma_m^{-1/2}\mathtt{X}_i\Sigma_p^{-1/2}, \quad \boldsymbol{e}_i = \Sigma_m^{-1/2}\mathbf{e}_i, \quad \boldsymbol{\beta}_0 = \Sigma_p^{1/2}\boldsymbol{\vartheta}_0, \quad \boldsymbol{\beta} = \Sigma_p^{1/2}\boldsymbol{\vartheta}. \tag{3.4}$$

The behavior of $\hat{\boldsymbol{\vartheta}}$ is now equivalent to the behavior of $\boldsymbol{\beta}$ given by

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} \rho(\boldsymbol{y}_i - X_i\boldsymbol{\beta}) + \frac{\tau}{2}\|\boldsymbol{\beta}\|^2, \tag{3.5}$$

where $\boldsymbol{y}_i = \boldsymbol{e}_i + X_i\boldsymbol{\beta}_0$, $X_i \overset{iid}{\sim} \mathcal{MN}_{m\times p}(\mathbf{0}, I_m, I_p)$, $\boldsymbol{e}_i \overset{iid}{\sim} (\mathbf{0}, \Sigma_{\boldsymbol{e}})$ with $\Sigma_{\boldsymbol{e}} = \Sigma_{\mathbf{e}}$, $\Sigma_m = I_m$, X_i are independent of $\boldsymbol{e}_i$, $\tau > 0$ is a constant, and ρ is a discrepancy convex function from $\mathbb{R}^m$ to $\mathbb{R}$. It is noted that if $\tau = 0$, (3.5) is reduced to the unregularized high-dimensional regression M-estimate, $\hat{\boldsymbol{\beta}}_M$, i.e.,

$$\hat{\boldsymbol{\beta}}_M = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} \rho(\boldsymbol{e}_i + X_i\boldsymbol{\beta}_0 - X_i\boldsymbol{\beta}). \tag{3.6}$$

By (3.4), the original estimate $\hat{\boldsymbol{\vartheta}}$ in (3.3) is obtained by $\hat{\boldsymbol{\vartheta}} = \Sigma_p^{-1/2}\hat{\boldsymbol{\beta}}$. Thus, we concentrate on the investigation of the behavior of $\hat{\boldsymbol{\beta}}$.

3.2.1 Assumptions

In this subsection, we make some assumptions. We first introduce the vector-valued proximal mapping of ρ, i.e., $\text{prox}_t(\rho)(\boldsymbol{z}) : \mathbb{R}^m \to \mathbb{R}^m$, for $t > 0$, and $\boldsymbol{u}, \boldsymbol{z} \in \mathbb{R}^m$ (see, e.g., Beck and Teboulle (2010); Moreau (1965) among others for more details)

$$\text{prox}_t(\rho)(\boldsymbol{z}) = \arg\min_{\boldsymbol{u}} \rho_t(\boldsymbol{z}; \boldsymbol{u}), \text{ where } \rho_t(\boldsymbol{z}; \boldsymbol{u}) = t\rho(\boldsymbol{u}) + \frac{1}{2}\|\boldsymbol{u} - \boldsymbol{z}\|^2.$$

We denote the power of $\log(n)$ by $\text{polyLog}(n)$, i.e. $\text{polyLog}(n) = O((\log n)^{m_0})$ for some constant $m_0 > 0$. For a sequence of random variables ξ_n, denote $\xi_n = O_{L_k}(1)$ if $(E|\xi_n|^k)^{1/k} = O(1)$, and $\xi_n = o_{L_k}(1)$ if $(E|\xi_n|^k)^{1/k} = o(1)$. m_ξ denotes the median of a random variable ξ. We make the following assumptions.

(A1) $p/n \to \kappa, 0 < \kappa < \infty$ as $p, n \to \infty$.

(A2) $\|\boldsymbol{\beta}_0\| \le c < \infty$, and $\|\boldsymbol{\beta}_0\|_\infty = O(n^{-\alpha})$ for $\alpha > 1/3$.

(A3) ρ is a twice differentiable convex function satisfying that $\rho(\boldsymbol{u}) \ge \rho(\boldsymbol{0}) = 0$ for any $\boldsymbol{u} \in \mathbb{R}^m$. $\sup_{\boldsymbol{u}} \|\boldsymbol{\psi}(\boldsymbol{u})\| \le c$, and $\sup_{\boldsymbol{u}} \lambda_{\max}(\nabla\boldsymbol{\psi}(\boldsymbol{u})) \le c$. For any vector $\boldsymbol{a}$, $\boldsymbol{b} \in \mathbb{R}^m$, $\|\nabla\boldsymbol{\psi}(\boldsymbol{a}) - \nabla\boldsymbol{\psi}(\boldsymbol{b})\|_{\max} \le c\|\boldsymbol{a} - \boldsymbol{b}\|$.

(A4) X_i are iid, $i = 1, \ldots, n$. All the elements of X_i are independently distributed. Given any positive integer $k < \infty$, the k-th moment of any element of X_i is bounded. Denote the j-th row of X_i by $\boldsymbol{x}_i(j)$. For $1 \le j \le m$, $\boldsymbol{x}_i(j) \sim (\boldsymbol{0}, I_p)$.

If f_1 is a convex function satisfying 1-Lipschitz condition, then $P(|f_1(\boldsymbol{x}_i(j)) - m_{f_1(\boldsymbol{x}_i(j))}| > t) \leq c\exp(-c_n t^2)$, where $t > 0$ and $1/c_n = O(\mathrm{polyLog}(n))$. Denote the ℓ-th column of $X = (X_1^\top, \ldots, X_n^\top)^\top$ by $\boldsymbol{x}_\ell$ for $1 \leq \ell \leq p$. For a convex function f_2 satisfying 1-Lipschitz condition, $P(|f_2(\boldsymbol{x}_\ell) - m_{f_2(\boldsymbol{x}_\ell)}| > t) \leq c\exp(-c_n t^2)$, where $t > 0$, and $1/c_n = O(\mathrm{polyLog}(n))$.

(A5) $\boldsymbol{e}_i \overset{iid}{\sim} (\mathbf{0}, \Sigma_{\boldsymbol{e}})$, are independent of X_i for $i = 1, \ldots, n$.

(F1) Given any vector $\boldsymbol{u} \in \mathbb{R}^m$, $\mathrm{tr}([I_m + x\nabla\boldsymbol{\psi}(\mathrm{prox}_x(\rho)(\boldsymbol{u}))]^{-1})$ is a decreasing function for $x \geq 0$.

The following two assumptions are the stronger versions of (A1) and (F1), which are required for Corollary 3.1.

(A1′) $p/n \to \kappa, 0 < \kappa < m$ as $p, n \to \infty$.

(F1′) ρ is strongly convex, and $\nabla\boldsymbol{\psi} \succ 0$. For any given vector $\boldsymbol{u} \in \mathbb{R}^m$, $\mathrm{tr}([I_m + x\nabla\boldsymbol{\psi}(\mathrm{prox}_x(\rho)(\boldsymbol{u}))]^{-1})$ is a strictly decreasing function for $x \geq 0$.

Remark 3.1 *Since we consider the case that $\boldsymbol{\beta}_0$ is not sparse, we limit each coordinate of $\boldsymbol{\beta}_0$ to a small value in (A2). (A3) is made for the theoretical study of M-estimation when p/n tends to a constant. Note that the requirement for $\sup_{\boldsymbol{u}} \|\boldsymbol{\psi}(\boldsymbol{u})\|$ having a bounded support no longer holds when $\rho(\boldsymbol{x}) = \boldsymbol{x}^\top A\boldsymbol{x}/2$. But our simulation experiments indicate that our system performs well on that case. By (A3),*

$\nabla\boldsymbol{\psi}(\boldsymbol{u}) \succeq 0$ *and* $\sup_{\boldsymbol{u}} \operatorname{tr}(\nabla\boldsymbol{\psi}(\boldsymbol{u})) \leq c$ *in view that* m *is fixed. The last one in (A3) can be viewed as a 1-Lipschitz condition. Motivated by El Karoui (2018), we make a general assumption on* X_i *in (A4), which holds true if* $X_i \sim \mathcal{MN}_{m\times p}(\mathbf{0}, I_m, I_p)$, *where* c_n *is a constant independent of* p, *or if* X_i *have independent entries with each bounded by a constant and with mean 0 and variance 1. More justifications on the assumption (A4) can be found in El Karoui (2009, 2013) and Ledoux (2001). (F1) implies that* $\operatorname{tr}([I_m + x\nabla\boldsymbol{\psi}(\operatorname{prox}_x(\rho)(\boldsymbol{u}))]^{-1})$ *has a unique root given* $\tau > 0$, *which plays a crucial role in proving the existence and uniqueness of* μ *in* (3.7). *(F1) holds true if* $\rho(\boldsymbol{x}) = \boldsymbol{x}^\top A\boldsymbol{x}/2$, *where* $A \succ 0$. *It is easy to see that* $\operatorname{tr}([I_m + x\nabla\boldsymbol{\psi}(\operatorname{prox}_x(\rho)(\boldsymbol{u}))]^{-1}) = \operatorname{tr}((I_m + xA)^{-1})$, *a decreasing function with respect to* $x \geq 0$. *(F1) also holds true if* $\rho(\boldsymbol{x}) = \|\boldsymbol{x}\|$.

3.2.2 Main results

We aim to characterize $E\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2$ by using the double leave-one-out approach for both observations and predictors, which was proposed by El Karoui et al. (2013). This approach relates $\hat{\boldsymbol{\beta}}$ to the leaving one observation out estimate $\hat{\boldsymbol{\beta}}_{(i)}$ in (3.9) and the leaving one predictor out estimate $\hat{\boldsymbol{\gamma}}$ in (3.15). The following theorem characterizes the risk of $\hat{\boldsymbol{\beta}}$, i.e., $E\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2$.

Theorem 3.1 *Assume that $\tau > 0$ is given. Under the assumptions (A1)-(A5) and (F1), $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|$ is asymptotically deterministic. If $\boldsymbol{z} = \boldsymbol{e} + \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|\tilde{\boldsymbol{z}}$ with $\boldsymbol{e}$ having the same distribution as $\boldsymbol{e}_i$, $\tilde{\boldsymbol{z}} \sim N(\boldsymbol{0}, I_m)$, and $\tilde{\boldsymbol{z}}$ being independent of $\boldsymbol{e}$, then there exists a constant $\mu > 0$ such that*

$$\begin{cases} E\mathrm{tr}(\nabla \mathrm{prox}_{\mu}(\rho)(\boldsymbol{z})) & = m - \kappa + \tau\mu, \\ \kappa^2 E\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2 & = \kappa E\|\boldsymbol{z} - \mathrm{prox}_{\mu}(\rho)(\boldsymbol{z})\|^2 + \tau^2\|\boldsymbol{\beta}_0\|^2\mu^2. \end{cases} \tag{3.7}$$

Both equations of this system establish functional relationships between μ and $E\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2$. In theory, the risk as well as μ can be found out via (3.7). It is noted that (i) the constant $\mu > 0$ is dependent on ρ, κ, τ; (ii) $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|$ is asymptotically deterministic as $p, n \to \infty$, namely $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2 \xrightarrow{p} E\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2$; (iii) under this regime, $\hat{\boldsymbol{\beta}}$ is biased, which is induced by the ridge regularization and the fluctuations of each coordinate as $p \to \infty$. When $\tau = 0$, Corollary 3.1 presented below characterizes the unregularized high-dimensional regression M-estimate $\hat{\boldsymbol{\beta}}_M$.

Corollary 3.1 *Under the assumptions (A1′), (A2)-(A5) and (F1′), $\|\hat{\boldsymbol{\beta}}_M - \boldsymbol{\beta}_0\|$ is asymptotically deterministic. If $\boldsymbol{z} = \boldsymbol{e} + \|\hat{\boldsymbol{\beta}}_M - \boldsymbol{\beta}_0\|\tilde{\boldsymbol{z}}$ with $\boldsymbol{e}$ having the same distribution as $\boldsymbol{e}_i$, $\tilde{\boldsymbol{z}} \sim N(\boldsymbol{0}, I_m)$, and $\tilde{\boldsymbol{z}}$ being independent of $\boldsymbol{e}$, then there exists a constant $\mu > 0$ such that*

$$\begin{cases} E\mathrm{tr}(\nabla \mathrm{prox}_{\mu}(\rho)(\boldsymbol{z})) & = m - \kappa, \\ \kappa E\|\hat{\boldsymbol{\beta}}_M - \boldsymbol{\beta}_0\|^2 & = E\|\boldsymbol{z} - \mathrm{prox}_{\mu}(\rho)(\boldsymbol{z})\|^2. \end{cases} \tag{3.8}$$

The proofs of Theorem 3.1 and Corollary 3.1 are given in Appendix B.2.

3.2.3 Main approximations

In this subsection, we state some important approximations that play vital roles in the inferential procedures of the double leave-one-out method, thus rendering us to obtain the systems (3.7) and (3.8).

3.2.3.1 Approximations of leaving one observation out

Let $\hat{\boldsymbol{\beta}}_{(i)}$ be the leaving one observation out estimate obtained by excluding the i-th observation $(X_i, \boldsymbol{y}_i)$, i.e.

$$\hat{\boldsymbol{\beta}}_{(i)} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{n}\sum_{j\neq i}^{n} \rho(\boldsymbol{e}_j + X_j\boldsymbol{\beta}_0 - X_j\boldsymbol{\beta}) + \frac{\tau}{2}\|\boldsymbol{\beta}\|^2. \tag{3.9}$$

Note that $\hat{\boldsymbol{\beta}}_{(i)}$ is independent of X_i. The corresponding leaving i-th observation out residual $\tilde{\boldsymbol{r}}_{j,[-i]} = \boldsymbol{e}_j + X_j\boldsymbol{\beta}_0 - X_j\hat{\boldsymbol{\beta}}_{(i)}$, for $j \neq i$. For $j = i$, $\tilde{\boldsymbol{r}}_{i,[-i]}$ is the prediction error of the i-th observation. Denote $S_i = n^{-1}\sum\limits_{j\neq i} X_j^\top \nabla\boldsymbol{\psi}(\tilde{\boldsymbol{r}}_{j,[-i]})X_j$, $C_i = n^{-1}X_i(S_i + \tau I_p)^{-1}X_i^\top$, $c_i = n^{-1}\text{tr}((S_i + \tau I_p)^{-1})$, $\tilde{\boldsymbol{g}}(\boldsymbol{u}) = \boldsymbol{u} + C_i\boldsymbol{\psi}(\boldsymbol{u})$. Define the residual as $\boldsymbol{r}_j = \boldsymbol{e}_j + X_j\boldsymbol{\beta}_0 - X_j\hat{\boldsymbol{\beta}},\ j = 1, \dots, n$.

With the purpose of approximating $\hat{\boldsymbol{\beta}}$ by $\hat{\boldsymbol{\beta}}_{(i)}$, we introduce a new quantity,

$$\tilde{\boldsymbol{\beta}}_i = \hat{\boldsymbol{\beta}}_{(i)} + \boldsymbol{\eta}_i, \text{ where } \boldsymbol{\eta}_i = \frac{1}{n}(S_i + \tau I_p)^{-1}X_i^\top \boldsymbol{\psi}(\tilde{\boldsymbol{g}}^{-1}(\tilde{\boldsymbol{r}}_{i,[-i]})). \tag{3.10}$$

The formulation of $\tilde{\boldsymbol{\beta}}_i$ will be provided afterward. El Karoui (2013, 2018) assumed intuitively that the difference between $\tilde{\boldsymbol{r}}_{j,[-i]}$ and $\boldsymbol{r}_j$ is negligible for $j \neq i$. For $j = i$, they showed that $\boldsymbol{r}_i$ could be well approximated by a function of $\tilde{\boldsymbol{r}}_{i,[-i]}$. One can expect reasonably that, therefore, those intuitive assumptions still hold true for the multivariate case ($m > 1$), which is given by the following theorem.

Theorem 3.2 *Under the assumptions (A1)-(A5), for any given $\tau > 0$, we have*

(i) $\sup_{1 \le i \le n} \|\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}_i\| = O_{L_k}(n^{-1}\text{polyLog}(n))$. *In particular, for* $1 \le i \le n$, $E(\|\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}_i\|^2) = O(n^{-2}\text{polyLog}(n))$, *and* $\|\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}\| = O_{L_k}(n^{-1/2})$.

(ii) $\sup_{1 \le i \le n} \sup_{j \neq i} \|\tilde{\boldsymbol{r}}_{j,[-i]} - \boldsymbol{r}_j\| = O_{L_k}(n^{-1/2}\text{polyLog}(n))$, $\sup_i \|\boldsymbol{r}_i - \tilde{\boldsymbol{g}}^{-1}(\tilde{\boldsymbol{r}}_{i,[-i]})\| = O_{L_k}\left(n^{-1/2}\text{polyLog}(n)\right)$, *and* $\sup_i \|\boldsymbol{r}_i - \text{prox}_{c_i}(\rho)(\tilde{\boldsymbol{r}}_{i,[-i]})\| = O_{L_k}(n^{-1/2}\text{polyLog}(n))$.

(iii) $\text{var}(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2) = O(n^{-1}\text{polyLog}(n))$.

This theorem establishes a bound on $\|\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}_i\|$, which thus induces the bounds on $\|\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}\|$ and $\text{var}(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2)$ tending to 0. It can be clearly noted that $\tilde{\boldsymbol{\beta}}_i$ approximates to $\hat{\boldsymbol{\beta}}$ sufficiently well with the accuracy bounded by $n^{-1}\text{polyLog}(n)$. If we exclude the n-th observation out, without loss of generality, by (3.10), the estimate $\tilde{\boldsymbol{\beta}}_n$ can be computed only based on the existing estimate (from the first $n-1$ observation) and a new observation (n-th observation). Together with the fact that $\|\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}_n\| = O_{L_k}(n^{-1}\text{polyLog}(n))$, an approximation of $\hat{\boldsymbol{\beta}}$ can be calculated if fast computing is required and/or data storage space is limited, which is very useful in the application of big data analysis.

Now we provide details on the formulation of $\tilde{\boldsymbol{\beta}}_i$. Firstly, we define

$$\boldsymbol{\phi}(\boldsymbol{\beta}) = -\frac{1}{n}\sum_{i=1}^{n} X_i^\top \boldsymbol{\psi}(\boldsymbol{e}_i + X_i\boldsymbol{\beta}_0 - X_i\boldsymbol{\beta}) + \tau\boldsymbol{\beta}. \tag{3.11}$$

Then we have

$$\boldsymbol{\phi}(\hat{\boldsymbol{\beta}}) = -\frac{1}{n}\sum_{i=1}^{n} X_i^\top \boldsymbol{\psi}(\boldsymbol{e}_i + X_i\boldsymbol{\beta}_0 - X_i\hat{\boldsymbol{\beta}}) + \tau\hat{\boldsymbol{\beta}} = \mathbf{0}, \tag{3.12}$$

and

$$\boldsymbol{\phi}(\hat{\boldsymbol{\beta}}_{(i)}) = -\frac{1}{n}\sum_{j\neq i} X_j^\top \boldsymbol{\psi}(\boldsymbol{e}_j + X_j\boldsymbol{\beta}_0 - X_j\hat{\boldsymbol{\beta}}_{(i)}) + \tau\hat{\boldsymbol{\beta}}_{(i)} = \mathbf{0}. \tag{3.13}$$

Take the first-order Taylor expansion to (3.12)-(3.13) of $\boldsymbol{\psi}(\boldsymbol{e}_j + X_j\boldsymbol{\beta}_0 - X_j\hat{\boldsymbol{\beta}})$ $(j \neq i)$ with respect to $\hat{\boldsymbol{\beta}}$ around $\hat{\boldsymbol{\beta}}_{(i)}$. By performing basic computations, we obtain that

$$\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)} \simeq \frac{1}{n}(S_i + \tau I_p)^{-1} X_i^\top \boldsymbol{\psi}(\boldsymbol{r}_i), \tag{3.14}$$

where '$\simeq$' denotes the approximation. Note that when we do the first-order Taylor expansion, the higher order terms are omitted. (3.14) establishes an approximating relationship of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_{(i)}$. When it comes to the issue that $\boldsymbol{r}_i$ is unknown since we exclude the i-th observation out, it is actually reasonable to replace $\boldsymbol{r}_i$ with its relatively proper approximation $\tilde{\boldsymbol{g}}^{-1}(\tilde{\boldsymbol{r}}_{i,[-i]})$. We thus define $\boldsymbol{\eta}_i = n^{-1}(S_i + \tau I_p)^{-1} X_i^\top \boldsymbol{\psi}(\tilde{\boldsymbol{g}}^{-1}(\tilde{\boldsymbol{r}}_{i,[-i]}))$, and the new quantity $\tilde{\boldsymbol{\beta}}_i$ follows.

3.2.3.2 Approximations of leaving one predictor out

We consider the leaving one predictor out estimate by omitting one of the predictors, say the p-th predictor. Before proceeding, we make some notations. Denote the first $p-1$ columns of X_i by $X_{i,-p} = (\boldsymbol{x}_{i,1}, \ldots, \boldsymbol{x}_{i,p-1})$, where $\boldsymbol{x}_{i,k}$ is the k-th column of X_i for $1 \le k \le p$. Write $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{0,-p}^\top, \beta_{0,p})^\top$, where $\boldsymbol{\beta}_{0,-p}$ denotes the first $p-1$ coordinates of $\boldsymbol{\beta}_0$. Let $\hat{\boldsymbol{\gamma}}$ be the leaving p-th predictor out estimate, i.e.,

$$\hat{\boldsymbol{\gamma}} = \arg\min_{\boldsymbol{\gamma}\in\mathbb{R}^{p-1}} \frac{1}{n}\sum_{i=1}^{n} \rho(\boldsymbol{e}_i + X_{i,-p}\boldsymbol{\beta}_{0,-p} - X_{i,-p}\boldsymbol{\gamma}) + \frac{\tau}{2}\|\boldsymbol{\gamma}\|^2. \tag{3.15}$$

Define the corresponding residuals as $\check{\boldsymbol{r}}_{i,-p} = \boldsymbol{e}_i + X_{i,-p}\boldsymbol{\beta}_{0,-p} - X_{i,-p}\hat{\boldsymbol{\gamma}}, i = 1, \ldots, n$. Denote $\Delta_p = n^{-1}\sum_i X_{i,-p}^\top \nabla\boldsymbol{\psi}(\check{\boldsymbol{r}}_{i,-p})X_{i,-p}$, $\boldsymbol{u}_p = n^{-1}\sum_i X_{i,-p}^\top \nabla\boldsymbol{\psi}(\check{\boldsymbol{r}}_{i,-p})\boldsymbol{x}_{i,p}$.

The aim of the work is to find an approximation of $\hat{\boldsymbol{\beta}}$. Decompose $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_{-p}^\top, \hat{\beta}_p)^\top$. We introduce another quantity $\tilde{\boldsymbol{b}} = (\tilde{\boldsymbol{b}}_{-p}^\top, \tilde{b}_p)^\top$,

$$\tilde{\boldsymbol{b}}_{-p} = \hat{\boldsymbol{\gamma}} - (\tilde{b}_p - \beta_{0,p})(\Delta_p + \tau I_{p-1})^{-1}\boldsymbol{u}_p, \tag{3.16}$$

$$\tilde{b}_p = \beta_{0,p}\frac{\xi_n}{\tau + \xi_n} + \frac{n^{-1}\sum_i \boldsymbol{x}_{i,p}^\top \boldsymbol{\psi}(\check{\boldsymbol{r}}_{i,-p})}{\tau + \xi_n}, \tag{3.17}$$

where $\xi_n = n^{-1}\sum_i \boldsymbol{x}_{i,p}^\top \nabla\boldsymbol{\psi}(\check{\boldsymbol{r}}_{i,-p})\boldsymbol{x}_{i,p} - \boldsymbol{u}_p^\top(\Delta_p + \tau I_{p-1})^{-1}\boldsymbol{u}_p$. Please be noted that the formulation of $\tilde{\boldsymbol{b}}$ is straightforward. We will briefly explain it afterward.

The following results yielded by leaving one predictor out are analogous to those in Theorem 3.2.

Theorem 3.3 *Under the assumptions (A1)-(A5), for any fixed* $\tau > 0$*, we have*

(i) $\|\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{b}}\| = O_{L_k}\left(\frac{\text{polyLog}(n)}{\min\{n,n^{2\alpha}\}}\right)$, *and* $|\hat{\beta}_p - \tilde{b}_p| = O_{L_k}\left(\frac{\text{polyLog}(n)}{\min\{n,n^{2\alpha}\}}\right)$.

(ii) $\sup_{1\leq i\leq n} \|X_i(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{b}})\| = O_{L_k}\left(\frac{n^{1/2}\text{polyLog}(n)}{\min\{n,n^{2\alpha}\}}\right)$.

(iii) $\sup_{1\leq i\leq n} \|\boldsymbol{r}_i - \check{\boldsymbol{r}}_{i,-p}\| = O_{L_k}\left(\frac{n^{1/2}\text{polyLog}(n)}{\min\{n,n^{2\alpha}\}}\right)$.

As pointed by El Karoui (2013, 2018), $\hat{\boldsymbol{\beta}}$ can be well approximated by $\tilde{\boldsymbol{b}}$ with the accuracy bounded by $\text{polyLog}(n)/\min\{n, n^{2\alpha}\}$. Moreover, the difference between $\check{\boldsymbol{r}}_{i,-p}$ and $\boldsymbol{r}_i$ is negligible. It is shown in the above theorem that those approximations still hold true under our model regime.

Now we briefly explain the formulation of $\tilde{\boldsymbol{b}}$. Denote the first $p-1$ coordinates and the p-th coordinate of $\boldsymbol{\phi}(\hat{\boldsymbol{\beta}})$ as $\boldsymbol{\phi}_{-p}(\hat{\boldsymbol{\beta}})$ and $\phi_p(\hat{\boldsymbol{\beta}})$, respectively. By the definition of $\hat{\boldsymbol{\gamma}}$, we have that $\boldsymbol{\phi}(\hat{\boldsymbol{\gamma}}) = -n^{-1}\sum_i X_{i,-p}^\top \boldsymbol{\psi}(\check{\boldsymbol{r}}_{i,-p}) + \tau\hat{\boldsymbol{\gamma}} = \mathbf{0}_{p-1}$, which, together with (3.12), yields that

$$\boldsymbol{\phi}_{-p}(\hat{\boldsymbol{\beta}}) = \frac{1}{n}\sum_i X_{i,-p}^\top(\boldsymbol{\psi}(\boldsymbol{r}_i) - \boldsymbol{\psi}(\check{\boldsymbol{r}}_{i,-p})) + \tau(\hat{\boldsymbol{\gamma}} - \hat{\boldsymbol{\beta}}_{-p}) = \mathbf{0}_{p-1}, \tag{3.18}$$

$$\phi_p(\hat{\boldsymbol{\beta}}) = \frac{1}{n}\sum_i \boldsymbol{x}_{i,p}^\top \boldsymbol{\psi}(\boldsymbol{r}_i) - \tau\hat{\beta}_p = 0. \tag{3.19}$$

We take the first-order Taylor expansions to (3.18) and (3.19) of $\boldsymbol{\psi}(\boldsymbol{r}_i)$ with respect to $\boldsymbol{r}_i$ around $\check{\boldsymbol{r}}_{i,-p}$, respectively. Together with the fact that $\boldsymbol{r}_i - \check{\boldsymbol{r}}_{i,-p} = X_{i,-p}(\hat{\boldsymbol{\gamma}} - \hat{\boldsymbol{\beta}}_{-p}) - (\hat{\beta}_p - \beta_{0,p})\boldsymbol{x}_{i,p}$, by basic computations, we have

$$\hat{\boldsymbol{\beta}}_{-p} \simeq \hat{\boldsymbol{\gamma}} - (\hat{\beta}_p - \beta_{0,p})(\Delta_p + \tau I_{p-1})^{-1}\boldsymbol{u}_p, \tag{3.20}$$

$$\hat{\beta}_p \simeq \beta_{0,p}\frac{\xi_n}{\tau+\xi_n} + \frac{n^{-1}\sum_i \boldsymbol{x}_{i,p}^\top \boldsymbol{\psi}(\check{\boldsymbol{r}}_{i,-p})}{\tau+\xi_n}. \tag{3.21}$$

Note that when we do the first-order Taylor expansion, the higher order terms are omitted. By (3.20) and (3.21), we thus define the quantity $\tilde{\boldsymbol{b}}$ as the approximation of $\hat{\boldsymbol{\beta}}$.

3.3 Details on double leave-one-out method

The double leave-one-out method allows us to understand the behavior of the ridge-regularized high-dimensional regression M-estimate $\hat{\boldsymbol{\beta}}$ in (3.5). Now we detail the procedures of the approach.

3.3.1 Derivation of leaving one observation out

We begin with $\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)} \simeq n^{-1}(S_i + \tau I_p)^{-1} X_i^\top \boldsymbol{\psi}(\boldsymbol{r}_i)$ in (3.14). Multiplying the expression above by X_i, we obtain that

$$\tilde{\boldsymbol{r}}_{i,[-i]} - \boldsymbol{r}_i \simeq \frac{1}{n} X_i (S_i + \tau I_p)^{-1} X_i^\top \boldsymbol{\psi}(\boldsymbol{r}_i). \tag{3.22}$$

Next, we show that

$$\tilde{\boldsymbol{r}}_{i,[-i]} - \boldsymbol{r}_i \simeq \mu \boldsymbol{\psi}(\boldsymbol{r}_i), \tag{3.23}$$

which is equal to show that the (s,t)-th entry of $n^{-1}X_i(S_i + \tau I_p)^{-1}X_i^\top$ (denoted as c_{st}) can be approximated by $\mu I_{\{s=t\}}$. Let $\boldsymbol{x}_i(s)$ and $\boldsymbol{x}_i(t)$ be the s-th and t-th rows of X_i, $s, t = 1, \ldots, m$, respectively. We thus have

$$c_{st} = n^{-1}\boldsymbol{x}_i^\top(s)(S_i + \tau I_p)^{-1}\boldsymbol{x}_i(t) \simeq \operatorname{tr}(n^{-1}(S_i + \tau I_p)^{-1})I_{\{s=t\}} \simeq \mu I_{\{s=t\}}.$$

The first '$\simeq$' in the above expression follows by (B.8). For the second '$\simeq$', it can be proved by the following two steps. Firstly, we show that $c_i = \operatorname{tr}(n^{-1}(S_i + \tau I_p)^{-1}) \simeq \operatorname{tr}(n^{-1}(S + \tau I_p)^{-1}) = c_\tau$ in view of (B.8) and (B.9). Secondly, we prove that c_τ is asymptotically deterministic (see Proposition B.8 in Appendix B.1). And μ is the limit of c_τ as both $n, p \to \infty$ with $p/n \to \kappa$.

3.3.2 Derivation of leaving one predictor out

Before proceeding, we state the following Sherman-Morrison-Woodbury (SMW) formula. For a matrix U and an invertible matrix A,

$$(A + UU^\top)^{-1} = A^{-1} - A^{-1}U(I + U^\top A^{-1}U)^{-1}U^\top A^{-1}.$$

Specially,

$$U^\top(A + UU^\top)^{-1}U = I - (I + U^\top A^{-1}U)^{-1}. \tag{3.24}$$

Denote $X_{-p} = (X_{1,-p}^\top, \ldots, X_{n,-p}^\top)^\top$, $\boldsymbol{x}_p = (\boldsymbol{x}_{1,p}^\top, \ldots, \boldsymbol{x}_{n,p}^\top)^\top$, and

$$D = \begin{pmatrix} \nabla\boldsymbol{\psi}(\check{\boldsymbol{r}}_{1,-p}) & & \\ & \ddots & \\ & & \nabla\boldsymbol{\psi}(\check{\boldsymbol{r}}_{n,-p}) \end{pmatrix}. \tag{3.25}$$

Then $\Delta_p = n^{-1}X_{-p}^\top DX_{-p}$ and $\boldsymbol{u}_p = n^{-1}X_{-p}^\top D\boldsymbol{x}_p$. We define that $B = n^{-1/2}D^{1/2}X_{-p}$ and $P = B(B^\top B + \tau I_{p-1})^{-1}B^\top$. By the definitions of Δ_p and $\boldsymbol{u}_p$, it is easy to deduce that

$$\begin{aligned}\boldsymbol{u}_p^\top(\Delta_p + \tau I_{p-1})^{-1}\boldsymbol{u}_p =& n^{-2}\boldsymbol{x}_p^\top DX_{-p}(n^{-1}X_{-p}{}^\top DX_{-p} + \tau I_{p-1})^{-1}X_{-p}{}^\top D\boldsymbol{x}_p \\ =& n^{-1}\boldsymbol{x}_p^\top D^{1/2}PD^{1/2}\boldsymbol{x}_p.\end{aligned}$$

Denote $\Delta_p(i) = \Delta_p - n^{-1}X_{i,-p}^\top \nabla\boldsymbol{\psi}(\check{\boldsymbol{r}}_{i,-p})X_{i,-p}$, and

$$P = \begin{pmatrix} P_{11} & & \\ & \ddots & \\ & & P_{nn} \end{pmatrix}, \tag{3.26}$$

where P_{ii} is the i-th entry matrix on the diagonal of P $(i = 1, \ldots, n)$, i.e.,

$$\begin{aligned}P_{ii} =& n^{-1}\nabla\boldsymbol{\psi}^{1/2}(\check{\boldsymbol{r}}_{i,-p})X_{i,-p}\left(\Delta_p(i) + \tau I_{p-1} + n^{-1}X_{i,-p}^\top\nabla\boldsymbol{\psi}(\check{\boldsymbol{r}}_{i,-p})X_{i,-p}\right)^{-1}X_{i,-p}^\top\nabla\boldsymbol{\psi}^{1/2}(\check{\boldsymbol{r}}_{i,-p}) \\ =& I_m - \Big(I_m + n^{-1}\nabla\boldsymbol{\psi}^{1/2}(\check{\boldsymbol{r}}_{i,-p})X_{i,-p}(\Delta_p(i) + \tau I_{p-1})^{-1}X_{i,-p}^\top\nabla\boldsymbol{\psi}^{1/2}(\check{\boldsymbol{r}}_{i,-p})\Big)^{-1}.\end{aligned} \tag{3.27}$$

The second '=' results from the SMW formula in (3.24) upon choosing $A = \Delta_p(i) + \tau I_{p-1}, U = X_{i,-p}^\top\nabla\boldsymbol{\psi}^{1/2}(\check{\boldsymbol{r}}_{i,-p})$. It follows that,

$$\begin{aligned}I_m - P_{ii} =& \Big(I_m + n^{-1}\nabla\boldsymbol{\psi}^{1/2}(\check{\boldsymbol{r}}_{i,-p})X_{i,-p}(\Delta_p(i) + \tau I_{p-1})^{-1}X_{i,-p}^\top\nabla\boldsymbol{\psi}^{1/2}(\check{\boldsymbol{r}}_{i,-p})\Big)^{-1} \\ \simeq& (I_m + \mu\nabla\boldsymbol{\psi}(\check{\boldsymbol{r}}_{i,-p}))^{-1},\end{aligned} \tag{3.28}$$

where $n^{-1}X_{i,-p}(\Delta_p(i)+\tau I_{p-1})^{-1}X_{i,-p}^\top \simeq \mathrm{tr}(n^{-1}(\Delta_p(i)+\tau I_{p-1})^{-1})I_m \simeq \mathrm{tr}(n^{-1}(\Delta_p + \tau I_{p-1})^{-1})I_m \simeq \mu I_m$. In view of the above definitions of P, Δ_p and B,

$$\mathrm{tr}(P) = \mathrm{tr}((\Delta_p + \tau I_{p-1})^{-1}\Delta_p) = p-1-\tau\mathrm{tr}((\Delta_p+\tau I_{p-1})^{-1}) \simeq p-1-n\tau\mu,$$

and thus, $\sum_{i=1}^n \mathrm{tr}(I_m - P_{ii}) = nm - [(p-1)-n\mu\tau] \simeq nm - p + n\mu\tau$, which, together with (3.28), entails that

$$\frac{1}{n}\sum_{i=1}^n \mathrm{tr}([I_m + \mu\nabla\boldsymbol{\psi}(\check{\boldsymbol{r}}_{i,-p})]^{-1}) \simeq m - \frac{p}{n} + \tau\mu. \tag{3.29}$$

Define the vector valued function $\boldsymbol{g}_\mu(\boldsymbol{x}) = \boldsymbol{x} + \mu\boldsymbol{\psi}(\boldsymbol{x}) = \boldsymbol{z}$. Then $\boldsymbol{x} = \boldsymbol{g}_\mu^{-1}(\boldsymbol{z}) = \mathrm{prox}_\mu(\rho)(\boldsymbol{z})$ and $\nabla\mathrm{prox}_\mu(\rho)(\boldsymbol{z}) = (I_m + \mu\nabla\boldsymbol{\psi}(\boldsymbol{z}))^{-1}$, which jointly with (3.23), (3.29), Theorem 3.3(iii) and Proposition B.7, concludes the first equation in (3.7).

We now drive the second equation in our system. Considering ξ_n in (3.17), it follows that,

$$\begin{aligned}
\xi_n =& \frac{1}{n}\sum_i \boldsymbol{x}_{i,p}^\top \nabla\boldsymbol{\psi}^{1/2}(\check{\boldsymbol{r}}_{i,-p})(I_m - P_{ii})\nabla\boldsymbol{\psi}^{1/2}(\check{\boldsymbol{r}}_{i,-p})\boldsymbol{x}_{i,p} \\
\simeq& \frac{1}{n}\sum_i \boldsymbol{x}_{i,p}^\top \nabla\boldsymbol{\psi}^{1/2}(\check{\boldsymbol{r}}_{i,-p})\left(I_m + \mu\nabla\boldsymbol{\psi}(\check{\boldsymbol{r}}_{i,-p})\right)^{-1}\nabla\boldsymbol{\psi}^{1/2}(\check{\boldsymbol{r}}_{i,-p})\boldsymbol{x}_{i,p} \\
=& \frac{1}{n\mu}\sum_i \boldsymbol{x}_{i,p}^\top \left(I_m - (I_m + \mu\nabla\boldsymbol{\psi}(\check{\boldsymbol{r}}_{i,-p}))^{-1}\right)\boldsymbol{x}_{i,p} \\
\simeq& \frac{1}{n\mu}\sum_i \mathrm{tr}(I_m - (I_m + \mu\nabla\boldsymbol{\psi}(\check{\boldsymbol{r}}_{i,-p}))^{-1}) \\
\simeq& \frac{1}{n\mu}(mn - mn + p - n\tau\mu) = \frac{p}{n\mu} - \tau.
\end{aligned} \tag{3.30}$$

The first and the last '$\simeq$' are derived from (3.28) and (3.29), respectively. Since $I_m - (I_m + \mu \nabla \boldsymbol{\psi}(\check{\boldsymbol{r}}_{i,-p}))^{-1}$ is a random symmetric matrix depending only on $X_{i,-p}$, by Lemma 3.37 in El Karoui (2018), the second approximation '$\simeq$' holds. The second '$=$' follows by applying the SMW formula given in (3.24). Plugging the approximation $\xi_n + \tau \simeq p/(n\mu)$ into (3.17) entails that

$$\sqrt{n}\mu(\xi_n + \tau)(\hat{\beta}_p - \beta_{0,p}) \simeq \frac{1}{\sqrt{n}} \sum_i \mu \boldsymbol{x}_{i,p}^\top \boldsymbol{\psi}(\check{\boldsymbol{r}}_{i,-p}) - \sqrt{n}\mu\tau\beta_{0,p}.$$

Since $\check{\boldsymbol{r}}_{i,-p}$ is independent of $\boldsymbol{x}_{i,p}$, $E\boldsymbol{x}_{i,p} = \boldsymbol{0}$, $\mathrm{Cov}(\boldsymbol{x}_{i,p}) = I_m$, and $\boldsymbol{r}_i \simeq \check{\boldsymbol{r}}_{i,-p}$ (by Theorem 3.3), we have

$$\begin{aligned} E\left(\left(\frac{p}{n}\right)^2 n(\hat{\beta}_p - \beta_{0,p})^2 | \{X_{i,-p}, \boldsymbol{e}_i; i = 1, \ldots, n\}\right) \simeq & \frac{1}{n}\sum_{i=1}^n \|\mu\boldsymbol{\psi}(\check{\boldsymbol{r}}_{i,-p})\|^2 + n\mu^2\tau^2\beta_{0,p}^2 \\ \simeq & \frac{1}{n}\sum_{i=1}^n \|\mu\boldsymbol{\psi}(\boldsymbol{r}_i)\|^2 + n\mu^2\tau^2\beta_{0,p}^2. \end{aligned}$$

Summing over all coordinates, it follows that

$$E\left(\left(\frac{p}{n}\right)^2 \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2\right) \simeq \frac{p}{n}\frac{1}{n}\sum_{i=1}^n E\|\mu\boldsymbol{\psi}(\boldsymbol{r}_i)\|^2 + \mu^2\tau^2\|\boldsymbol{\beta}_0\|^2. \tag{3.31}$$

By (3.23), $\boldsymbol{\psi}(\boldsymbol{r}_i) \simeq \mu^{-1}(\tilde{\boldsymbol{r}}_{i,[-i]} - \boldsymbol{r}_i)$, (3.31) therefore becomes

$$\left(\frac{p}{n}\right)^2 E\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2 \simeq \frac{p}{n}\frac{1}{n}\sum_i E\|\tilde{\boldsymbol{r}}_{i,[-i]} - \boldsymbol{r}_i\|^2 + \mu^2\tau^2\|\boldsymbol{\beta}_0\|^2.$$

By Theorem 3.2 and Proposition B.7 in Appendix B.1, the second equation in (3.7) follows.

3.4 Examples

We present some examples to explore the behaviors of ridge-regularized ($\tau > 0$) and/or unregularized ($\tau = 0$) high-dimensional regression M-estimates via the nonlinear systems stated in Theorem 3.1 and Corollary 3.1.

3.4.1 $\rho(\boldsymbol{x}) = \boldsymbol{x}^\top A\boldsymbol{x}/2$

In this example, $\rho(\boldsymbol{x}) = \boldsymbol{x}^\top A\boldsymbol{x}/2$, $A \succ 0$, and $\boldsymbol{\psi}(\boldsymbol{x}) = A\boldsymbol{x}$. We thus have $\text{prox}_\mu(\rho)(\boldsymbol{z}) = (I_m + \mu A)^{-1}\boldsymbol{z}$ and $\nabla\text{prox}_\mu(\rho)(\boldsymbol{z}) = (I_m + \mu A)^{-1}$. Assume that $\boldsymbol{z}$ has mean vector $\mathbf{0}$ and covariance matrix, $Cov(\boldsymbol{z}) = \Sigma_{\boldsymbol{e}} + E\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2 I_m$ (Cov is the abbreviation of covariance). Hence,

$$E\|\boldsymbol{z} - \text{prox}_\mu(\rho)(\boldsymbol{z})\|^2 = \text{tr}((I_m - (I_m + \mu A)^{-1})^2\Sigma_{\boldsymbol{e}}) + \text{tr}((I_m - (I_m + \mu A)^{-1})^2)E\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2.$$

Together with (3.7) in Theorem 3.1, by elementary computations, we obtain that, for a given $\tau > 0$,

$$\text{tr}((I_m + \mu A)^{-1}) = m - \frac{p}{n} + \tau\mu,$$

and

$$E\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2 = \frac{\frac{p}{n}\text{tr}((I_m - (I_m + \mu A)^{-1})^2\Sigma_{\boldsymbol{e}}) + \tau^2\|\boldsymbol{\beta}_0\|^2\mu^2}{\frac{p^2}{n^2} - \frac{p}{n}\text{tr}((I_m - (I_m + \mu A)^{-1})^2)}.$$

When $\tau = 0$, it is reduced to the result in (3.8). We thus obtain μ and $E\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2$ by solving the above two equations. One should note that the risk of $\hat{\boldsymbol{\beta}}$ depends on the covariance matrix of $\boldsymbol{e}$, say $\Sigma_{\boldsymbol{e}}$, rather than its distribution.

We explore and compare the risks of the high-dimensional regression M-estimates that we obtain using system prediction (R) and numerical simulations under the cases of multivariate normal errors ($ER - \mathcal{N}$) and multivariate t errors ($ER - \mathcal{T}$). We details the simulation settings as follows.

- Let $m = 5; n = 100; p/n = 0.1, 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5; \tau = 0, 0.1, 1, 10$.

- Consider $\boldsymbol{e}_i \sim \mathcal{N}(\mathbf{0}, \Sigma_{\boldsymbol{e}})$ and $\boldsymbol{e}_i \sim \mathcal{T}_3(\mathbf{0}, \Sigma_{\boldsymbol{e}}/3)$, where $\Sigma_{\boldsymbol{e}} = (\sigma_{\ell j})$, the diagonal entries $(\sigma_{11}, \sigma_{22}, \sigma_{33}, \sigma_{44}, \sigma_{55})^\top = (1, 1, 1.2, 1.5, 1.1)^\top$ and else $\sigma_{\ell j} = 0.8^{|\ell - j|}$ for $\ell \neq j$. $\mathcal{T}_3(\mathbf{0}, \Sigma_{\boldsymbol{e}}/3)$ denotes a multivariate t-distribution with degree of freedom 3 and covariance matrix $\Sigma_{\boldsymbol{e}}$.

- Let $A = \Sigma_{\boldsymbol{e}}^{-1}$, $\boldsymbol{\beta}_0 = (0.5, \ldots, 0.5) \in \mathbb{R}^p$. Note that each element of $\boldsymbol{\beta}_0$ is $5/\sqrt{n}$. We can also set each entry to be $c/\sqrt{p}$.

- For simplicity, we limit our attention to the case that $X_i \sim \mathcal{MN}_{m \times p}(\mathbf{0}, I_m, I_p)$ for $i = 1, \ldots, n$.

We carry out 500 simulations. Table 3.1 displays the risks of $\hat{\boldsymbol{\beta}}$ estimated by the system and numerical simulations. As expected, for a given $\tau \geq 0$, the discrepancies between the system prediction risk R and the empirical risk by simulations, $ER - \mathcal{N}$ (or $ER - \mathcal{T}$), are relatively small, which implies that the nonlinear system captures the behaviors of M-estimate very well under our model settings.

Table 3.1: $\rho(\boldsymbol{x}) = \boldsymbol{x}^\top A\boldsymbol{x}/2$: results for the risks of $\hat{\boldsymbol{\beta}}$ obtained by system prediction (R) and numerical simulations in the cases of multivariate normal errors ($ER - \mathcal{N}$) and multivariate t errors ($ER - \mathcal{T}$) for different τ, p/n.

τ	Method	p/n									
		0.1	0.5	1	1.5	2	2.5	3	3.5	4	4.5
0	R	0.170	0.863	1.657	2.280	2.779	3.274	3.909	4.912	6.874	12.696
	$ER - \mathcal{N}$	0.167	0.855	1.636	2.256	2.787	3.271	3.952	4.941	6.972	12.827
	$ER - \mathcal{T}$	0.167	0.898	1.534	2.318	2.716	3.193	3.779	5.369	6.850	12.372
0.1	R	0.165	0.830	1.580	2.167	2.653	3.173	3.908	5.127	7.255	10.855
	$ER - \mathcal{N}$	0.166	0.842	1.571	2.203	2.668	3.177	3.919	5.139	7.325	10.929
	$ER - \mathcal{T}$	0.163	0.817	1.483	2.152	2.873	3.252	3.790	5.177	7.174	10.571
1	R	0.181	1.010	2.348	4.157	6.598	9.800	13.844	18.763	24.547	31.152
	$ER - \mathcal{N}$	0.180	1.025	2.364	4.195	6.637	9.857	13.731	18.701	24.612	31.157
	$ER - \mathcal{T}$	0.178	1.002	2.317	4.286	6.658	9.883	13.863	18.680	24.411	31.159
10	R	1.050	5.561	11.918	18.949	26.546	34.592	43.061	51.869	60.963	70.307
	$ER - \mathcal{N}$	1.043	5.552	11.920	18.929	26.480	34.691	43.093	51.772	61.037	70.510
	$ER - \mathcal{T}$	1.047	5.526	11.897	18.932	26.510	34.611	43.163	51.981	60.964	70.299

3.4.2 $\rho(\boldsymbol{x}) = \sum_{\ell=1}^{m} \rho(x_\ell)$

For simplicity, in this example, we consider a discrepancy function of special type that was studied by Koenker and Portnoy (1990): $\rho(\boldsymbol{x}) = \sum_{\ell=1}^{m} \rho(x_\ell)$, where $\rho(x_\ell)$ is a univariate convex function, say, L_1 discrepancy function $\rho(x_\ell) = |x_\ell|$, or Huber

Table 3.2: Results for the risks of $\hat{\boldsymbol{\beta}}$ obtained by system prediction ($R_{[\mathcal{H}]}$) and by numerical simulations ($ER_{[\mathcal{H}]}$) in the case of Huber discrepancy function and multivariate normal errors for different $\tau, p/n$.

τ	Method	p/n									
		0.1	0.5	1	1.5	2	2.5	3	3.5	4	4.5
0	$R_{[\mathcal{H}]}$	0.025	0.135	0.303	0.517	0.799	1.191	1.773	2.736	4.656	10.441
	$ER_{[\mathcal{H}]}$	0.026	0.137	0.307	0.523	0.811	1.202	1.783	2.752	4.732	10.571
0.1	$R_{[\mathcal{H}]}$	0.025	0.137	0.309	0.528	0.819	1.224	1.819	2.762	4.382	7.326
	$ER_{[\mathcal{H}]}$	0.025	0.138	0.316	0.527	0.823	1.229	1.847	2.777	4.415	7.381
1	$R_{[\mathcal{H}]}$	0.126	0.838	2.499	5.508	10.058	15.925	22.798	30.431	38.655	47.351
	$ER_{[\mathcal{H}]}$	0.127	0.845	2.505	5.534	10.032	15.855	22.681	30.315	38.617	47.343
10	$R_{[\mathcal{H}]}$	1.493	9.302	20.199	31.500	43.004	54.637	66.359	78.149	89.992	101.878
	$ER_{[\mathcal{H}]}$	1.493	9.291	20.182	31.493	42.993	54.637	66.364	78.135	90.001	101.887

discrepancy function,

$$\rho_k(x_\ell) = \begin{cases} x_\ell^2/2 & if\ |x_\ell| \leq k, \\ k(|x_\ell| - \frac{k}{2}) & if\ |x_\ell| > k, \end{cases}$$

where $k > 0$ shows where the transitions from quadratics to linear take place. The Huber discrepancy function becomes more similar to $|\cdot|$, L_1 discrepancy function, for small values of k. Therefore, we focus only on the case of Huber discrepancy function. We have, $\psi_k(x_\ell) = x_\ell I_{\{|x_\ell| \leq k\}} + k\text{sign}(x_\ell) I_{\{|x_\ell|>k\}}$ for $\ell = 1, \ldots, m$, and

$$\text{prox}_\mu(\rho_k)(z_\ell) = \begin{cases} z_\ell/(1+\mu) & if\ |z_\ell| \leq (1+\mu)k, \\ z_\ell - \mu k \text{sign}(z_\ell) & if\ |z_\ell| > (1+\mu)k. \end{cases}$$

Assume that $\boldsymbol{e}_i \overset{iid}{\sim} \mathcal{N}(\mathbf{0}, \Sigma_{\boldsymbol{e}})$ for $i = 1, \ldots, n$. Denote $s_\ell^2 = \sigma_\ell^2 + E\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2$, and $\alpha_\ell = (1+\mu)k/s_\ell$, where σ_ℓ^2 is the ℓ-th diagonal entry in $\Sigma_{\boldsymbol{e}}$. Then $z_\ell \sim N(0, s_\ell^2)$, and $Ez_\ell^2 I_{\{|z_\ell| \leq (1+\mu)k\}} = s_\ell^2(2\Phi(\alpha_\ell) - 1 - 2\alpha_\ell \phi(\alpha_\ell))$, where ϕ and Φ are respective the standard normal density and distribution function. Thus the first and second equations in (3.7) become

$$\sum_{\ell=1}^{m} \Phi(\alpha_\ell) = \frac{1}{2}\left(m + \frac{1+\mu}{\mu}\left(\frac{p}{n} - \mu\tau\right)\right),$$

and

$$\begin{aligned} &\frac{p}{n}E\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2 - \frac{n}{p}\mu^2\tau^2\|\boldsymbol{\beta}_0\|^2 \\ =& \left(\frac{\mu}{1+\mu}\right)^2 \sum_{\ell=1}^{m} s_\ell^2(2\Phi(\alpha_\ell) - 1 - 2\alpha_\ell\phi(\alpha_\ell)) + (\mu k)^2\left(m - \left(\frac{p}{n} - \mu\tau\right)\frac{1+\mu}{\mu}\right), \end{aligned}$$

respectively. Now we compare the risk estimated by the system to the empirical risk obtained by numerical simulations. The settings of $m, n, p/n, \tau, \Sigma_{\boldsymbol{e}}, \ \boldsymbol{\beta}_0, X_i$ are the same as in Section 3.4.1. In the case of Huber discrepancy function and $\boldsymbol{e}_i \sim \mathcal{N}(\mathbf{0}, \Sigma_{\boldsymbol{e}})$, we denote the system prediction risk by $R_{[\mathcal{H}]}$ and the empirical risk by $ER_{[\mathcal{H}]}$. Herein we take $k = 1.345$. The simulations are repeated 500 times, and the results are presented in Table 3.2. It can be seen from the table that our system predictions match very well with the numerical simulation results, which demonstrates the capacity of our system prediction in capturing the behaviors of $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2$ in expectation. We remark that the common distribution of random errors can also be extended to a general symmetric distribution. We omit the case here. For more details on the univariate case, one can refer to El Karoui et al. (2013).

3.4.3 $\rho(\boldsymbol{x}) = \|\boldsymbol{x}\|$

The third case that we are interested in is $\rho(\boldsymbol{x}) = \|\boldsymbol{x}\|$ that was studied by Bai et al. (1992). In this case, $\boldsymbol{\psi}(\boldsymbol{x}) = \boldsymbol{x}/\|\boldsymbol{x}\|$ (except at $\mathbf{0}$). Then $\boldsymbol{x} + \mu\boldsymbol{\psi}(\boldsymbol{x}) = \boldsymbol{x} + \mu\boldsymbol{x}/\|\boldsymbol{x}\| = \boldsymbol{x}(1 + \mu/\|\boldsymbol{x}\|), (\boldsymbol{x} \neq \mathbf{0})$. Define $\boldsymbol{z} = \boldsymbol{x}(1 + \mu/\|\boldsymbol{x}\|)$. We thus have $\|\boldsymbol{x}\|(1 + \mu/\|\boldsymbol{x}\|) = \|\boldsymbol{z}\|$, or, $\|\boldsymbol{x}\| = \|\boldsymbol{z}\| - \mu$. It follows that

$$\boldsymbol{x} = \text{prox}_\mu(\rho)(\boldsymbol{z}) = \boldsymbol{z} - \mu\frac{\boldsymbol{z}}{\|\boldsymbol{z}\|} \qquad (\boldsymbol{z} \neq \mathbf{0}),$$

and

$$\nabla\text{prox}_\mu(\rho)(\boldsymbol{z}) = I_m - \mu\frac{1}{\|\boldsymbol{z}\|}\left(I_m - \frac{\boldsymbol{z}\boldsymbol{z}^\top}{\|\boldsymbol{z}\|^2}\right).$$

One can obtain the first equation in (3.7)

$$E\text{tr}\left(I_m - \mu\frac{1}{\|\boldsymbol{z}\|}\left(I_m - \frac{\boldsymbol{z}\boldsymbol{z}^\top}{\|\boldsymbol{z}\|^2}\right)\right) = m - \mu(m-1)E\frac{1}{\|\boldsymbol{z}\|} = m - \frac{p}{n} + \tau\mu,$$

namely $\mu(m-1)E(\|\boldsymbol{z}\|^{-1}) = p/n - \tau\mu$. On the other hand, since $E\|\boldsymbol{z} - \text{prox}_\mu(\rho)(\boldsymbol{z})\|^2 = E\|\boldsymbol{z} - (\boldsymbol{z} - \mu\boldsymbol{z}/\|\boldsymbol{z}\|)\|^2 = \mu^2$, thus

$$\left(\frac{p}{n}\right)^2 E\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2 = \frac{p}{n}E\|\boldsymbol{z} - \text{prox}_\mu(\rho)(\boldsymbol{z})\|^2 + \tau^2\|\boldsymbol{\beta}_0\|^2\mu^2 = \left(\frac{p}{n} + \tau^2\|\boldsymbol{\beta}_0\|^2\right)\mu^2,$$

which, jointly with $\mu(m-1)E(\|\boldsymbol{z}\|^{-1}) = p/n - \tau\mu$, deduces $E\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2$ and μ for $\tau \geq 0$.

4 General matching quantiles M-estimation

4.1 Introduction

MQE was first proposed by Sgouropoulos et al. (2015) as a way to address the problem of estimating representative portfolios for backtesting counterparty credit risks. However, financial applications can entail many challenges stemming from the complexity of data. For example, financial modeling with outliers may result in naive interpretation of statistics and unreliable scientific conclusions, which may further lead to large economic loss. To overcome the challenge, a robust method is developed to construct representative portfolios.

In this chapter, we propose a general enhancement of MQE by replacing the OLS estimation with an M-estimation. We show that in addition to being resistant to outliers, the MQME estimate is consistent, as is MQE. The proposed MQME can handle situations when both n and p are large, but the number of informative variables is small. This is common in many modern problems. This suggests that a ‘sparse’ matching quantiles estimate is highly desirable. Therefore, a sparse MQME is also developed by combining MQME with an adaptive Lasso penalty. As with the original MQME, we expect the ‘sparse’ variant to be robust to outlier observations.

The rest of this chapter is organized as follows. In Section 4.2, we introduce the MQME method. We discuss its theoretical properties in Section 4.3. Numerical experiments of varying designs are explored in Section 4.4, followed by a real case study of the stock market index in Hong Kong during the period of 2013-2016 in Section 4.5. All proofs to any presented theoretical results can be found in Appendix C.

4.2 The methods

4.2.1 Matching Quantiles M-Estimation (MQME)

Consider all linear combinations of p random variables $\{X_1, \ldots, X_p\}$. The goal is to match the distributions of Y and $\boldsymbol{\beta}^\top \boldsymbol{X}$ for some $\boldsymbol{\beta}$, i.e.,

$$\mathcal{L}(Y) = \mathcal{L}(\boldsymbol{\beta}^\top \boldsymbol{X}). \tag{4.1}$$

A straightforward approach one could take is to match the distribution (or probability density) functions of Y and $\boldsymbol{\beta}^\top \boldsymbol{X}$ by only matching the center parts of their distributions well because both distributions are close to 1 or 0 for extremely large or small values. In some fields including risk management, however, those extreme values may be very important and useful. It is noted that the distribution of Y is rarely known. In order to find a $\boldsymbol{\beta}$ such that the distribution of $\boldsymbol{\beta}^\top \boldsymbol{X}$ matches the distribution of Y not only in the middle but also at the tails, Sgouropoulos et al. (2015) proposed searching for $\boldsymbol{\beta}$ by minimizing the squared difference of the two quantiles functions across all levels, given by

$$\tilde{S}(\boldsymbol{\beta}) = \int_0^1 \Big(Q_Y(\alpha) - Q_{\boldsymbol{\beta}^\top \boldsymbol{X}}(\alpha)\Big)^2 d\alpha, \tag{4.2}$$

and defined the matching quantiles estimate $\tilde{\boldsymbol{\beta}}_n$ of $\boldsymbol{\beta}$ to be the one minimizing the following matching sample quantiles

$$\tilde{S}_n(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \Big(Y_{(i)} - (\boldsymbol{\beta}^\top \boldsymbol{X})_{(i)}\Big)^2. \tag{4.3}$$

However, in some real applications, we may encounter outliers due to uncontrollable factors. For such data, procedures based on OLS estimation behave badly (El Karoui et al., 2013; Huber, 1973). Therefore, we develop a more general matching quantiles estimation procedure based on M-estimation, i.e., by searching for $\boldsymbol{\beta}$ such that it minimizes

$$\check{S}(\boldsymbol{\beta}) = \int_0^1 \rho\Big(Q_Y(\alpha) - Q_{\boldsymbol{\beta}^\top \boldsymbol{X}}(\alpha)\Big)\, d\alpha \tag{4.4}$$

for a convex discrepancy function $\rho(\cdot)$ that satisfies the assumption (A1) in Section 4.3. We define the matching quantiles M-estimate $\check{\boldsymbol{\beta}}_n$ of $\boldsymbol{\beta}$ to be the one minimizing the following matching sample quantiles

$$\check{S}_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \rho\Big(Y_{(i)} - (\boldsymbol{\beta}^\top \boldsymbol{X})_{(i)}\Big). \tag{4.5}$$

The particular cases of $\rho(\cdot)$ are respectively the L_1 discrepancy function, and the Huber discrepancy function $\rho_c(\cdot)$ given in (4.7), which have been extensively studied and applied in modeling against outliers (Lambert-Lacroix and Zwald, 2011). Specifically, the L_1 matching quantiles estimate $\check{\boldsymbol{\beta}}_{L_1}$ is defined by minimizing the convex function

$$\check{S}_n^{L_1}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \Big|Y_{(i)} - (\boldsymbol{\beta}^\top \boldsymbol{X})_{(i)}\Big|. \tag{4.6}$$

We now define the Huber discrepancy function-based matching quantiles estimate $\check{\boldsymbol{\beta}}_H$. For any $c > 0$, the Huber discrepancy function is defined as

$$\rho_c(x) = \begin{cases} x^2/2 & \text{if } |x| \leq c, \\ \\ c(|x| - c/2) & \text{if } |x| > c. \end{cases} \tag{4.7}$$

The parameter c shows where the transitions from quadratics to linear take place. The Huber discrepancy function $\rho_c(\cdot)$ becomes more similar to the L_1 discrepancy function, for small values of c while it becomes more similar to $(\cdot)^2$, i.e., the L_2 discrepancy function, for large values of c (Lambert-Lacroix and Zwald, 2011). The Huber discrepancy function-based matching quantiles estimate $\check{\boldsymbol{\beta}}_H$ is defined by minimizing the convex function, i.e.,

$$\check{S}_n^H(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\rho_c\Big(Y_{(i)} - (\boldsymbol{\beta}^\top \boldsymbol{X})_{(i)}\Big). \tag{4.8}$$

It is sensible that the choice of c has an impact on the estimation, and thus the c value should be properly chosen. The details on how to select c will be given in Section 4.2.4.

4.2.2 Sparse MQME

In recent years, there has been a great focus on the sparse modeling where p and/or n are very large. A 'sparse' matching quantiles estimate is highly desirable. We thus combine the MQME with the adaptive Lasso penalty, which is named as sparse MQME. The purpose of sparse MQME is to search for $\boldsymbol{\beta}$ such that it minimizes $S(\boldsymbol{\beta})$, i.e.,

$$S(\boldsymbol{\beta}) = \int_0^1 \rho\Big(Q_Y(\alpha) - Q_{\boldsymbol{\beta}^\top \boldsymbol{X}}(\alpha)\Big)d\alpha + \lambda\sum_{j=1}^{p}\omega_j|\beta_j|. \tag{4.9}$$

The sparse matching quantiles M-estimate $\hat{\boldsymbol{\beta}}_n$ is thus obtained by minimizing

$$S_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\rho\Big(Y_{(i)} - (\boldsymbol{\beta}^\top \boldsymbol{X})_{(i)}\Big) + \lambda\sum_{j=1}^{p}\omega_j|\beta_j|, \tag{4.10}$$

where $\lambda \geq 0$ and $\omega_j \geq 0$, $j = 1, 2, \ldots, p$, are tuning parameters. Zou (2006) showed that the adaptive Lasso enjoys the consistency and oracle property in variable selection by choosing a proper λ and ω_j, $j = 1, 2, \ldots, p$. The procedure of how to select proper tuning parameters will be discussed in Section 4.2.4.

Remark 4.1 *In practice, one may interest in matching a part of distribution of Y, say, that between the a_nth quantile and the b_nth quantile, $0 \leq a_n < b_n \leq 1$. Herein both a_n and b_n are dependent on n, while they both are fixed in Sgouropoulos et al. (2015). We hence replace (4.9) and (4.10) by*

$$S(\boldsymbol{\beta}; a_n, b_n) = \int_{a_n}^{b_n} \rho\Big(Q_Y(\alpha) - Q_{\boldsymbol{\beta}^\top \boldsymbol{X}}(\alpha)\Big) d\alpha + \lambda \sum_{j=1}^{p} \omega_j |\beta_j|, \tag{4.11}$$

and

$$S_n(\boldsymbol{\beta}; n_1, n_2) = \frac{1}{n} \sum_{i=n_1+1}^{n_2} \rho\Big(Y_{(i)} - (\boldsymbol{\beta}^\top \boldsymbol{X})_{(i)}\Big) + \lambda \sum_{j=1}^{p} \omega_j |\beta_j|, \tag{4.12}$$

respectively. In (4.12), $n_1(n) = [na_n]$ and $n_2(n) = [nb_n]$, where $[x]$ denotes the greatest integer less than or equal to x. Note that $n_1(n), n_2(n)$ are dependent on n. We drop the suffix n for notational convenience.

4.2.3 Iterative algorithm for (sparse) MQME

The matching quantiles M-estimate $\hat{\boldsymbol{\beta}}_n$ does not admit an explicit solution. In light of Sgouropoulos et al. (2015), we provide an iterative algorithm for computing $\hat{\boldsymbol{\beta}}_n$ that minimizes $S_n(\boldsymbol{\beta})$ in (4.10), which also works for computing $\check{\boldsymbol{\beta}}_n$ as $\check{S}_n(\boldsymbol{\beta})$ in (4.5) is a special case of $S_n(\boldsymbol{\beta})$ with $\lambda = 0$.

Let $\hat{\boldsymbol{\beta}}^{(k)}$ denote an optimal estimate in the kth iteration such that $\hat{\boldsymbol{\beta}}^{(k)} = \arg\min_{\boldsymbol{\beta}} S_n^k(\boldsymbol{\beta})$, where $S_n^k(\boldsymbol{\beta})$ is defined by

$$S_n^k(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \rho\left(Y_{(i)} - \boldsymbol{\beta}^\top \boldsymbol{X}_{(i)}^{(k-1)}\right) + \lambda \sum_{j=1}^{p} \omega_j |\beta_j|. \tag{4.13}$$

Here $\left\{\boldsymbol{X}_{(i)}^{(k)}\right\}$ is a permutation of $\{\boldsymbol{X}_i\}$ at the kth iteration such that $(\hat{\boldsymbol{\beta}}^{(k)})^\top \boldsymbol{X}_{(1)}^{(k)} \leq (\hat{\boldsymbol{\beta}}^{(k)})^\top \boldsymbol{X}_{(2)}^{(k)} \leq \cdots \leq (\hat{\boldsymbol{\beta}}^{(k)})^\top \boldsymbol{X}_{(n)}^{(k)}$. The iterative algorithm consists of the following three steps:

Step 1: Set an initial estimate $\hat{\boldsymbol{\beta}}^{(0)}$.

Step 2: At the kth ($k \geq 1$) iteration, given a $\hat{\boldsymbol{\beta}}^{(k-1)}$, obtain $\hat{\boldsymbol{\beta}}^{(k)}$ by minimizing $S_n^k(\boldsymbol{\beta})$.

Step 3: Iterate *Step 2* until $|S_n^k(\hat{\boldsymbol{\beta}}^{(k)}) - S_n^k(\hat{\boldsymbol{\beta}}^{(k-1)})|$ is less than a prespecified small positive number.

The sparse matching quantiles M-estimate $\hat{\boldsymbol{\beta}}_n$ is $\hat{\boldsymbol{\beta}}^{(k)}$. The convergence of the iterative algorithm was proved by Sgouropoulos et al. (2015) when $\rho(\cdot)$ is L_2 discrepancy function. In this chapter, we show that the convergence of the algorithm still holds for a general discrepancy function $\rho(\cdot)$, which is given in Theorem 4.1.

We remark that in the step 1, $\hat{\boldsymbol{\beta}}^{(0)}$ may be taken as an M-estimate that minimizes $n^{-1}\sum_{i=1}^{n} \rho(Y_i - \boldsymbol{\beta}^\top \boldsymbol{X}_i)$, which becomes the OLS estimate if $\rho(\cdot) = (\cdot)^2$, or the minimum L_1-norm estimate if $\rho(\cdot) = |\cdot|$.

4.2.4 Selection of tuning parameters

The constant c in Huber function regulates the amount of robustness as argued in Huber (1981). By Wang et al. (2007), the choice of c in Huber function should reflect the possible proportion of outliers in the data. Huber (1981) recommended the value of $c = 1.345$ for location problems to achieve about 95% efficiency when the data are normally distributed. Some other values of c, say 1.25 or 1.2, can be found in Street et al. (1988), Chi (1994), and Cantoni and Ronchetti (2001) among others. Since the choice of c may impact the estimation efficiency, Wang et al. (2007) and Jiang et al. (2019) proposed data driven approaches to adjust the values of c. Recently, a popular method to select the tuning constant c is through cross-validations (Chen et al., 2017; Fan et al., 2017).

The regularization parameter λ in (4.10) may be chosen by a grid search using cross-validation or information-based criterion, for example, Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC), or the cross validated measurement, such as Mean Absolute Error (MAE), Mean Squared Error (MSE)(Fan et al., 2017).

In this chapter, if ρ is the Huber discrepancy function, we run a two-dimensional grid search using five-fold cross-validation to find the optimal pair (c, λ) that minimizes the mean absolute matching errors (MAME) (see the definition (4.22)) of the validation datasets. Similarly, we use one-dimensional grid search to find the best λ if ρ is L_2 or L_1 discrepancy function. The weights ω_j, $j = 1, \ldots, p$, in (4.10), are usually constructed based on unpenalized estimates (Zou, 2006). For example, let $\hat{\boldsymbol{\beta}}^{[M]}$ be an M-estimate of $\boldsymbol{\beta}$. Then ω_j can be set as $|\hat{\beta}_j^{[M]}|^{-\gamma}$, where $\gamma > 0$ and $\hat{\beta}_j^{[M]}$ is the jth element of the M-estimate $\hat{\boldsymbol{\beta}}^{[M]}$.

4.3 Theoretical properties

In this section, the convergence of the above iterative algorithm, and the statistical properties the matching quantiles M-estimate are presented. Before proceeding, we make the following assumptions.

(A1) $\rho(x)$ is a convex function satisfying that $\rho(x) \geq \rho(0) = 0$, and is Lipschitz continuous, that is, there exists a constant $M \geq 0$, such that for any $x_1, x_2 \in \mathbb{R}$, $|\rho(x_1) - \rho(x_2)| \leq M|x_1 - x_2|$.

(A2) For any $0 < \tau_0 < \tau_1 < 1/2$, there exists Ω_n such that (i) $\inf_{Q_\xi(\alpha) \in \Omega_n} f_\xi(Q_\xi(\alpha)) = n^{-(\tau_1 - \tau_0)}$; (ii) $\sup_{Q_\xi(\alpha) \in \Omega_n} |f'_\xi(Q_\xi(\alpha))| < \infty$; (iii) $P(\xi \in \Omega_n^c) = o(n^{-\tau_0})$, where $\xi = Y$ or $\boldsymbol{\beta}^\top \boldsymbol{X}$ for any fixed $\boldsymbol{\beta}$.

Remark 4.2 *(A1) is commonly made in the theoretical study of M-estimation. (A2) controls the behavior of the tail densities of the random variable ξ, which is weaker than the Conditions B (ii)-(iii) made in Sgouropoulos et al. (2015), where the bounded* *supports of Y and $\boldsymbol{X}$ are needed. If $\xi \sim N(0,1)$, one can take $\Omega_n = [-(2(\tau_1 - \tau_0)\log n - \log 2\pi)^{1/2}, (2(\tau_1 - \tau_0)\log n - \log 2\pi)^{1/2}]$ for any $2\tau_0 < \tau_1 < 1/2$, then (A2) holds true.*

Under the assumption (A2) (i)-(ii), we can show that f_ξ and f'_ξ satisfy the Kiefer conditions on Ω_n, and we have (see Kulik (2007))

$$f_\xi(Q(\alpha))n^{1/2}(Q_\xi(\alpha) - Q_{n,\xi}(\alpha)) - n^{1/2}(F_{n,\xi}(Q_\xi(\alpha)) - \alpha) = R_n(\alpha), \tag{4.14}$$

where $R_n(\alpha) \xrightarrow{a.s.} n^{-1/4}(\log n)^{1/2}(\log\log n)^{1/4}$.

We now present our lemmas and theorems .

Theorem 4.1 *Under the assumption (A1), for a fixed n, the iterative algorithm proposed in Section 4.2.3 converges, i.e., the following holds true:*

$$S_n^k(\hat{\boldsymbol{\beta}}^{(k)}) \to \eta, \text{ as } k \to +\infty, \tag{4.15}$$

where η is a nonnegative constant, and $\hat{\boldsymbol{\beta}}^{(k)} = \arg\min_{\boldsymbol{\beta}} S_n^k(\boldsymbol{\beta})$.

Remark 4.3 *Theorem 4.1 no longer holds if we search for $\boldsymbol{\beta}^\top \boldsymbol{X}$ that matches a part of distribution of Y by employing L_2 discrepancy function (Sgouropoulos et al., 2015). This issue remains if the L_2 discrepancy function is replaced by a general discrepancy function $\rho(\cdot)$.*

The following lemma is needed in the proof of Theorem 4.1.

Lemma 4.1 *Under the assumption (A1), let $\{a_i\}_{i=1}^n$ and $\{b_i\}_{i=1}^n$ be any two sequences. Then we have*

$$\sum_{i=1}^{n} \rho(a_{(i)} - b_{(i)}) \leq \sum_{i=1}^{n} \rho(a_i - b_i), \tag{4.16}$$

where $a_{(i)}$ and $b_{(i)}$ $(i = 1, 2, \ldots, n)$ are the ith ordered values of $\{a_i\}_{i=1}^n$ and $\{b_i\}_{i=1}^n$, respectively.

Theorem 4.2 *Under the assumptions (A1)-(A2), we have*

$$\sup_{\boldsymbol{\beta} \in \mathcal{B}} |S_n(\boldsymbol{\beta}) - S(\boldsymbol{\beta})| \xrightarrow{p} 0, \text{ as } n \to \infty, \tag{4.17}$$

where $\mathcal{B}$ is any compact subset in $\mathbb{R}^p$.

The following lemma is needed in the proof of Theorem 4.2.

Lemma 4.2 *Under the assumptions (A1)-(A2), for any fixed $\boldsymbol{\beta} \in \mathbb{R}^p$ and $\tau_0 < 1/2$,*

$$n^{\tau_0}\{S_n(\boldsymbol{\beta}) - S(\boldsymbol{\beta})\} \xrightarrow{p} 0, \text{ as } n \to \infty. \tag{4.18}$$

Let $\boldsymbol{\beta}^{\mathbf{0}} = \arg\min_{\boldsymbol{\beta}} S(\boldsymbol{\beta})$ for $\lambda \geq 0$. Then we have the following two theorems.

Theorem 4.3 *Under the assumptions (A1)-(A2), we have*

$$S_n(\hat{\boldsymbol{\beta}}_n) \xrightarrow{p} S(\boldsymbol{\beta}^0), \text{ as } n \to \infty. \tag{4.19}$$

Note that $\boldsymbol{\beta}^0$ might not be unique for L_2 discrepancy function (Sgouropoulos et al., 2015), which remains true for a general discrepancy function $\rho(\cdot)$. Namely, $\hat{\boldsymbol{\beta}}_n$ that minimizes (4.10) may not be unique. Nevertheless, by writing $\mathcal{B}_0 = \{\boldsymbol{\beta},\ S(\boldsymbol{\beta}) = S(\boldsymbol{\beta}^0)\}$ and defining $d(\hat{\boldsymbol{\beta}}_n, \mathcal{B}_0) = \min_{\boldsymbol{\beta}\in\mathcal{B}_0} \|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\|$, we have the following theorem.

Theorem 4.4 *Under the assumptions (A1)-(A2), for any $\varepsilon > 0$, we have*

$$P\{d(\hat{\boldsymbol{\beta}}_n, \mathcal{B}_0) \geq \varepsilon\} \to 0, \ \textit{as } n \to \infty. \tag{4.20}$$

We remark that these proofs are in light of Sgouropoulos et al. (2015). However, compared to Sgouropoulos et al. (2015), our work overcomes the challenges in the proofs of theoretical results as we replace the L_2 discrepancy function with a general discrepancy function ρ, and we no longer require $\boldsymbol{X}$ and Y to have bounded supports. In fact, the most notable differences are in the assumptions and the proofs of Lemmas 4.1-4.2 and Theorem 4.2, which are not a straightforward extension of Sgouropoulos et al. (2015).

4.4 Simulations study

The simulations are performed under different scenarios, without or with outliers. The L_2, L_1, and Huber discrepancy functions are chosen for comparison purpose. We remark that the tuning parameters for all methods are chosen by using five-fold cross-validation. For convenience, the MQME method based on Huber ρ_c $(c > 0)$, L_1, and L_2 discrepancy functions are abbreviated as HUBER, LAD, and LS, respectively.

4.4.1 Evaluation measures

The objective of this study is to find a $\hat{\boldsymbol{\beta}}_n$ such that the distribution of $\hat{\boldsymbol{\beta}}_n^\top \boldsymbol{X}$ matches the distribution of Y sufficiently well. Thus, we use three measures to evaluate the performance of the proposed method by using a post-sample of size n, which is generated with each drawn sample as in Sgouropoulos et al. (2015), and denoted by $\{(\tilde{y}_j, \tilde{\boldsymbol{x}}_j), j = 1, \ldots, n\}$. The first measure is the root mean squared matching error (rMSME), and another one is the mean absolute matching error (MAME), which are respectively defined as

$$\text{rMSME} = \left\{ \frac{1}{n} \sum_{i=1}^{n} \left[\tilde{y}_{(i)} - (\hat{\boldsymbol{\beta}}_n^\top \tilde{\boldsymbol{x}})_{(i)} \right]^2 \right\}^{1/2}, \tag{4.21}$$

and

$$\text{MAME} = \frac{1}{n} \sum_{i=1}^{n} \left| \tilde{y}_{(i)} - (\hat{\boldsymbol{\beta}}_n^\top \tilde{\boldsymbol{x}})_{(i)} \right|, \tag{4.22}$$

where $\hat{\boldsymbol{\beta}}_n$ is the (sparse) matching quantiles M-estimate. The third one is the measure for the partial goodness-of-match (Sgouropoulos et al., 2015),

$$\hat{R} = 1 - \frac{1}{2} \sum_{j=1}^{[n/k]} \left| C_j - k/n \right|, \tag{4.23}$$

where

$$C_j = \frac{1}{n} \sum_{i=1}^{n} I\left\{ \frac{(j-1)k}{n} < U_i \leq \frac{jk}{n} \right\}, \quad U_i = \frac{1}{n} \sum_{j=1}^{n} I\left\{ \tilde{y}_j \leq \hat{\boldsymbol{\beta}}_n^\top \tilde{\boldsymbol{x}}_i \right\}.$$

The smaller the difference between the distributions of Y and $\hat{\boldsymbol{\beta}}_n^\top \boldsymbol{X}$, the smaller the values of rMSME and MAME, and the larger the value of $\hat{R}$, where $\hat{R} \in [0,1]$. The ideal case is that $\hat{R} = 1$, which happens if and only if n/k is an integer and each interval $((j-1)k/n, jk/n],\ j = 1, \ldots, n/k$ contains exactly k points from $U_1, \ldots, U_n$. It is easy to see that k should be large enough to guarantee that there are enough sample points in each of $[n/k]$ intervals (see Sgouropoulos et al. (2015) for details).

4.4.2 Example 1

To make the proposed MQME method applicable in the financial market, we generate synthetic data that mimic the styles of assets returns. The simulation design is similar to Sgouropoulos et al. (2015). This example is conducted to illustrate the finite-sample properties of MQME via a linear model,

$$Y_i = \boldsymbol{\beta}^\top \boldsymbol{X}_i + Z_i,\ i = 1, 2, \ldots, n, \tag{4.24}$$

and $\boldsymbol{X}_i$, $1 \le i \le n$, are defined by a multi-factor model

$$\boldsymbol{X}_i = A\boldsymbol{V}_i + \boldsymbol{\epsilon}_i,\ i = 1, \ldots, n,$$

where A is a $p \times 3$ constant factor loading matrix with the elements drawn independently from $U[-1,1]$; the three elements of $\boldsymbol{V}_i$ are independently linear $AR(1)$ processes with Lognormal (0,1) innovations in which the three autoregressive coefficients are drawn independently from $U[-0.95, 0.95]$; $\boldsymbol{\epsilon}_i$ is a $p \times 1$ random error vector

Table 4.1: Mean (standard deviation) of MAME, rMSME and $\hat{R}$ for post-samples from 1000 simulations in Example 1.

(Scenario, p)	MAME			rMSME			$\hat{R}$		
	LS	LAD	HUBER	LS	LAD	HUBER	LS	LAD	HUBER
(I, 10)	0.5614	0.2808	0.2027	0.6309	0.4077	0.2878	0.9229	0.9542	0.9646
	(0.4240)	(0.0955)	(0.0485)	(0.4103)	(0.1367)	(0.0649)	(0.0613)	(0.0230)	(0.0175)
(I, 20)	0.5137	0.3791	0.2279	0.5961	0.5509	0.3193	0.9465	0.9557	0.9700
	(0.3495)	(0.1272)	(0.0511)	(0.3395)	(0.1967)	(0.0698)	(0.0359)	(0.0207)	(0.0150)
(I, 50)	0.5737	0.5564	0.3089	0.6910	0.8116	0.4309	0.9581	0.9562	0.9724
	(0.3491)	(0.1879)	(0.0796)	(0.3536)	(0.2736)	(0.1147)	(0.0237)	(0.0210)	(0.0146)
(II, 10)	1.1096	0.3072	0.2358	1.2028	0.4504	0.3390	0.8663	0.9536	0.9628
	(1.0644)	(0.1136)	(0.0774)	(1.0496)	(0.1658)	(0.1129)	(0.1250)	(0.0229)	(0.0191)
(II, 20)	0.9510	0.4017	0.2572	1.0559	0.5904	0.3632	0.9100	0.9552	0.9684
	(0.8841)	(0.1391)	(0.0679)	(0.8749)	(0.2159)	(0.0936)	(0.0838)	(0.0211)	(0.0160)
(II, 50)	0.7905	0.5778	0.3298	0.9276	0.8426	0.4602	0.9457	0.9561	0.9722
	(0.5593)	(0.2039)	(0.0919)	(0.5585)	(0.2945)	(0.1244)	(0.0384)	(0.0216)	(0.0147)
(III, 10)	4.7797	0.3983	0.3230	5.0725	0.5771	0.4502	0.6460	0.9479	0.9565
	(4.8848)	(0.1732)	(0.1360)	(5.0240)	(0.2581)	(0.1935)	(0.2352)	(0.0242)	(0.0193)
(III, 20)	6.1347	0.5400	0.3931	6.4841	0.7775	0.5352	0.6656	0.9486	0.9611
	(6.5863)	(0.2432)	(0.1814)	(6.7499)	(0.3479)	(0.2570)	(0.0237)	(0.0239)	(0.0176)
(III, 50)	7.7869	0.7948	0.6353	8.4185	1.1393	0.8680	0.7074	0.9486	0.9593
	(8.2518)	(0.3733)	(0.3452)	(8.5867)	(0.5129)	(0.4920)	(0.2105)	(0.0234)	(0.0187)

with mean $\mathbf{0}$ whose components are independently distributed as $t(4)$; the underlying regression coefficient vector $\boldsymbol{\beta}^0$ in (4.24) are drawn independently from uniform distribution, $\beta_j^0 \sim U[-0.5, 0.5]$, $j = 1, \ldots, p$, with $p = 10, 20, 50$. As p increases, it is sensible to consider the sparse structure of the parameter vector $\boldsymbol{\beta}^0$. We thus assume that $\boldsymbol{\beta}^0 = (\beta_1^0, \ldots, \beta_{10}^0, 0, \ldots, 0)^\top$ for $p = 50, 100$, where $\beta_j^0 \sim U[-0.5, 0.5]$, $1 \leq j \leq 10$. We consider the following three scenarios without or with outliers.

- Scenario I: There are no outlier observations. The random errors Z_i, $i = 1, 2, \ldots, n$, are generated from $N(0, 1)$.
- Scenario II: There exist Type A outliers. The random errors Z_i, $i = 1, 2, \ldots, n$, follow a contaminated-normal distribution $(1 - \pi)N(0, 1) + \pi N(0, 64)$ with $\pi < 0.5$.
- Scenario III: There exist Type B outliers. $\{Y_i\}$ are contaminated as follows: randomly take $[\pi n]$ observations from $\{Y_i\}$, and replace their values by m multiples of $\max_{1 \leq i \leq n}\{Y_i\}$ with $m > 1$.

We carry out 1000 simulations for each setting with $n = 300$, $\pi = 0.025$, and $m = 1.5$. We first perform MQME when the underlying true regression vector $\boldsymbol{\beta}^0$ is not sparse for $p = 10$, 20, 50. We compute the MAME, rMSME, and $\hat{R}$ with $k/n = 0.025$ using the post-samples in Table 4.1. For all the settings, HUBER

Table 4.2: Mean (standard deviation) of MAME, rMSME and $\hat{R}$ from 1000 simulations in Example 1 with $\boldsymbol{\beta}^0$ being sparse.

(Scenario, p)	MAME			rMSME			$\hat{R}$		
	LS	LAD	HUBER	LS	LAD	HUBER	LS	LAD	HUBER
(I, 50)	0.4351	0.2937	0.2042	0.5109	0.4310	0.2908	0.9390	0.9538	0.9653
	(0.2839)	(0.0970)	(0.0479)	(0.2721)	(0.1561)	(0.0663)	(0.0373)	(0.0234)	(0.0173)
(I, 100)	0.7673	0.2955	0.2157	0.8347	0.4349	0.3047	0.8982	0.9534	0.9638
	(0.6259)	(0.1009)	(0.1173)	(0.6079)	(0.1569)	(0.1523)	(0.0893)	(0.0234)	(0.0194)
(II, 50)	1.3780	0.3207	0.2388	1.4673	0.4753	0.3429	0.8390	0.9528	0.9635
	(1.2293)	(0.1148)	(0.0716)	(1.2181)	(0.1851)	(0.1010)	(0.1350)	(0.0233)	(0.0179)
(II, 100)	1.2961	0.3209	0.2486	1.3837	0.4764	0.3550	0.8479	0.9523	0.9612
	(1.1361)	(0.1171)	(0.1069)	(1.1253)	(0.1791)	(0.1444)	(0.1323)	(0.0240)	(0.0200)
(II, 50)	4.8764	0.4170	0.3600	5.1257	0.5987	0.5006	0.6292	0.9461	0.9537
	(4.5731)	(0.1841)	(0.1631)	(4.6942)	(0.2708)	(0.2370)	(0.2233)	(0.0246)	(0.0195)
(III, 100)	4.0397	0.4123	0.3810	4.3105	0.5960	0.5333	0.6842	0.9465	0.9514
	(3.8521)	(0.1799)	(0.3203)	(3.9498)	(0.2581)	(0.4248)	(0.2023)	(0.0246)	(0.0263)

outperforms the other two methods, i.e., LAD and LS, and LAD performs better than LS in general. Another interesting finding is that even though the target random variable Y is not contaminated, both LAD and HUBER outperforms LS, which can be due to the large volatility in the data that simulate assets returns. We also report the MAME, rMSME, and $\hat{R}$ in Table 4.2 when $\boldsymbol{\beta}^0$ is sparse for $p = 50, 100$. We draw the same conclusion as above, that is, HUBER outperforms both LAD and LS in all the settings, and LAD performs better than LS.

4.4.3 Example 2

Our purpose is to match the distribution of a target random variable Y by a linear combination of p random variables $\boldsymbol{X}$, $\mathcal{L}(Y) = \mathcal{L}(\boldsymbol{\beta}^\top \boldsymbol{X}), \boldsymbol{\beta} \in \mathbb{R}^p$. As the linear regression relationship between Y and $\boldsymbol{X}$ is too obvious in Example 1, we thus weaken the simulation setting by generating $\boldsymbol{X}_i, Y_i, i = 1, \ldots, n$ as follows: $\boldsymbol{X}_i \sim N_p(\mathbf{0}, \Sigma)$, $\Sigma = (\sigma_{jj'})$ with $\sigma_{jj'} = 0.5^{|j-j'|}$, $j, j' = 1, \ldots, p$; $Y_i \sim \mathcal{L}(\boldsymbol{\beta}^\top \boldsymbol{X}) = N(0, \boldsymbol{\beta}^\top \Sigma \boldsymbol{\beta})$. Since such generated data set does not contain an outlier, this scenario is still named as Scenario I for simple presentation. Similar to the previous scenario II, we contaminate Y_i by generating from the distribution $(1-\pi)N(0, \boldsymbol{\beta}^\top \Sigma \boldsymbol{\beta}) + \pi N(0, 64\boldsymbol{\beta}^\top \Sigma \boldsymbol{\beta})$ with $\pi < 0.5$ to create Type A outliers in the data. For simple presentation, this scenario remains named as Scenario II. The scenario III for Example 2, and the settings of $\boldsymbol{\beta}^0$, π, m, n, p are the same as those in Example 1.

The MAME, rMSME, and $\hat{R}$ with $k/n = 0.25$ are calculated based on the post-samples for HUBER, LAD, and LS. We report their values in Tables 4.3-4.4. By these tables, it can be observed that when there are no outliers, the three methods have similar performance since their differences in matching errors and $\hat{R}$ values are

Table 4.3: Mean (standard deviation) of MAME, rMSME and $\hat{R}$ from 1000 simulations in Example 2.

(Scenario, p)	MAME			rMSME			$\hat{R}$		
	LS	LAD	HUBER	LS	LAD	HUBER	LS	LAD	HUBER
(I, 10)	0.0882	0.0884	0.0880	0.1160	0.1165	0.1161	0.9406	0.9409	0.9402
	(0.0311)	(0.0321)	(0.0314)	(0.0344)	(0.0355)	(0.0349)	(0.0254)	(0.0260)	(0.0253)
(I, 20)	0.1404	0.1363	0.1380	0.1845	0.1810	0.1830	0.9394	0.9414	0.9408
	(0.0511)	(0.0476)	(0.0497)	(0.0566)	(0.0531)	(0.0548)	(0.0252)	(0.0258)	(0.0268)
(I, 50)	0.3089	0.2721	0.2891	0.4011	0.3555	0.3769	0.9323	0.9383	0.9372
	(0.1209)	(0.1026)	(0.1127)	(0.1384)	(0.1163)	(0.1295)	(0.0297)	(0.0258)	(0.0285)
(II, 10)	0.2733	0.0990	0.0980	0.3500	0.1309	0.1288	0.8813	0.9377	0.9381
	(0.1294)	(0.0395)	(0.0433)	(0.1591)	(0.0463)	(0.0532)	(0.0483)	(0.0269)	(0.0272)
(II, 20)	0.4706	0.1591	0.1493	0.5993	0.2104	0.1967	0.8695	0.9366	0.9383
	(0.2192)	(0.0608)	(0.0575)	(0.2711)	(0.0714)	(0.0668)	(0.0495)	(0.0269)	(0.0276)
(II, 50)	1.1554	0.3275	0.3260	1.4644	0.4301	0.4249	0.8467	0.9312	0.9323
	(0.4714)	(0.1314)	(0.1753)	(0.5834)	(0.1566)	(0.2140)	(0.0482)	(0.0311)	(0.0314)
(III, 10)	0.1533	0.0997	0.0926	0.1997	0.1315	0.1219	0.9193	0.9372	0.9388
	(0.0556)	(0.0402)	(0.0349)	(0.0659)	(0.0468)	(0.0399)	(0.0336)	(0.0278)	(0.0270)
(III, 20)	0.2608	0.1580	0.1439	0.3379	0.2093	0.1896	0.9106	0.9368	0.9393
	(0.0917)	(0.0600)	(0.0557)	(0.1088)	(0.0706)	(0.0649)	(0.0353)	(0.0273)	(0.0274)
(III, 50)	0.6234	0.3296	0.3169	0.8002	0.4318	0.4140	0.8940	0.9314	0.9328
	(0.1924)	(0.1308)	(0.1360)	(0.2302)	(0.1535)	(0.1644)	(0.0377)	(0.0300)	(0.0298)

Table 4.4: Mean (standard deviation) of MAME, rMSME and $\hat{R}$ from 1000 simulations in Example 2 with $\boldsymbol{\beta}^0$ being sparse.

(Scenario, p)	MAME			rMSME			$\hat{R}$		
	LS	LAD	HUBER	LS	LAD	HUBER	LS	LAD	HUBER
(I, 50)	0.0937	0.0933	0.0946	0.1220	0.1217	0.1235	0.9371	0.9381	0.9384
	(0.0352)	(0.0345)	(0.0360)	(0.0387)	(0.0381)	(0.0404)	(0.0282)	(0.0267)	(0.0278)
(I, 100)	0.0923	0.0910	0.1021	0.1207	0.1193	0.1339	0.9382	0.9397	0.9306
	(0.0339)	(0.0416)	(0.0380)	(0.0379)	(0.0484)	(0.0445)	(0.0280)	(0.0399)	(0.0333)
(II, 50)	0.2506	0.1049	0.1001	0.3210	0.1371	0.1311	0.8870	0.9355	0.9324
	(0.1326)	(0.0427)	(0.0388)	(0.1642)	(0.0498)	(0.0448)	(0.0486)	(0.0293)	(0.0318)
(II, 100)	0.2808	0.1058	0.1026	0.3583	0.1389	0.1348	0.8795	0.9337	0.9342
	(0.1593)	(0.0536)	(0.0417)	(0.1976)	(0.0645)	(0.0490)	(0.0523)	(0.0516)	(0.0302)
(II, 50)	0.1353	0.1045	0.0998	0.1768	0.1367	0.1300	0.9262	0.9357	0.9318
	(0.0540)	(0.0420)	(0.0385)	(0.0646)	(0.0487)	(0.0441)	(0.0335)	(0.0278)	(0.0317)
(III, 100)	0.1405	0.1050	0.1025	0.1839	0.1377	0.1348	0.9237	0.9361	0.9333
	(0.0564)	(0.0441)	(0.0424)	(0.0687)	(0.0526)	(0.0500)	(0.0324)	(0.0361)	(0.0303)

small. But for the scenarios II and III, our methods LAD and HUBER outperform LS overwhelmingly in terms of the MAMEs, rMSMEs, and $\hat{R}$s. In addition, Huber generally outperforms LAD in terms of their MAMEs and rMSMEs though their differences are small. Moreover, it is hard to conclude that one method dominates others via their $\hat{R}$ values.

4.5 A real case study

In this section, a real example is considered for investigating the performance of (sparse) MQME. Our purpose is to assemble a representative portfolio with different securities that matches various characteristics of a benchmark index. The function $\rho(\cdot)$ is chosen to be L_2, L_1 or Huber discrepancy function. The tuning parameter c in Huber function is selected via five-fold cross-validation.

We apply the proposed method to the stock market index in Hong Kong during the period of 2013-2016. These data are from *Yahoo!Finance* (https://finance.yahoo.com/). Hang Seng Index (HSI) is the main indicator of the overall market performance in Hong Kong. It records and monitors daily changes of the largest companies of the Hong Kong stock market. Let the target random variable Y be the net returns of HSI, and $\mathcal{X} = \{X_1, X_2, \dots, X_p\}$ be the whole set of returns of 41 actively traded stocks included in HSI ($p = 41$) from 2013 to 2016. The returns are calculated using the adjusted daily closing prices. By eliminating the missing data, the data set contains 974 recorded net returns of HSI between 2013 and 2016, which are displayed in Figure 4.1. This figure visually shows that the overall market behavior during the period is not homogenized. After applying the R package `changepoint` to the data, we divide the financial market into Market I (overall positive average return) of size 552 and Market II (overall negative average return) of size 422. The Market II has a higher volatility than the Market I. Given the different characteristics of the two markets, the representative portfolio may be different. For Market I, we further split the data set into the in-sample of size 252, and the out-sample of size 300. We also divide the data set in Market II into the in-sample of size 222, and the post-sample of size 200. We apply the in-samples to estimate portfolios, and post-samples to compare their performances with the returns of HSI from each market.

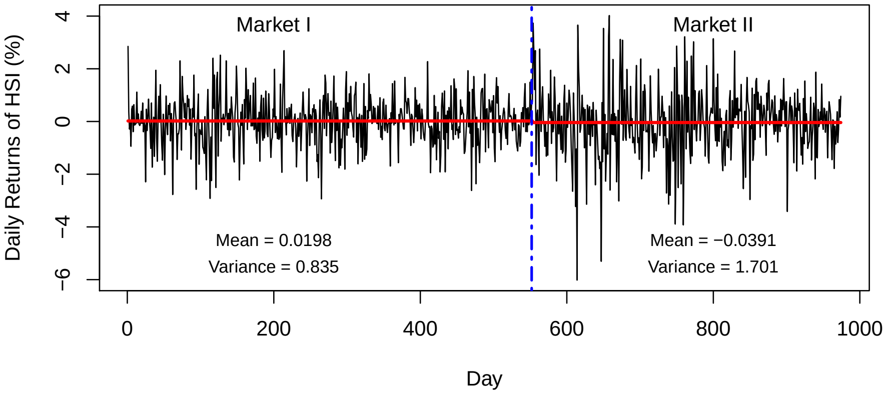

Figure 4.1: Daily net returns of HSI. The data set is divided into two parts corresponding to Markets I and II by the different overall means and variances.

In order to examine the outliers resistance of MQME, we contaminate HSI by Type B outliers. In each market, randomly select one point from the in-sample, and reset it to be m multiples of the maximum of the sample. We perform (sparse) MQME on the contaminated data for both markets. In some cases, an investor has prior information and he/she is interested in forming a portfolio based on a specific subset $\mathcal{X}_1$ of $\mathcal{X}$, where $|\mathcal{X}_1| < |\mathcal{X}|$ with $|\cdot|$ denoting the size of a set. In other cases, however, an investor may not have any prior information, and he/she regards every stock in $\mathcal{X}$ evenly. Herein, two modeling schemes are carried out to estimate the representative portfolios.

- Scheme 1: Firstly, randomly select a subset $\mathcal{X}_1$ of $\mathcal{X}$. Let $|\mathcal{X}_1| = 3,\ 5,\ 8$. Then form portfolios based on $\mathcal{X}_1$ by computing MQME for both markets.
- Scheme 2: Form a portfolio based on $\mathcal{X}$ by computing sparse MQME. Thereby a subset of $\mathcal{X}$ is selected.

Table 4.5: Matching errors (MAME and rMSME) under the scheme 1 for Markets I and II.

Method	No.S	Market I		Market II	
		MAME	rMSME	MAME	rMSME
LS	3	0.1144	0.2050	0.2237	0.3094
LAD	3	0.0725	0.1132	0.1213	0.1882
HUBER	3	0.0664	0.0992	0.1353	0.2044
LS	5	0.1609	0.3404	0.3563	0.4658
LAD	5	0.0535	0.1171	0.0649	0.0961
HUBER	5	0.0507	0.1198	0.0695	0.1004
LS	8	0.2084	0.3763	0.3708	0.4964
LAD	8	0.0700	0.1267	0.0611	0.0926
HUBER	8	0.0596	0.1233	0.0647	0.0947

We compare the distribution matching for three different discrepancy functions via MAME and rMSME, which are given in Table 4.5 (Scheme 1) and Table 4.6

Table 4.6: Matching errors (MAME and rMSME) under the scheme 2 for both markets.

Method	Market I			Market II		
	No.S	MAME	rMSME	No.S	MAME	rMSME
LS	5	0.1246	0.3292	5	0.4663	0.6619
LAD	5	0.0642	0.0826	5	0.0784	0.1078
HUBER	5	0.0754	0.0917	5	0.0663	0.0964
LS	10	0.1250	0.3046	9	0.3760	0.5590
LAD	10	0.0436	0.0610	9	0.0751	0.1042
HUBER	10	0.0468	0.0628	9	0.0604	0.0813
LS	13	0.1449	0.3647	11	0.3247	0.4785
LAD	13	0.0368	0.0589	11	0.0610	0.0819
HUBER	13	0.0402	0.0551	11	0.0518	0.0686

(Scheme 2), where 'No.S' stands for the number of stocks. We also plot the $\hat{R}$ defined in (4.23) with $k/n = 0.05, 0.01, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5$ in Figure 4.2 (Scheme 1) and Figure 4.3 (Scheme 2). Inspecting Tables 4.5-4.6 and Figures 4.2-4.3 reveals both LAD and HUBER outperform LS in terms of the distribution matching, showing smaller MAME and rMSME than LS overall, and larger values of $\hat{R}$ than LS in most cases. Although HUBER has smaller matching errors than those by LAD in general, the performance difference between LAD and HUBER is relatively small, as the values of MAME and rMSME do not differ significantly. This phenomenon agrees with the $\hat{R}$ values in Figures 4.2-4.3.

In addition, we provide the summary statistics of the target portfolio (HSI) and the representative portfolios, which include the number of stocks (No.S) assembled in

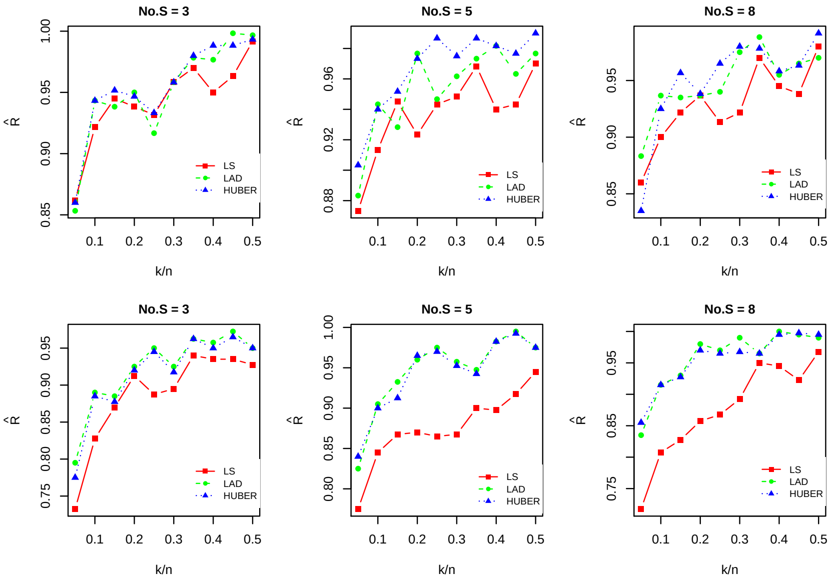

Figure 4.2: Matching goodness ($\hat{R}$) under the scheme 1 for Market I (upper three panels) and Market II (lower three panels).

Table 4.7: Summary statistics under the scheme 1 for Markets I and II.

Method	No.S	Market I						Market II					
		Mean	Min	Max	STD	NM	SS	Mean	Min	Max	STD	NM	SS
HIS	50	0.032	-2.936	2.273	0.860	-0.675	−−	0.038	-3.415	3.141	0.978	-0.768	−−
LS	3	0.049	-3.202	3.561	1.019	-0.734	-0.453	0.019	-3.622	5.732	1.238	-0.909	0.000
LAD	3	0.020	-2.898	2.785	0.927	-0.757	0.000	0.013	-2.912	5.019	1.075	-0.775	0.000
HUBER	3	0.019	-2.842	2.692	0.909	-0.724	0.000	0.013	-2.994	5.140	1.102	-0.794	0.000
LS	5	0.067	-3.534	5.184	1.113	-0.727	-0.248	0.004	-4.341	4.192	1.403	-0.981	-0.684
LAD	5	0.042	-2.666	3.453	0.903	-0.646	0.000	0.035	-2.830	3.572	0.962	-0.748	0.000
HUBER	5	0.049	-2.904	3.405	0.904	-0.637	0.000	0.034	-2.945	3.763	0.979	-0.754	0.000
LS	8	0.037	-3.990	4.870	1.180	-0.789	-0.678	0.048	-4.949	5.174	1.460	-1.189	-0.358
LAD	8	0.046	-3.276	2.945	0.934	-0.659	-0.009	0.037	-3.073	3.521	0.983	-0.772	0.000
HUBER	8	0.045	-3.251	3.092	0.929	-0.671	0.000	0.036	-3.089	3.610	0.991	-0.764	0.000

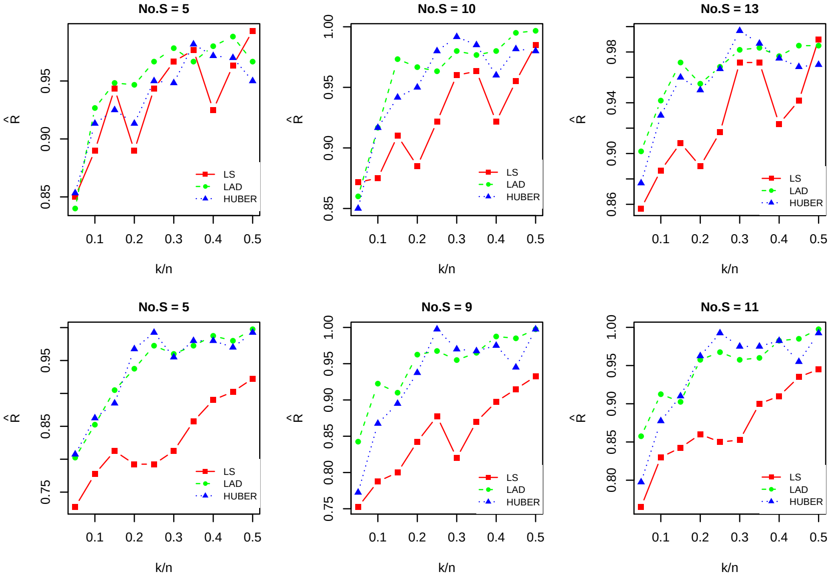

Figure 4.3: Matching goodness ($\hat{R}$) under the scheme 2 for Market I (upper three panels) and Market II (lower three panels).

Table 4.8: Summary statistics under the scheme 2 for Markets I and II.

Method	Market I							Market II						
	No.S	Mean	Min	Max	Sd	NM	SS	No.S	Mean	Min	Max	Sd	NM	SS
HIS	50	0.032	-2.936	2.273	0.860	-0.675	−−	50	0.038	-3.415	3.141	0.978	-0.768	−−
LS	5	0.086	-3.255	5.836	1.075	-0.701	0.000	5	-0.036	-4.149	7.293	1.550	-1.123	-2.082
LAD	5	0.021	-2.880	2.434	0.808	-0.620	0.000	5	0.074	-3.675	2.706	0.999	-0.799	0.000
HUBER	5	0.002	-2.811	2.173	0.797	-0.599	0.000	5	0.070	-3.613	3.213	1.030	-0.835	0.000
LS	10	0.093	-2.984	5.559	1.048	-0.686	-0.112	9	-0.021	-3.627	6.762	1.437	-1.053	-2.149
LAD	10	0.017	-2.898	2.402	0.855	-0.688	0.000	9	0.060	-3.949	3.483	1.058	-0.833	0.000
HUBER	10	0.020	-2.846	2.316	0.830	-0.621	0.000	9	0.064	-3.589	3.135	1.015	-0.826	0.000
LS	13	0.104	-3.295	6.435	1.107	-0.687	-0.272	11	-0.002	-3.135	6.232	1.366	-1.000	-2.114
LAD	13	0.020	-2.953	2.524	0.875	-0.673	0.000	11	0.051	-3.775	3.459	1.026	-0.815	0.000
HUBER	13	0.021	-2.877	2.353	0.839	-0.628	0.000	11	0.061	-3.510	3.126	0.998	-0.798	0.000

each portfolio, mean, maximum (Max), minimum (Min), standard deviations (STD), the negative mean (NM) of the daily returns, and the percentages for short sales (SS). Herein NM is defined as the mean of all the negative returns (Sgouropoulos et al., 2015). As seen from Tables 4.7-4.8, we can easily tell that the portfolios of either LAD or HUBER are superior to the LS by presenting similar statistics, say, Max, Min, STD and NM, to the target portfolio HSI. The larger absolute values of STD and NM indicate that the LS portfolios are more risky. The risk of a portfolio is also reflected by the large short sales, the goal of which is to hedge the risk. LS yields larger average daily returns than LAD and HUBER in most cases, together with the larger STD, NM, and SS, which indicates the LS portfolios are more risky. While in the Market II under scheme 2 (see Table 4.8), the LS portfolio shows negative average daily returns with a higher risk though the average return of HSI is positive. For both markets, the LAD or HUBER portfolios yield positive average returns with a lower risk, and they have similar characters in majority cases, which agrees with the fact that both LAD and HUBER work well when there are large outliers.

5 Conclusions and future work

5.1 Conclusions

In this dissertation, we investigate the sign-constrained high-dimensional linear regression model, high-dimensional multivariate regression M-estimation and matching quantiles M-estimation.

Firstly, we propose a method for high-dimensional regression problems where the regression coefficients are non-negative, sparse, or even carrying a structure with homogeneous subgroups. We aim to identify the underlying optimal grouping and obtain the optimal estimate that satisfies the sign constraints. Specifically, we formulate a regularized minimization problem with a non-convex, but the difference of convex, objective function. By using the difference of convex programming, a subproblem at each iteration is reformulated as a constrained minimization problem with a convex objective, which is solved by applying the augmented Lagrange and the coordinate descent methods. The theoretical results show that the nnFSG estimate can consistently identify the underlying true grouping and features associated with non-zero coefficients. In addition, the numerical studies show that our method achieves high accuracy in model prediction, feature grouping and/or selection.

Second, we study the properties of ridge-regularized and un-regularized M-estimate under the matrix framework as p/n tends to a finite number. We first establish a nonlinear system of two deterministic equations that characterizes the behaviours of M-estimate. We also provide some examples that demonstrate the remarkable accuracy of our proposed system in measuring the bounds of $\hat{\boldsymbol{\beta}}$.

Finally, we propose an MQME method that is resistant to outliers. The MQME combined with the adaptive Lasso penalty encourages sparsity in the estimate. Since the MQME does not admit an explicit solution, an iterative algorithm is thus developed to solve it. The theoretical properties of the MQME estimate are investigated based on the assumptions that are weaker than those of MQE made in Sgouropoulos et al. (2015). Experimental results on both simulated and real data demonstrate the effectiveness of the MQME.

5.2 Future work

In the era of data explosion, data sets with outliers or heavy tails are ubiquitous in high-dimensional statistical modeling. This means that innovative methods should be explored to explain and analyze those complex data, such as high-dimensional regression M-estimation with general constraints. It may lead to another problem, non-convex M-estimation which has been rarely studied. My short-term plan is to study the statistical properties of M-estimates and develop an algorithm to solve the corresponding non-convex optimization problem.

In terms of multivariate high-dimensional M-estimation, we use a nonlinear system of two deterministic equations to characterize the properties of the M-estimates. Since the proximal mapping function $\text{prox}_t(\rho)(\cdot)$ is involved in our system, it is of great challenge to solve it for many choices of the discrepancy function ρ. Regarding the application of the proposed nonlinear system, an additional task to perform is to develop a numerical algorithm to solve the equations. A potential research topic is to consider 'elliptical-like' explanatory variables inspired by El Karoui (2018).

Bibliography

Arnold, T. B., and Tibshirani, R.J. (2016). Efficient implementations of the generalized lasso dual path algorithm. *Journal of Computational and Graphical Statistics*, 25(1), 1–27.

Bai, Z.D. (1999). Methodologies in spectral analysis of large-dimensional random matrices, a review, *Statistica Sinica*, 611–677.

Bai, Z.D., Rao, C.R., and Wu,Y. (1992). M-estimation of multivariate linear regression parameters under a convex discrepancy function.*Statistica Sinica*, 2, 237–254.

Bai, Z.D., and Wu, Y. (1994). Limiting behavior of M-estimators of regression coefficients in high dimensional linear models I. scale-dependent case. *Journal of Multivariate Analysis*, 51, 211–239.

Bean, D., Bickel, P. J., El Karoui, N., and Yu, B. (2013). Optimal M-estimation in high-dimensional regression. *Proceedings of the National Academy of Sciences*, 110(36), 14563–14568.

Beck, A., and Teboulle, M. (2010). Gradient-based algorithms with application in signal recovery problems, in Convex Optimization in Signal Processing and Communications, D. Palomar and Y. Eldar, eds, Cambridge University Press.

Bickel, P.J. (1975). One-Step Huber Estimates in the Linear Model. *Journal of the American Statistical Association*, 70, 428–434.

Cantoni, E., and Ronchetti, E. (2001). Robust inference for generalized linear models. *Journal of the American Statistical Association*, 96(455), 1022–1030.

Chen, K., Lv, Q., Lu, Y., and Dou, Y. (2017). Robust regularized extreme learning machine for regression using iteratively reweighted least squares. *Neurocomputing* 230, 345–358.

Chi, E. M. (1994). M-estimation in cross-over trials. *Biometrics*, 486–493.

Dawid, A.P. (1981). Some matrix-variate distribution theory: notational considerations and a Bayesian application. *Biometrika*, 68, 265–274.

Dominicy, Y., and Veredas, D. (2013). The method of simulated quantiles. *Journal of Econometrics*, 172(2), 235–247.

El Karoui, N. (2009). Concentration of measure and spectra of random matrices: Applications to correlation matrices, elliptical distributions and beyond. *Annals of Applied Probability*, 19, 2362–2405.

El Karoui, N. (2013). Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators, rigorous results, arXiv preprint arXiv:1311.2445.

El Karoui, N. (2018). On the impact of predictor geometry on the performance on high-dimensional ridge regularized generalized robust regression estimators. *Probability Theory and Related Fields*, 170, 95–175.

El Karoui, N., Bean, D., Bickel, P. J., Lim, C., and Yu, B. (2013). On robust regression with high-dimensional predictors. *Proceedings of the National Academy of Sciences*, 110(36), 14557–14562.

Esser, E., Lou, Y.F., and Xin, J. (2013). A method for finding structured sparse solutions to nonnegative least squares problems with applications. *SIAM Journal on Imaging Sciences*, 6.4, 2010–2046.

Fan, J., and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456), 1348–1360.

Fan, J., Li, Q., and Wang, Y. (2017). 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.

Frank, L.E., and Friedman J.H.(1993). A statistical view of some chemometrics regression tools. *Technometrics*, 35(2), 109–135.

Friedman, J., Hastie, T., Simon, N., and Tibshirani, R. (2016). Lasso and Elastic-Net Regularized Generalized Linear Models. R-package version 2.0-5. 2016.

Fu, A., Narasimhan, B., and Boyd, S. (2017). CVXR: An R package for disciplined convex optimization. arXiv preprint arXiv:1711.07582.

Goeman, J.J. (2010). L_1 penalized estimation in the Cox proportional hazards model. *Biometrical Journal*, 52(1), 70–84.

He, X.M., and Shao, Q.M. (1996). A general Bahadur representation of M-estimators and its application to linear regression with nonstochastic designs. *The Annals of Statistics*, 24, 2608–2630.

He, X.M., and Shao, Q.M. (2000). On parameters of increasing dimensions. *Journal of Multivariate Analysis*, 73, 120–135.

Hu, Z., Follmann, D. A., and Miura, K. (2015). Vaccine design via nonnegative lasso-based variable selection. *Statistics in Medicine*, 34(10), 1791–1798.

Huang, J., Ma, S., Xie, H., and Zhang, C. H. (2009). A group bridge approach for variable selection. *Biometrika*, 96(2), 339–355.

Huber, P.J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35, 73–101.

Huber, P.J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. *The Annals of Statistics*, 1, 799–821.

Huber, P.J. (1981). Robust Statistics 1981.New York: John Wiley.

Itoh, Y., Duarte, M. F., and Parente, M. (2016). Perfect recovery conditions for non-negative sparse modeling. *IEEE Transactions on Signal Processing*, 65(1), 69–80.

Jang, W., Lim, J., Lazar, N., Loh, J. M., McDowell, J., and Yu, D. (2011). Regression shrinkage and equality selection for highly correlated predictors with HORSES. *Biometrics*, 64, 1–23.

Jiang, Y., Wang, Y. G., Fu, L., and Wang, X. (2019). Robust Estimation Using Modified Huber's Functions With New Tails. *Technometrics*, 61(1), 111–122.

Kiefer, J. (1970). Deviations between the sample quantile process and the sample df. *Nonparametric Techniques in Statistical Inference*, 299–319.

Koenker, R., and Portnoy, S. (1990). M-estimation of multivariate regressions. *Journal of the American Statistical Association*, 85, 1060–1068.

Koike, Y., and Tanoue, Y. (2019). Oracle inequalities for sign constrained generalized linear models. *Econometrics and Statistics*, 11, 145–157.

Kulik, R. (2007). Bahadur-Kiefer theory for sample quantiles of weakly dependent linear processes. *Bernoulli*, 13, 1071–1090.

Küng, R., and Jung, P. (2016). Robust nonnegative sparse recovery and 0/1-bernoulli measurements. *2016 IEEE Information Theory Workshop (ITW)*, 260–264. IEEE.

Lambert-Lacroix, S., and Zwald, L. (2011). Robust regression through the Huber's criterion and adaptive lasso penalty. *Electronic Journal of Statistics*, 5, 1015–1053.

Ledoux, M. (2001). The concentration of measure phenomenon, volume 89 of Mathematical Surveys and Monographs, Providence, RI: American Mathematical Society.

Lei, L., Bickel, P.J., and El Karoui, N. (2018). Asymptotics for high dimensional regression M-estimates: fixed design results. *Probability Theory and Related Fields*, 172.3-4, 983–1079.

Li, G., Peng, H., and Zhu, L. (2011). Nonconcave penalized M-estimation with a diverging number of parameters. *Statistica Sinica*, 21, 391–419.

Li, H., Sheffield, J., and Wood, E. F. (2010). Bias correction of monthly precipitation and temperature fields from Intergovernmental Panel on Climate Change AR4 models using equidistant quantile matching. *Journal of Geophysical Research: Atmospheres*, 115(D10).

Loh, P.L. (2017). Statistical consistency and asymptotic normality for high-dimensional robust M-estimators. *The Annals of Statistics*, 45, 866–896.

Loh, P.L., and Wainwright, M.J. (2015). Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *Journal of Machine Learning Research*, 16, 559–616.

Luenberger, D.G., and Ye, Y. (2015). Linear and nonlinear programming (Vol. 228). Springer.

Mammen, E. (1989). Asymptotics with increasing dimension for robust regression with applications to the bootstrap. *The Annals of Statistics*, 17, 382–400.

Mandal, B. N., and Ma, J. (2016). l_1 regularized multiplicative iterative path algorithm for non-negative generalized linear models. *Computational Statistics & Data Analysis*, 101, 289–299.

Meinshausen, N. (2013). Sign-constrained least squares estimation for high-dimensional regression. *Electronic Journal of Statistics*, 7, 1607–1631.

Moreau, J.J. (1965). Proximité et dualité dans un espace hilbertien. *Bulletin de la Societe Mathematique de France*, 93, 273–299.

Mullen, K. M., and van Stokkum, I. H. (2012). The LawsonšCHanson algorithm for nonnegative least squares (NNLS). *Technical report*, CRAN. http://cran. r-proje ct. org/web/packa ges/nnls/nnls. pdf.

Negahban, S.N., Ravikumar, P., Wainwright, M.J., and Yu, B. (2012). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. Science, 27, 538–557.

Ollila, E., Soloveychik, I., Tyler, D. E., and Wiesel, A. (2016). Simultaneous penalized M-estimation of covariance matrices using geodesically convex optimization. arXiv preprint arXiv:1608.08126.

Portnoy, S. (1984). Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large. I. Consistency. *The Annals of Statistics*, 12, 1298–1309.

Portnoy, S. (1985). Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large. II. Normal approximation. *The Annals of Statistics*, 13, 1403–1417.

Rekabdarkolaee, H. M., Boone, E., and Wang, Q. (2017). Robust estimation and variable selection in sufficient dimension reduction. *Computational Statistics Data Analysis*, 108, 146–157.

Renard, B. Y., Kirchner, M., Steen, H., Steen, J. A., and HAMPRECHT, F. A. (2008). NITPICK: peak identification for mass spectrometry data. *BMC bioinformatics*, 9(1), 355.

Sgouropoulos, N., Yao, Q., and Yastremiz, C. (2015). Matching a distribution by matching quantiles estimation. *Journal of the American Statistical Association* 110(510), 742–759.

Shadmi, Y., Jung, P., and Caire, G. (2019). Sparse Non-Negative Recovery from Biased Subgaussian Measurements using NNLS. arXiv preprint arXiv:1901.05727.

She, Y. (2010). Sparse regression with exact clustering. *Electronic Journal of Statistics*, 4, 1055–1096.

Shen, X., Huang, H. C., and Pan, W. (2012). Simultaneous supervised clustering and feature selection over a graph. *Biometrika*, 99(4), 899–914.

Shen, X., Pan, W., and Zhu, Y. (2012). Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association*, 107(497), 223–232.

Shen, X., Pan, W., Zhu, Y., and Zhou, H. (2013). On constrained and regularized high-dimensional regression. *Annals of the Institute of Statistical Mathematics*, 65(5), 807–832.

Slawski, M., Hein, M., and Campus, E. (2010). Sparse recovery for protein mass spectrometry data. *In Practical Applications of Sparse Modeling*, 79–98. MIT Press.

Slawski, M., and Hein, M. (2013). Non-negative least squares for high-dimensional linear models: Consistency and sparse recovery without regularization. *Electronic Journal of Statistics*, 7, 3004–3056.

Slawski, M., Hussong, R., Tholey, A., Jakoby, T., Gregorius, B., Hildebrandt, A., and Hein, M. (2012). Isotope pattern deconvolution for peptide mass spectrometry by non-negative least squares/least absolute deviation template matching. *BMC bioinformatics*,13(1), 291.

Srivastav, R. K., Schardong, A., and Simonovic, S. P. (2014). Equidistance quantile matching method for updating IDFCurves under climate change. *Water resources management* 28(9), 2539–2562.

Street, J. O., Carroll, R. J., and Ruppert, D. (1988). A note on computing robust regression estimates via iteratively reweighted least squares. *The American Statistician*, 42(2), 152–154.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267–288.

Tibshirani, R., and Taylor, J. (2011). The solution path of the generalized lasso. *The Annals of Statistics*, 39(3), 1335–1371.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1), 91–108.

Tibshirani, R., and Wang, P. (2007). Spatial smoothing and hot spot detection for CGH data using the fused lasso. *Biostatistics*, 9(1), 18–29.

Tseng, P. (2001). Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of optimization theory and applications*, 109(3), 475–494.

Wang, Y. G., Lin, X., Zhu, M., and Bai, Z. (2007). Robust estimation using the Huber function with a data-dependent tuning constant. *Journal of Computational and Graphical Statistics*, 16(2), 468–481.

Welsh, A.H. (1989). On M-processes and M-estimation. *The Annals of Statistics*, 17, 337–361.

Wen, Y. W., Wang, M., Cao, Z., Cheng, X., Ching, W. K., and Vassiliadis, V. S. (2015). Sparse solution of nonnegative least squares problems with applications in the construction of probabilistic Boolean networks. *Numerical Linear Algebra with Applications*, 22(5), 883–899.

Wright, S., and Nocedal, J. (1999). Numerical optimization. *Springer Science*, 35(67–68), 7.

Wu, L., and Yang, Y. (2014). Nonnegative elastic net and application in index tracking. *Applied Mathematics and Computation*, 227, 541–552.

Wu, L., Yang, Y., and Liu, H. (2014). Nonnegative-lasso and application in index tracking. *Computational Statistics & Data Analysis*, 70, 116–126.

Xiang, S., Shen, X., and Ye, J. (2015). Efficient nonconvex sparse group feature selection via continuous and discrete optimization. *Artificial Intelligence*, 224, 28–50.

Yang, S., Yuan, L., Lai, Y. C., Shen, X., Wonka, P., and Ye, J. (2012). Feature grouping and selection over an undirected graph. *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 922–930. ACM.

Yang, Y., and Wu, L. (2016). Nonnegative adaptive lasso for ultra-high dimensional regression models and a two-stage method applied in financial modeling. *Journal of Statistical Planning and Inference*, 174, 52–67.

Yuan, M., and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1), 49–67.

Yohai, V.J., and Maronna, R.A. (1979). Asymptotic behavior of M-estimators for the linear model. *The Annals of Statistics*, 7, 258–268.

Zellner, A. (1962). An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias. *Journal of the American Statistical Association*, 57, 348–368.

Zhang, C.H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, 38(2), 894–942.

Zhang, C., Zhi, R., Li, T., and Corchado, J. (2016). Adaptive m-estimation for robust cubature kalman filtering. *In 2016 Sensor Signal Processing for Defence (SSPD)*, 1–5. IEEE.

Zhu, Y., Shen, X., and Pan, W. (2013). Simultaneous grouping pursuit and feature selection over an undirected graph. *Journal of the American Statistical Association*, 108(502), 713–725.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476), 1418–1429.

Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: Series B (statistical methodology)*, 67(2), 301–320.

A Appendix

This part contains the proofs of those lemmas and theorems in Chapter 2. These proofs are all under the assumptions presented in Chapter 2.

Proof of Lemma 2.1. Since $\hat{\boldsymbol{\alpha}}^{ols} = (\hat{\alpha}_1^{ols}, \ldots, \hat{\alpha}_{K^0}^{ols})^\top = (Z_{\mathcal{G}_0^{0c}}^\top Z_{\mathcal{G}_0^{0c}})^{-1} Z_{\mathcal{G}_0^{0c}}^\top \boldsymbol{y} = \boldsymbol{\alpha}^0 + (Z_{\mathcal{G}_0^{0c}}^\top Z_{\mathcal{G}_0^{0c}})^{-1} Z_{\mathcal{G}_0^{0c}}^\top \boldsymbol{\epsilon}$, $\hat{\boldsymbol{\alpha}}^{ols} \sim N\left(\boldsymbol{\alpha}^0, \sigma^2 (Z_{\mathcal{G}_0^{0c}}^\top Z_{\mathcal{G}_0^{0c}})^{-1}\right)$, namely,

$$\hat{\alpha}_k^{ols} - \alpha_k^0 \sim N\left(0, \sigma^2 (Z_{\mathcal{G}_0^{0c}}^\top Z_{\mathcal{G}_0^{0c}})_{kk}^{-1}\right), k = 1, \ldots, K^0,$$

where $(Z_{\mathcal{G}_0^{0c}}^\top Z_{\mathcal{G}_0^{0c}})_{kk}^{-1}$ denotes the k-th diagonal element of matrix $(Z_{\mathcal{G}_0^{0c}}^\top Z_{\mathcal{G}_0^{0c}})^{-1}$.

By the assumption (A2), it yields that the variance of $\hat{\alpha}_k^{ols}$ is bounded from above by $\sigma^2/(nc_0)$ for all $k = 1, \ldots, K^0$. In view of the assumption (A3), $\min_{1 \le k \le K^0} \alpha_k^0 = \min_{j \in \mathcal{G}_0^{0c}} \beta_j^0 > c_n$, where $c_n = [2\sigma^2 \log\{2nK^0/(2\pi)^{1/2}\}/(nc_0)]^{1/2}$. Similar to Meinshausen (2013), by Bonferroni's inequality, we thus have

$$\|\hat{\boldsymbol{\alpha}}^{ols} - \boldsymbol{\alpha}^0\|_\infty \le c_n,$$

with probability at least

$$1 - 2K^0 \left\{1 - \Phi\left(c_n (nc_0)^{1/2}/\sigma\right)\right\} = 1 - 2K^0 \left\{1 - \Phi\left([2\log\{2nK^0/(2\pi)^{1/2}\}]^{1/2}\right)\right\}.$$

It implies that, with the same probability, $\min_{1\le k\le K^0} \hat{\alpha}_k^{ols} > 0$, and thus $\hat{\boldsymbol{\alpha}}^{ora} = \hat{\boldsymbol{\alpha}}^{ols}$ or $\hat{\boldsymbol{\beta}}^{ora} = \hat{\boldsymbol{\beta}}^{ols}$. That is,

$$P\left(\hat{\boldsymbol{\beta}}^{ora} \neq \hat{\boldsymbol{\beta}}^{ols}\right) \le 2K^0 \left\{1 - \Phi\left([2\log\{2nK^0/(2\pi)^{1/2}\}]^{1/2}\right)\right\}.$$

Since $1-\Phi(x) \le (2\pi)^{-1/2}x^{-1}\exp(-x^2/2)$ for any $x > 0$, it follows that

$$P\left(\hat{\boldsymbol{\beta}}^{ora} \neq \hat{\boldsymbol{\beta}}^{ols}\right) \le \frac{1}{n}\frac{1}{[2\log\{2nK^0/(2\pi)^{1/2}\}]^{1/2}} = O\left(\frac{1}{n(\log n)^{1/2}}\right).$$

Proof of Theorem 2.1. Let $\mathcal{G}$ be a grouping of the constrained problem in Section 2.2, where $\mathcal{G} = (\mathcal{G}_0, \mathcal{G}_1, \ldots, \mathcal{G}_K)$, satisfying that $0 \le \hat{\beta}_j^{cons} \le \tau$ if $j \in \mathcal{G}_0$, $|\hat{\beta}_j^{cons} - \hat{\beta}_{j'}^{cons}| > \tau$ if $j \in \mathcal{G}_k$, $j' \in \mathcal{G}_{k'}$, $j = 1, \ldots, p; 1 \le k \neq k' \le K$.

If $\mathcal{G} = \mathcal{G}^0$, then $|\mathcal{G}_0^c| = s_1^0$. By the first constraint $\sum_{j=1}^p \min\{\beta_j/\tau, 1\} \le s_1$, $\sum_{j\in\mathcal{G}_0} \hat{\beta}_j^{cons}/\tau + s_1^0 \le s_1^0$, which implies that $\hat{\beta}_j^{cons} = 0$, $j \in \mathcal{G}_0$. By the second constraint $\sum_{(j,j')\in\varepsilon} \min\{|\beta_j - \beta_{j'}|/\tau, 1\} \le s_2$, similarly, we obtain that $\hat{\beta}_j^{cons} = \hat{\beta}_{j'}^{cons}$, $j, j' \in \mathcal{G}_k = \mathcal{G}_k^0$, $(j,j') \in \varepsilon$, $k = 1, \ldots, K$. Thus, $\hat{\boldsymbol{\beta}}^{cons} = \hat{\boldsymbol{\beta}}^{ora}$ if $\mathcal{G} = \mathcal{G}^0$, which, together with the fact that $P(\hat{\boldsymbol{\beta}}^{cons} \neq \hat{\boldsymbol{\beta}}^{ora}) = P(\hat{\boldsymbol{\beta}}^{cons} \neq \hat{\boldsymbol{\beta}}^{ora}, \mathcal{G} \neq \mathcal{G}^0) + P(\hat{\boldsymbol{\beta}}^{cons} \neq \hat{\boldsymbol{\beta}}^{ora}, \mathcal{G} = \mathcal{G}^0)$, yields that

$$P\left(\hat{\boldsymbol{\beta}}^{cons} \neq \hat{\boldsymbol{\beta}}^{ora}\right) = P\left(\hat{\boldsymbol{\beta}}^{cons} \neq \hat{\boldsymbol{\beta}}^{ora}, \mathcal{G} \neq \mathcal{G}^0\right). \tag{A.1}$$

Denote $\bar{S}(\boldsymbol{\beta}) = 2^{-1}\|Y - X\boldsymbol{\beta}\|^2$. In view that $P(\hat{\boldsymbol{\beta}}^{cons} \neq \hat{\boldsymbol{\beta}}^{ora}, \mathcal{G} \neq \mathcal{G}^0) = P(\hat{\boldsymbol{\beta}}^{cons} \neq$

$\hat{\boldsymbol{\beta}}^{ora}, \hat{\boldsymbol{\beta}}^{ora} = \hat{\boldsymbol{\beta}}^{ols}, \mathcal{G} \neq \mathcal{G}^0) + P(\hat{\boldsymbol{\beta}}^{cons} \neq \hat{\boldsymbol{\beta}}^{ora}, \hat{\boldsymbol{\beta}}^{ora} \neq \hat{\boldsymbol{\beta}}^{ols}, \mathcal{G} \neq \mathcal{G}^0)$, (A.1) thus becomes

$$P\left(\hat{\boldsymbol{\beta}}^{cons} \neq \hat{\boldsymbol{\beta}}^{ora}\right) \leq P\left(\bar{S}(\hat{\boldsymbol{\beta}}^{cons}) - \bar{S}(\hat{\boldsymbol{\beta}}^{ora}) \leq 0, \hat{\boldsymbol{\beta}}^{ora} = \hat{\boldsymbol{\beta}}^{ols}, \mathcal{G} \neq \mathcal{G}^0\right) + P\left(\hat{\boldsymbol{\beta}}^{ora} \neq \hat{\boldsymbol{\beta}}^{ols}\right). \tag{A.2}$$

The second term in (A.2) has already proved in Lemma 2.1. Next, we work on the first term in (A.2), and denote it by Γ.

Consider the case where $\hat{\boldsymbol{\beta}}^{ora} = \hat{\boldsymbol{\beta}}^{ols}$ and $\mathcal{G} \neq \mathcal{G}^0$. Define $\bar{\boldsymbol{\beta}} = (\bar{\beta}_1, \ldots, \bar{\beta}_p)^\top$, satisfying

$$\bar{\beta}_j = \begin{cases} \frac{\sum_{j' \in \mathcal{G}_k} \hat{\beta}_{j'}^{cons}}{|\mathcal{G}_k|}, & if\ j \in \mathcal{G}_k, k = 1, \ldots, K, \\ 0, & if\ j \in \mathcal{G}_0. \end{cases}$$

It follows that $|\bar{\beta}_j - \hat{\beta}_j^{cons}| \leq \tau$, $\|\bar{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^{cons}\|^2 \leq \tau^2 p$, and thus

$$\|X(\bar{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^{cons})\|^2 \leq \lambda_{\max}(X^\top X)\tau^2 p. \tag{A.3}$$

Note that

$$\|Y - X\bar{\boldsymbol{\beta}}\|^2 \geq \|Y - P_{Z_{\mathcal{G}_0^c}} Y\|^2 = \|(I - P_{Z_{\mathcal{G}_0^c}})X\boldsymbol{\beta}^0 + (I - P_{Z_{\mathcal{G}_0^c}})\boldsymbol{\epsilon}\|^2. \tag{A.4}$$

For any vector $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^p$ and $a > 0$, it holds that $\|\boldsymbol{u} + \boldsymbol{v}\|^2 \geq a^{-1}(a-1)\|\boldsymbol{u}\|^2 - (a-1)\|\boldsymbol{v}\|^2$ (Shen et al., 2012). We thus have

$$\bar{S}(\hat{\boldsymbol{\beta}}^{cons}) = \frac{1}{2}\left\|Y - X\hat{\boldsymbol{\beta}}^{cons}\right\|^2 \geq \frac{a-1}{2a}\left\|Y - X\bar{\boldsymbol{\beta}}\right\|^2 - \frac{a-1}{2}\left\|X(\bar{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^{cons})\right\|^2. \tag{A.5}$$

By substituting (A.3)-(A.4) into (A.5) and together with $\bar{S}(\hat{\boldsymbol{\beta}}^{ols}) = 2^{-1}\|(I-P_{Z_{\mathcal{G}_0^{0c}}})\boldsymbol{\epsilon}\|^2 \leq 2^{-1}\boldsymbol{\epsilon}^\top\boldsymbol{\epsilon}$, we obtain that, for any $a > 1$,

$$2a\left\{\bar{S}(\hat{\boldsymbol{\beta}}^{cons}) - \bar{S}(\hat{\boldsymbol{\beta}}^{ora})\right\} = 2a\left\{\bar{S}(\hat{\boldsymbol{\beta}}^{cons}) - \bar{S}(\hat{\boldsymbol{\beta}}^{ols})\right\} \geq -L_1 - L_2 + L_3,$$

where $L_1 = \{\boldsymbol{\epsilon} - (a-1)(I - P_{Z_{\mathcal{G}_0^c}})X\boldsymbol{\beta}^0\}^\top(I - P_{Z_{\mathcal{G}_0^c}})\{\boldsymbol{\epsilon} - (a-1)(I - P_{Z_{\mathcal{G}_0^c}})X\boldsymbol{\beta}^0\}$, and $L_1\sigma^{-2}$ follows noncentral chi-squared distribution $\chi^2_{k,\Lambda}$ with degrees of freedom $k = \max\{n - K(\mathcal{G}_0^c), 0\}$, and noncentral parameter $\Lambda = (a-1)^2\|(I - P_{Z_{\mathcal{G}_0^c}})X\boldsymbol{\beta}^0\|^2/\sigma^2$; $L_2 = a\boldsymbol{\epsilon}^\top P_{Z_{\mathcal{G}_0^c}}\boldsymbol{\epsilon}$ is independent of L_1, and $a^{-1}\sigma^{-2}L_2$ follows chi-squared distribution χ^2_κ with degrees of freedom $\kappa = K(\mathcal{G}_0^c)$; $L_3 = a(a-1)\|(I - P_{Z_{\mathcal{G}_0^c}})X\boldsymbol{\beta}^0\|^2 - a(a-1)\lambda_{\max}(X^\top X)\tau^2 p$. Note that, by the definition of $C_{\min}$, $\|(I - P_{Z_{\mathcal{G}_0^c}})X\boldsymbol{\beta}^0\|^2 \geq nC_{\min}$.

For Γ, by Markov inequality and moment-generating function of chi-squared distribution, it holds that, for any $0 < t < 1/(2a)$ and $1 - 2at < 1 - 2t < 1$ $(a > 1)$, by Shen et al. (2012),

$$\begin{aligned}\Gamma \leq &\sum_{i=1}^{s_1^0}\sum_{j=0}^{i}\binom{p - s_1^0}{j}\binom{s_1^0}{s_1^0 - i}T_i \times \\ &\frac{\exp\left\{\frac{t(a-1)^2 niC_{\min}}{(1-2t)\sigma^2}\right\}\exp\left[-\frac{t}{\sigma^2}\left\{-a(a-1)\lambda_{\max}(X^\top X)p\tau^2 + a(a-1)niC_{\min}\right\}\right]}{(1-2t)^{\frac{n-K_i^*}{2}}(1-2at)^{\frac{K_i^*}{2}}} \\ \leq &\sum_{i=1}^{s_1^0}\sum_{j=0}^{i}\binom{p - s_1^0}{j}\binom{s_1^0}{s_1^0 - i}T_i \exp\left\{\frac{(a-1)\log p}{4n} - n\frac{t(a-1)iC_{\min}}{\sigma^2}\frac{1-2at}{1-2t}\right\} \\ &\left(\frac{1-2t}{1-2at}\right)^{K_i^*/2}\frac{1}{(1-2t)^{n/2}},\end{aligned}$$

where $K_i^* = \max_{\{\mathcal{G}\in\mathcal{T}, |\mathcal{G}_0 \backslash \mathcal{G}_0^0|=i\}} K(\mathcal{G}_0^c)$. Note that the last inequality holds true because

$$\frac{t}{\sigma^2} a(a-1)\lambda_{\max}(X^\top X)p\tau^2 \le \frac{2ta(a-1)\log p}{4n} \le \frac{(a-1)\log p}{4n}$$

for any $\tau \le \sigma[\log p/\{2np\lambda_{\max}(X^\top X)\}]^{1/2}$. We choose $a = 4+n/4$, $t = 4^{-1}(a-1)^{-1}$, and define $b = (1-2t)/(1-2at)$. Then $b = (2a-3)/(a-2) < 5/2$, and $(a-1)/(4n) \le 1$. Since $-\log(1-x) \le x(1-x)^{-1}$ for $0 < x < 1$, and $0 < 2t = 2^{-1}(a-1)^{-1} < 1$, it follows that

$$-\frac{n}{2}\log(1-2t) \le \frac{n}{2}\frac{1/\{2(a-1)\}}{1-1/\{2(a-1)\}} \le \frac{n}{2}\frac{1}{2(4+n/4)-3} \le 1,$$

which, jointly with the facts

$$\binom{s_1^0}{s_1^0 - i} \le (s_1^0)^i, \sum_{j=0}^{i}\binom{p-s_1^0}{j} \le (p-s_i^0)^i \text{ and } (p-s_1^0)s_1^0 \le p^2/4$$

yields that

$$\begin{aligned}\Gamma \le& \sum_{i=1}^{s_1^0}\left(\frac{p^2}{4}\right)^i T_i \exp\left(\frac{(a-1)\log p}{4n} - n\frac{iC_{\min}}{4c\sigma^2}\right) b^{K_i^*/2}\frac{1}{(1-2t)^{n/2}} \\ \le& \exp(1)\sum_{i=1}^{s_1^0}\exp\left(i(3\log p + \bar{T} - \frac{nC_{\min}}{10\sigma^2} + \frac{\bar{K}}{2})\right) \\ \le& \exp(1)\sum_{i=1}^{s_1^0}\exp\left(-i\frac{n}{10\sigma^2}\left(C_{\min} - \frac{10\sigma^2}{n}(3\log p + \bar{T} + \bar{K}/2)\right)\right). \qquad \text{(A.6)}\end{aligned}$$

Since $(1-z)^{-1} = \sum_{i=0}^{\infty} z^i$ for $|z| < 1$, we thus obtain that, for $x < 0$,

$$\sum_{i=1}^{s_1^0}\exp(ix) \le -1 + \frac{1}{1-\exp(x)} = \frac{\exp(x)}{1-\exp(x)}.$$

We take $x = -10^{-1}\sigma^{-2}n\{C_{\min} - 10\sigma^2 n^{-1}(3\log p + \bar{T} + \bar{K}/2)\}$ if $C_{\min} > 10\sigma^2 n^{-1}(3\log p + \bar{T} + \bar{K}/2)$. Together with $\Gamma \leq 1$, (A.6) becomes

$$\Gamma \leq \{\exp(1) + 1\} \exp\left(-\frac{n}{10\sigma^2}\left(C_{\min} - \frac{10\sigma^2}{n}(3\log p + \bar{T} + \bar{K}/2)\right)\right). \tag{A.7}$$

Similarly, we can show that (A.7) still holds for $C_{\min} \leq 10\sigma^2 n^{-1}(3\log p + \bar{T} + \bar{K}/2)$. By Lemma 2.1 and (A.7), (A.2) becomes

$$\begin{aligned} P\left(\hat{\boldsymbol{\beta}}^{cons} \neq \hat{\boldsymbol{\beta}}^{ora}\right) \leq & \{\exp(1) + 1\} \exp\left(-\frac{n}{10\sigma^2}\left\{C_{\min} - \frac{10\sigma^2}{n}(3\log p + \bar{T} + \bar{K}/2)\right\}\right) \\ & + \frac{c}{n(\log n)^{1/2}}. \end{aligned} \tag{A.8}$$

(1) If $C_{\min} \geq 10\sigma^2 n^{-1}\left(\log n + 2^{-1}\log\log n + 3\log p + \bar{T} + \bar{K}/2\right)$, by (A.8),

$$P\left(\hat{\boldsymbol{\beta}}^{cons} \neq \hat{\boldsymbol{\beta}}^{ora}\right) = O\left(\frac{1}{n(\log n)^{1/2}}\right).$$

(2) We denote $T_1 = n^{-1}E(\|X\hat{\boldsymbol{\beta}}^{cons} - X\boldsymbol{\beta}^0\|^2 I_{\{G\}})$, and $T_2 = n^{-1}E(\|X\hat{\boldsymbol{\beta}}^{cons} - X\boldsymbol{\beta}^0\|^2 I_{\{G^c\}})$, where $G = \{n^{-1}\|X\hat{\boldsymbol{\beta}}^{cons} - X\boldsymbol{\beta}^0\|^2 \geq 25\sigma^2\}$. It is easy to see that

$$\frac{1}{n}E\left\|X\hat{\boldsymbol{\beta}}^{cons} - X\boldsymbol{\beta}^0\right\|^2 = T_1 + T_2.$$

Now, we work on T_1. By the definition, $T_1 = \int_{25\sigma^2}^{\infty} P(n^{-1}\|X\hat{\boldsymbol{\beta}}^{cons} - X\boldsymbol{\beta}^0\|^2 \geq x)dx +$

$25\sigma^2 P(n^{-1}\|X\hat{\boldsymbol{\beta}}^{cons} - X\boldsymbol{\beta}^0\|^2 \geq 25\sigma^2)$. For the first term of T_1,

$$\begin{aligned}
&\int_{25\sigma^2}^{\infty} P\left(n^{-1}\|X\hat{\boldsymbol{\beta}}^{cons} - X\boldsymbol{\beta}^0\|^2 \geq x\right) dx \\
\leq &\int_{25\sigma^2}^{\infty} P\left(4n^{-1}\|\boldsymbol{\epsilon}\|^2 \geq x\right) dx \\
\leq &\int_{25\sigma^2}^{\infty} E\left\{\exp\left(\frac{\|\boldsymbol{\epsilon}\|^2}{3\sigma^2}\right)\right\} \exp\left(-\frac{nx}{12\sigma^2}\right) dx \\
= &\int_{25\sigma^2}^{\infty} \exp\left(-\frac{n}{12\sigma^2}\{x - 6(\log 3)\sigma^2\}\right) dx \\
< &\int_{25\sigma^2}^{\infty} \exp\left\{-\frac{n}{12\sigma^2}(x - 24\sigma^2)\right\} dx \\
= &\frac{12\sigma^2}{n} \exp\left(-\frac{n}{12}\right) = o\left(\frac{K^0\sigma^2}{n}\right). \qquad \text{(A.9)}
\end{aligned}$$

Since $\|X\hat{\boldsymbol{\beta}}^{cons} - X\boldsymbol{\beta}^0\|^2 \leq 2(\|Y - X\hat{\boldsymbol{\beta}}^{cons}\|^2 + \|Y - X\boldsymbol{\beta}^0\|^2) \leq 4\|Y - X\boldsymbol{\beta}^0\|^2 = 4\|\boldsymbol{\epsilon}\|^2$, the first '$\leq$' follows. The second '$\leq$' is obtained by the Markov inequality. In view of the moment generating function for Chi-squared distribution, the first '$=$' holds. For the second term of T_1,

$$25\sigma^2 P(n^{-1}\|X\hat{\boldsymbol{\beta}}^{cons} - X\boldsymbol{\beta}^0\|^2 \geq 25\sigma^2) \leq 25\sigma^2 \exp(-n/12) = o\left(\frac{K^0\sigma^2}{n}\right). \qquad \text{(A.10)}$$

By (A.9) and (A.10), we thus have $T_1 = o(K^0\sigma^2/n)$.

On the other hand,

$$\begin{aligned}
T_2 = &E\left(\frac{1}{n}\left\|X\hat{\boldsymbol{\beta}}^{cons} - X\boldsymbol{\beta}^0\right\|^2 I_{\{G^c\}} I_{\{\hat{\boldsymbol{\beta}}^{cons} \neq \hat{\boldsymbol{\beta}}^{ols}\}}\right) \qquad \text{(A.11)} \\
&+ E\left(\frac{1}{n}\left\|X\hat{\boldsymbol{\beta}}^{cons} - X\boldsymbol{\beta}^0\right\|^2 \left(1 - I_{\{G\}}\right) I_{\{\hat{\boldsymbol{\beta}}^{cons} = \hat{\boldsymbol{\beta}}^{ols}\}}\right). \qquad \text{(A.12)}
\end{aligned}$$

For (A.11), it follows that

$$
\begin{aligned}
&E\left(\frac{1}{n}\left\|X\hat{\boldsymbol{\beta}}^{cons}-X\boldsymbol{\beta}^0\right\|^2 I_{\{G^c\}}I_{\{\hat{\boldsymbol{\beta}}^{cons}\neq\hat{\boldsymbol{\beta}}^{ols}\}}\right)\\
&\leq 25\sigma^2 P\left(\hat{\boldsymbol{\beta}}^{cons}\neq\hat{\boldsymbol{\beta}}^{ols}\right)\leq 25\sigma^2 P\left(\hat{\boldsymbol{\beta}}^{cons}\neq\hat{\boldsymbol{\beta}}^{ora}\right)+25\sigma^2 P\left(\hat{\boldsymbol{\beta}}^{ora}\neq\hat{\boldsymbol{\beta}}^{ols}\right)\\
&\leq 25\sigma^2\left\{\exp(1)+1\right\}\exp\left(-\frac{n}{10\sigma^2}\left\{C_{\min}-\frac{10\sigma^2}{n}(3\log p+\bar{T}+\bar{K}/2)\right\}\right)+25\sigma^2\frac{2c}{n\sqrt{\log n}}\\
&\leq\frac{100\sigma^2}{n(\log n)^{1/2}}+\frac{50\sigma^2 c}{n(\log n)^{1/2}}=o\left(\frac{K^0\sigma^2}{n}\right). \qquad (A.13)
\end{aligned}
$$

For (A.12),

$$
\begin{aligned}
E\left(\frac{1}{n}\left\|X\hat{\boldsymbol{\beta}}^{cons}-X\boldsymbol{\beta}^0\right\|^2 I_{\{G\}}I_{\{\hat{\boldsymbol{\beta}}^{cons}=\hat{\boldsymbol{\beta}}^{ols}\}}\right)\leq&E\left(\frac{1}{n}\left\|X\hat{\boldsymbol{\beta}}^{cons}-X\boldsymbol{\beta}^0\right\|^2 I_{\{G\}}\right)\\
=&o\left(\frac{K^0\sigma^2}{n}\right), \qquad (A.14)
\end{aligned}
$$

and

$$
\begin{aligned}
E\left(\frac{1}{n}\left\|X\hat{\boldsymbol{\beta}}^{cons}-X\boldsymbol{\beta}^0\right\|^2 I_{\{\hat{\boldsymbol{\beta}}^{cons}=\hat{\boldsymbol{\beta}}^{ols}\}}\right)=&\frac{1}{n}E\left(\left\|X\hat{\boldsymbol{\beta}}^{ols}-X\boldsymbol{\beta}^0\right\|^2\right)\\
=&\frac{1}{n}E\left(\left\|P_{Z_{\mathcal{G}_0^{0c}}}\boldsymbol{\epsilon}\right\|^2\right)=\frac{K^0\sigma^2}{n}. \qquad (A.15)
\end{aligned}
$$

By (A.11)-(A.15), $T_2=n^{-1}K^0\sigma^2(1+o(1))$. Therefore,

$$
\frac{1}{n}E\left(\left\|X\hat{\boldsymbol{\beta}}^{cons}-X\boldsymbol{\beta}^0\right\|^2\right)=T_1+T_2=\frac{K^0\sigma^2}{n}(1+o(1)).
$$

Proof of Theorem 2.2. This proof mimics the proof of Theorem 1 in (Shen et al., 2012). By theorem 4.1 in Tseng (2001), the coordinate descent method converges. By Theorem 17.6 in Wright and Nocedal (1999), the augmented Lagrange method converges as well, and the convergence rate is linear. Then for $m \in N_+,\ 0 \leq S(\hat{\boldsymbol{\beta}}^{(m)}) = S^{(m+1)}(\hat{\boldsymbol{\beta}}^{(m)}) \leq S^{(m)}(\hat{\boldsymbol{\beta}}^{(m)}) \leq S^{(m)}(\hat{\boldsymbol{\beta}}^{(m-1)}) = S(\hat{\boldsymbol{\beta}}^{(m-1)})$. Hence $S(\hat{\boldsymbol{\beta}}^{(m)}) \to c, c \geq 0$ as $m \to \infty$, which leads to the convergence of the algorithm 1.

Proof of Theorem 2.3. By the algorithm 1 presented in Section 2.3.2, there exists a finite m^* such that $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{(m^*)}$. Denote the grouping of $\hat{\boldsymbol{\beta}}$ by $\mathcal{G} = (\mathcal{G}_0, \mathcal{G}_1, \ldots, \mathcal{G}_K)$ with $K < K^*$. Then $\hat{\boldsymbol{\beta}}$ satisfies that, for grouping $\mathcal{G}$,

$$\begin{cases} -(X_{\mathcal{G}_k}\mathbf{1})^\top(\boldsymbol{y} - X\boldsymbol{\beta}) + n\sum\limits_{j\in\mathcal{G}_k} \Delta_j(\boldsymbol{\beta}) = 0 & k = 1, \ldots, K \\ |(X_A\mathbf{1})^\top(\boldsymbol{y} - X\boldsymbol{\beta}) - n\sum\limits_{j\in A} \Delta_j(\boldsymbol{\beta})| \leq n\frac{\lambda_2}{\tau}|\varepsilon \cap \{A \times (\mathcal{G}_k \setminus A)\}| & A \subset \mathcal{G}_k, |\mathcal{G}_k| > 1 \\ |\boldsymbol{x}_{(j)}^\top(\boldsymbol{y} - X\boldsymbol{\beta}) - n\Delta_j(\boldsymbol{\beta})| \leq n\frac{\lambda_1}{\tau} & j \in \mathcal{G}_0, \end{cases} \tag{A.16}$$

where

$$\Delta_j(\boldsymbol{\beta}) = \lambda_1\tau^{-1}\text{sign}(\beta_j)I_{\{|\beta_j|\leq\tau\}} + \lambda_2\tau^{-1}\sum_{j':(j',j)\in\varepsilon} \text{sign}(\beta_j - \beta_{j'})I_{\{|\beta_j-\beta_{j'}|\leq\tau\}} + 2\lambda_3\beta_j I_{\{\beta_j<0\}}.$$

Denote $\mathcal{J} = \mathcal{J}_{11} \cap \mathcal{J}_{12} \cap \mathcal{J}_{21} \cap \mathcal{J}_{22}$, where $\mathcal{J}_{11} = \{\min\limits_{j\notin\mathcal{G}_0^0} \hat{\beta}_j^{ols} > 2\tau\}$, $\mathcal{J}_{12} = \{\max\limits_{j\in\mathcal{G}_0^0} |\boldsymbol{x}_{(j)}^\top(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}^{ols})| \leq n\lambda_1\tau^{-1}\}$, $\mathcal{J}_{21} = \{\min\limits_{1\leq k<l\leq K^0} |\hat{\alpha}_k^{ols} - \hat{\alpha}_l^{ols}| > 2\tau\}$, $\mathcal{J}_{22} = \cap_{k=1,\ldots,K^0:|\mathcal{G}_k^0|>1}\{\max_{A\subset\mathcal{G}_k^0} |(X_A\mathbf{1})^\top(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}^{ols})| \leq n\lambda_2\tau^{-1}|\varepsilon \cap \{A \times (\mathcal{G}_k^0 \setminus A)\}|\}$. First, we show that $\hat{\boldsymbol{\beta}}^{ols}$ is a solution to (A.16) on $\mathcal{J}$. Note that, $\sum_{j\in\mathcal{G}_k^0} \Delta_j(\hat{\boldsymbol{\beta}}^{ols}) = 0$ on the set $\mathcal{J}_{11} \cap \mathcal{J}_{21}$. By the definition of $\hat{\boldsymbol{\beta}}^{ols}$, $(X_{\mathcal{G}_k^0}\mathbf{1})^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^{ols}) = 0$. Thus the first equation in (A.16) holds for $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}^{ols}$. Since $\sum_{j \in \mathcal{G}_k^0} \Delta_j(\hat{\boldsymbol{\beta}}^{ols}) = 0$ on $\mathcal{J}$, one can easily see that the second and third inequalities also hold for $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}^{ols}$.

Next, we show that (A.16) has a unique solution on $\mathcal{J}$, and thus $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{ols}$. We provide the proof by contradiction. Assume that $\hat{\boldsymbol{\beta}} \neq \hat{\boldsymbol{\beta}}^{ols}$. Let $\mathcal{H} = (\mathcal{H}_1, \ldots, \mathcal{H}_L) = \mathcal{G}_0^c \vee \mathcal{G}_0^{0c}$. Herein, we give an example to explain the sign '$\vee$'. Define two sets $A_1 = \{\{1,2,3,4\},\{5,6\}\}$, and $A_2 = \{\{1,2\},\{3,4,5,6\},\{7\}\}$. Then $A_1 \vee A_2 = \{\{1,2\},\{3,4\},\{5,6\},\{7\}\}$. Denote $\hat{\boldsymbol{\alpha}}_{\mathcal{H}}^{ols} = (\hat{\alpha}_{\mathcal{H}_1}^{ols}, \ldots, \hat{\alpha}_{\mathcal{H}_L}^{ols})^\top$, and $\hat{\boldsymbol{\alpha}}_{\mathcal{H}} = (\hat{\alpha}_{\mathcal{H}_1}, \ldots, \hat{\alpha}_{\mathcal{H}_L})^\top$ the coefficients estimated by OLS and the algorithm 1, respectively. Then $S(\boldsymbol{\alpha}_{\mathcal{H}}) = (2n)^{-1}\|\boldsymbol{y} - Z_{\mathcal{H}}\boldsymbol{\alpha}_{\mathcal{H}}\|^2 + J(\boldsymbol{\alpha}_{\mathcal{H}})$, where

$$\begin{aligned} J(\boldsymbol{\alpha}_{\mathcal{H}}) =& \lambda_1 \sum_{k=1}^{L} |\mathcal{H}_k| \min\left\{\frac{|\alpha_{\mathcal{H}_k}|}{\tau}, 1\right\} + \lambda_2 \sum_{1 \le k < l \le L} |\varepsilon_{kl}| \min\left\{\frac{|\alpha_{\mathcal{H}_k} - \alpha_{\mathcal{H}_l}|}{\tau}, 1\right\} \\ &+ \lambda_3 \sum_{k=1}^{L} |\mathcal{H}_k| (\min\{\alpha_{\mathcal{H}_k}, 0\})^2 \end{aligned}$$

for $\boldsymbol{\alpha}_{\mathcal{H}} = (\alpha_{\mathcal{H}_1}, \ldots, \alpha_{\mathcal{H}_L})^\top$, where ε_{kl} is the set of undirected edge between $\mathcal{H}_k$ and $\mathcal{H}_l$. We thus have

$$\frac{\partial S(\hat{\boldsymbol{\alpha}}_{\mathcal{H}})}{\partial \boldsymbol{\alpha}_{\mathcal{H}}} - \frac{\partial S(\hat{\boldsymbol{\alpha}}_{\mathcal{H}}^{ols})}{\partial \boldsymbol{\alpha}_{\mathcal{H}}} = \frac{1}{n} Z_{\mathcal{H}}^\top Z_{\mathcal{H}} (\hat{\boldsymbol{\alpha}}_{\mathcal{H}} - \hat{\boldsymbol{\alpha}}_{\mathcal{H}}^{ols}) + \boldsymbol{\varphi},$$

where $\boldsymbol{\varphi} = (\varphi_1, \ldots, \varphi_L)^\top = \boldsymbol{\varphi}_1 + \boldsymbol{\varphi}_2$, $\boldsymbol{\varphi}_1 = (\varphi_{11}, \ldots, \varphi_{L1})^\top$, $\boldsymbol{\varphi}_2 = (\varphi_{12}, \ldots, \varphi_{L2})^\top$. For $k = 1, \ldots, L$, $\varphi_{k1} = \lambda_1 \tau^{-1} |\mathcal{H}_k| (a_k I_{\{|\hat{\alpha}_{\mathcal{H}_k}| \le \tau\}} - a_k^{ols} I_{\{|\hat{\alpha}_{\mathcal{H}_k}^{ols}| \le \tau\}}) + \lambda_2 \tau^{-1} \sum_{l \neq k} |\varepsilon_{kl}| (b_{kl} I_{\{|\hat{\alpha}_{\mathcal{H}_k} - \hat{\alpha}_{\mathcal{H}_l}| \le \tau\}} - b_{kl}^{ols} I_{\{|\hat{\alpha}_{\mathcal{H}_k}^{ols} - \hat{\alpha}_{\mathcal{H}_l}^{ols}| \le \tau\}})$, $\varphi_{k2} = 2\lambda_3 (|\mathcal{H}_k| \hat{\alpha}_{\mathcal{H}_k} I_{\{\hat{\alpha}_{\mathcal{H}_k} < 0\}} - |\mathcal{H}_k| \hat{\alpha}_{\mathcal{H}_k}^{ols} I_{\{\hat{\alpha}_{\mathcal{H}_k}^{ols} < 0\}})$, where $a_k = \text{sign}(\hat{\alpha}_{\mathcal{H}_k})$, if $\hat{\alpha}_{\mathcal{H}_k} \neq 0$, $a_k \in [-1, 1]$ otherwise; $b_{kl} = \text{sign}(\hat{\alpha}_{\mathcal{H}_k} - \hat{\alpha}_{\mathcal{H}_l})$ if $\hat{\alpha}_{\mathcal{H}_k} - \hat{\alpha}_{\mathcal{H}_l} \neq 0$, $b_{kl} \in [-1, 1]$

otherwise. Similarly, we have a_k^{ols} and b_{kl}^{ols}. Note that $\|\boldsymbol{\varphi}_1\|^2 \leq 4\tau^{-2}(\lambda_1 s^* + \lambda_2|\mathcal{N}|)^2$.

Now, we consider two cases: (i) $\|\hat{\boldsymbol{\alpha}}_{\mathcal{H}} - \hat{\boldsymbol{\alpha}}_{\mathcal{H}}^{ols}\| < \tau/2$ and (ii) $\|\hat{\boldsymbol{\alpha}}_{\mathcal{H}} - \hat{\boldsymbol{\alpha}}_{\mathcal{H}}^{ols}\| \geq \tau/2$. For each case, we show that both $\hat{\boldsymbol{\alpha}}_{\mathcal{H}}$ and $\hat{\boldsymbol{\alpha}}_{\mathcal{H}}^{ols}$ are the local minimizers of $S(\boldsymbol{\alpha}_{\mathcal{H}})$ and $\hat{\boldsymbol{\alpha}}_{\mathcal{H}} = \hat{\boldsymbol{\alpha}}_{\mathcal{H}}^{ols}$ on $\mathcal{J}$.

(i) $\|\hat{\boldsymbol{\alpha}}_{\mathcal{H}} - \hat{\boldsymbol{\alpha}}_{\mathcal{H}}^{ols}\| < \tau/2$. On the set $\mathcal{J}$, $\hat{\alpha}_{\mathcal{H}_k} \geq \hat{\alpha}_{\mathcal{H}_k}^{ols} - |\hat{\alpha}_{\mathcal{H}_k} - \hat{\alpha}_{\mathcal{H}_k}^{ols}| \geq 2\tau - \tau/2 > \tau$ if $\hat{\alpha}_{\mathcal{H}_k}^{ols} > 2\tau$; $|\hat{\alpha}_{\mathcal{H}_k}| < |\hat{\alpha}_{\mathcal{H}_k}^{ols}| + |\hat{\alpha}_{\mathcal{H}_k} - \hat{\alpha}_{\mathcal{H}_k}^{ols}| < \tau/2$ if $|\hat{\alpha}_{\mathcal{H}_k}^{ols}| = 0$; $|\hat{\alpha}_{\mathcal{H}_k} - \hat{\alpha}_{\mathcal{H}_l}| \geq -|\hat{\alpha}_{\mathcal{H}_k} - \hat{\alpha}_{\mathcal{H}_k}^{ols}| - |\hat{\alpha}_{\mathcal{H}_l} - \hat{\alpha}_{\mathcal{H}_l}^{ols}| + |\hat{\alpha}_{\mathcal{H}_k}^{ols} - \hat{\alpha}_{\mathcal{H}_l}^{ols}| \geq \tau$ if $|\hat{\alpha}_{\mathcal{H}_k}^{ols} - \hat{\alpha}_{\mathcal{H}_l}^{ols}| \geq 2\tau$; $|\hat{\alpha}_{\mathcal{H}_k} - \hat{\alpha}_{\mathcal{H}_l}| \leq |\hat{\alpha}_{\mathcal{H}_k} - \hat{\alpha}_{\mathcal{H}_k}^{ols}| + |\hat{\alpha}_{\mathcal{H}_l} - \hat{\alpha}_{\mathcal{H}_l}^{ols}| + |\hat{\alpha}_{\mathcal{H}_k}^{ols} - \hat{\alpha}_{\mathcal{H}_l}^{ols}| < \tau$ if $|\hat{\alpha}_{\mathcal{H}_k}^{ols} - \hat{\alpha}_{\mathcal{H}_l}^{ols}| = 0$. It implies that both $\hat{\boldsymbol{\alpha}}_{\mathcal{H}}$ and $\hat{\boldsymbol{\alpha}}_{\mathcal{H}}^{ols}$ are the local minimizers of $S(\boldsymbol{\alpha}_{\mathcal{H}})$ and $\hat{\boldsymbol{\alpha}}_{\mathcal{H}} = \hat{\boldsymbol{\alpha}}_{\mathcal{H}}^{ols}$ on $\mathcal{J}$.

(ii) $\|\hat{\boldsymbol{\alpha}}_{\mathcal{H}} - \hat{\boldsymbol{\alpha}}_{\mathcal{H}}^{ols}\| \geq \tau/2$. By Cauchy-Schwarz inequality,

$$\left|\boldsymbol{\varphi}_1^\top(\hat{\boldsymbol{\alpha}}_{\mathcal{H}} - \hat{\boldsymbol{\alpha}}_{\mathcal{H}}^{ols})\right| \leq \frac{2}{\tau}\left(\lambda_1 s^* + \lambda_2|\mathcal{N}|\right)\|\hat{\boldsymbol{\alpha}}_{\mathcal{H}} - \hat{\boldsymbol{\alpha}}_{\mathcal{H}}^{ols}\|.$$

It is easy to verify that $(\hat{\alpha}_{\mathcal{H}_k} I_{\{\hat{\alpha}_{\mathcal{H}_k}<0\}} - \hat{\alpha}_{\mathcal{H}_k}^{ols} I_{\{\hat{\alpha}_{\mathcal{H}_k}^{ols}<0\}})(\hat{\alpha}_{\mathcal{H}_k} - \hat{\alpha}_{\mathcal{H}_k}^{ols}) \geq 0$, followed by

$$\boldsymbol{\varphi}_2^\top(\hat{\boldsymbol{\alpha}}_{\mathcal{H}} - \hat{\boldsymbol{\alpha}}_{\mathcal{H}}^{ols}) \geq 0.$$

By the assumption (A4),

$$\begin{aligned} &\left(\frac{\partial S(\hat{\boldsymbol{\alpha}}_{\mathcal{H}})}{\partial \boldsymbol{\alpha}_{\mathcal{H}}} - \frac{\partial S(\hat{\boldsymbol{\alpha}}_{\mathcal{H}}^{ols})}{\partial \boldsymbol{\alpha}_{\mathcal{H}}}\right)^\top \frac{\hat{\boldsymbol{\alpha}}_{\mathcal{H}} - \hat{\boldsymbol{\alpha}}_{\mathcal{H}}^{ols}}{\|\hat{\boldsymbol{\alpha}}_{\mathcal{H}} - \hat{\boldsymbol{\alpha}}_{\mathcal{H}}^{ols}\|} \\ \geq &\min_{K(\mathcal{H})\leq K^*} \frac{\tau}{2}\lambda_{\min}\left(\frac{1}{n}Z_{\mathcal{H}}^\top Z_{\mathcal{H}}\right) - \frac{2}{\tau}\left(\lambda_1 s^* + \lambda_2|\mathcal{N}|\right) > 0. \end{aligned} \tag{A.17}$$

On the other hand, $\frac{\partial S(\hat{\boldsymbol{\alpha}}_{\mathcal{H}})}{\partial \boldsymbol{\alpha}_{\mathcal{H}}} = 0$ and $\frac{\partial S(\hat{\boldsymbol{\alpha}}_{\mathcal{H}}^{ols})}{\partial \boldsymbol{\alpha}_{\mathcal{H}}} = 0$ on $\mathcal{J}$ if $\hat{\boldsymbol{\alpha}}_{\mathcal{H}} \neq \hat{\boldsymbol{\alpha}}_{\mathcal{H}}^{ols}$, which contracts to (A.17).

Therefore, the problem (A.16) has a unique solution on $\mathcal{J}$. That is $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{ols}$ on $\mathcal{J}$, which yields that

$$P(\hat{\boldsymbol{\beta}} \neq \hat{\boldsymbol{\beta}}^{ols}) \leq P(J^c) \leq P(\mathcal{J}_{11}^c) + P(\mathcal{J}_{12}^c) + P(\mathcal{J}_{21}^c) + P(\mathcal{J}_{12}^c). \tag{A.18}$$

Next, we show the bounds of $P(\mathcal{J}_{11}^c), P(\mathcal{J}_{12}^c), P(\mathcal{J}_{21}^c), P(\mathcal{J}_{12}^c)$.

Before proceeding, we provide the following inequality, for $x > 0$, $\Phi(-x) \leq (2\pi)^{-1/2}x^{-1}\exp(-x^2/2)$. If $x^2 \geq 2\log\{2na/(2\pi)^{1/2}\}$, $a \geq 1$, then $2a\Phi(-x) \leq cn^{-1}(\log n)^{-1/2}$.

For $\mathcal{J}_{11}^c$, by the assumptions (A1)-(A2), $\hat{\beta}_j^{ols} \sim N(\beta_j^0, var(\hat{\beta}_j^{ols}))$, where $var(\hat{\beta}_j^{ols}) \leq n^{-1}\sigma^2\lambda_{\min}^{-1}(n^{-1}Z_{\mathcal{G}_0^{0c}}^\top Z_{\mathcal{G}_0^{0c}})$. If $\gamma_{\min} > 2\tau$, and $\{(\gamma_{\min}-2\tau)n^{1/2}\lambda_{\min}^{1/2}(n^{-1}Z_{\mathcal{G}_0^{0c}}^\top Z_{\mathcal{G}_0^{0c}})\sigma^{-1}\}^2 \geq 2\log\{2n(p-|\mathcal{G}_0^0|)/(2\pi)^{1/2}\}$, then

$$\begin{aligned}
P(\mathcal{J}_{11}^c) \leq & \sum_{j\in\mathcal{G}_0^{0c}} P\left(\hat{\beta}_j^{ols} \leq 2\tau\right) \leq \sum_{j\in\mathcal{G}_0^{0c}} P(\beta_j^0 - |\hat{\beta}_j^{ols} - \beta_j^0| \leq 2\tau) \\
\leq & 2\left(p - |\mathcal{G}_0^0|\right)\Phi\left(-(\gamma_{\min} - 2\tau)n^{1/2}\lambda_{\min}^{1/2}(n^{-1}Z_{\mathcal{G}_0^{0c}}^\top Z_{\mathcal{G}_0^{0c}})\sigma^{-1}\right) \\
= & O\left(\frac{1}{n(\log n)^{1/2}}\right).
\end{aligned} \tag{A.19}$$

For $\mathcal{J}_{12}^c$, by the assumptions (A1)-(A2), $\boldsymbol{x}_{(j)}^\top(\boldsymbol{y} - X^\top\hat{\boldsymbol{\beta}}^{ols}) = \boldsymbol{x}_{(j)}^\top(I - P_{Z_{\mathcal{G}_0^{0c}}})\boldsymbol{\epsilon} \sim N(0, \sigma^2\|(I-P_{Z_{\mathcal{G}_0^{0c}}})\boldsymbol{x}_{(j)}\|^2)$, and $\|(I-P_{Z_{\mathcal{G}_0^{0c}}})\boldsymbol{x}_{(j)}\|^2 \leq \|\boldsymbol{x}_{(j)}\|^2$. If $(n\lambda_1\tau^{-1}\sigma^{-1}/\max_{1\leq i\leq p}\|\boldsymbol{x}_{(j)}\|)^2 \geq$

$2\log\{2n|\mathcal{G}_0^0|/(2\pi)^{1/2}\}$, then

$$P(\mathcal{J}_{12}^c) \leq \sum_{j\in\mathcal{G}_0^0} P\left(\left|\boldsymbol{x}_{(j)}^\top(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}^{ols})\right| > n\frac{\lambda_1}{\tau}\right)$$

$$\leq 2|\mathcal{G}_0^0|\Phi\left(-\frac{n\lambda_1/\tau}{\sigma \max\limits_{1\leq i\leq p}\|\boldsymbol{x}_{(j)}\|}\right) = O\left(\frac{1}{n(\log n)^{1/2}}\right). \qquad \text{(A.20)}$$

For $\mathcal{J}_{21}^c$, by the assumptions (A1)-(A2), $\hat{\alpha}_k^{ols} - \hat{\alpha}_l^{ols} \sim N(\alpha_k^0 - \alpha_l^0, var(\hat{\alpha}_k^{ols} - \hat{\alpha}_l^{ols}))$, where $\mathrm{var}(\hat{\alpha}_k^{ols} - \hat{\alpha}_l^{ols}) \leq 4n^{-1}\sigma^2\lambda_{\min}^{-1}(n^{-1}Z_{\mathcal{G}_0^{0c}}^\top Z_{\mathcal{G}_0^{0c}})$. If $\gamma_{\min} > 2\tau$, and $\{2^{-1}\sigma^{-1}(\gamma_{\min} - 2\tau)n^{1/2}\lambda_{\min}^{1/2}(n^{-1}Z_{\mathcal{G}_0^{0c}}^\top Z_{\mathcal{G}_0^{0c}})\}^2 \geq 2\log\{nK^0(K^0-1)/(2\pi)^{1/2}\}$, then

$$\begin{aligned}
P(\mathcal{J}_{21}^c) \leq & \sum_{1\leq k<l\leq K^0} P(|\hat{\alpha}_k - \hat{\alpha}_l| \leq 2\tau) \\
\leq & \sum_{1\leq k<l\leq K^0} P(|\alpha_k^0 - \alpha_l^0| - |(\hat{\alpha}_k - \hat{\alpha}_l) - (\alpha_k^0 - \alpha_l^0)| \leq 2\tau) \\
\leq & K^0(K^0-1)\Phi\left(-2^{-1}\sigma^{-1}(\gamma_{\min} - 2\tau)n^{1/2}\lambda_{\min}^{1/2}(n^{-1}Z_{\mathcal{G}_0^{0c}}^\top Z_{\mathcal{G}_0^{0c}})\right) \\
= & O\left(\frac{1}{n(\log n)^{1/2}}\right). \qquad \text{(A.21)}
\end{aligned}$$

For $\mathcal{J}_{22}^c$, by the assumptions (A1)-(A2), $(X_A\mathbf{1})^\top(\boldsymbol{y} - X^\top\hat{\boldsymbol{\beta}}^{ols}) = (X_A\mathbf{1})^\top(I - P_{Z_{\mathcal{G}_0^{0c}}})\boldsymbol{\epsilon} \sim N(0, \sigma^2\|(I - P_{Z_{\mathcal{G}_0^{0c}}})X_A\mathbf{1}\|^2)$, and $\|(I - P_{Z_{\mathcal{G}_0^{0c}}})X_A\mathbf{1}\|^2 \leq \|X_A\mathbf{1}\|^2$. Denote $\mathcal{D} = \max\limits_{k,A\subset\mathcal{G}_k^0}\|X_A\mathbf{1}\|/|\varepsilon \cap \{A \times (\mathcal{G}_k^0 \setminus A)\}|$. If $(2^{-1}n\lambda_2\tau^{-1}\sigma^{-1}/\mathcal{D})^2 \geq 2\log\{2n|\mathcal{N}|/(2\pi)^{1/2}\}$, then

$$P(\mathcal{J}_{22}^c) \leq \sum_{k=1,\ldots,K^0;A\subset\mathcal{G}_k^0} P\left(\left|(X_A\mathbf{1})^\top(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}^{ols})\right| > n\frac{\lambda_2}{\tau}\left|\varepsilon \cap \{A \times (\mathcal{G}_k^0 \setminus A)\}\right|\right)$$

$$\leq 2|\mathcal{N}|\Phi\left(-\frac{n\lambda_2/\tau}{2\sigma\mathcal{D}}\right) = O\left(\frac{1}{n(\log n)^{1/2}}\right). \qquad \text{(A.22)}$$

By (A.18)-(A.22), we thus have

$$P(\hat{\boldsymbol{\beta}} \neq \hat{\boldsymbol{\beta}}^{ols}) = O\left(\frac{1}{n(\log n)^{1/2}}\right),$$

which, together with Lemma 2.1, yields that

$$P(\hat{\boldsymbol{\beta}} \neq \hat{\boldsymbol{\beta}}^{ora}) \leq P(\hat{\boldsymbol{\beta}} \neq \hat{\boldsymbol{\beta}}^{ols}) + P(\hat{\boldsymbol{\beta}}^{ora} \neq \hat{\boldsymbol{\beta}}^{ols}) = O\left(\frac{1}{n(\log n)^{1/2}}\right).$$

(2) Note that, $\hat{\boldsymbol{\alpha}}$ satisfies that

$$-Z_{\mathcal{G}_0^c}^{\top}\left(\boldsymbol{y} - Z_{\mathcal{G}_0^c}\hat{\boldsymbol{\alpha}}\right) + 2n\lambda_3 M_0\hat{\boldsymbol{\alpha}} + n\hat{\boldsymbol{\delta}} = 0,$$

where M_0 is a $K \times K$ diagonal matrix with diagonal elements $|\mathcal{G}_k| I_{\{\hat{\boldsymbol{\alpha}}_k < 0\}}$ for $k = 1, \ldots, K$; $\hat{\boldsymbol{\delta}} = (\hat{\delta}_1, \ldots, \hat{\delta}_K)^{\top}$, $\hat{\delta}_k = \sum_{j \in \mathcal{G}_k} \Upsilon_j(\hat{\boldsymbol{\beta}})$, and $\Upsilon_j(\boldsymbol{\beta}) = \lambda_1 \tau^{-1}\text{sign}(\beta_j) I_{\{|\beta_j| \leq \tau\}} + \lambda_2 \tau^{-1} \sum\limits_{j':(j',j)\in\varepsilon} \text{sign}(\beta_j - \beta_{j'}) I_{\{|\beta_j - \beta_{j'}| \leq \tau\}}$. Note that $\|\hat{\boldsymbol{\delta}}\|^2 \leq \tau^{-2}(\lambda_1 s^* + \lambda_2|\mathcal{N}|)^2$. We obtain that

$$\hat{\boldsymbol{\alpha}} = (Z_{\mathcal{G}_0^c}^{\top} Z_{\mathcal{G}_0^c} + 2n\lambda_3 M_0)^{-1}(Z_{\mathcal{G}_0^c}^{\top}\boldsymbol{y} - n\hat{\boldsymbol{\delta}}),$$

followed by

$$\begin{aligned}
&\|X\hat{\boldsymbol{\beta}} - X\boldsymbol{\beta}^0\|^2 \\
=&\|Z_{\mathcal{G}_0^c}(Z_{\mathcal{G}_0^c}^{\top} Z_{\mathcal{G}_0^c} + 2n\lambda_3 M_0)^{-1}(Z_{\mathcal{G}_0^c}^{\top}\boldsymbol{y} - n\hat{\boldsymbol{\delta}}) - Z_{\mathcal{G}_0^{0c}}\boldsymbol{\alpha}^0\|^2 \\
=&\|\{I - Z_{\mathcal{G}_0^c}(Z_{\mathcal{G}_0^c}^{\top} Z_{\mathcal{G}_0^c} + 2n\lambda_3 M_0)^{-1} Z_{\mathcal{G}_0^c}^{\top}\} Z_{\mathcal{G}_0^{0c}}\boldsymbol{\alpha}^0 - Z_{\mathcal{G}_0^c}(Z_{\mathcal{G}_0^c}^{\top} Z_{\mathcal{G}_0^c} + 2n\lambda_3 M_0)^{-1} Z_{\mathcal{G}_0^c}^{\top}\boldsymbol{\epsilon} \\
&+ nZ_{\mathcal{G}_0^c}(Z_{\mathcal{G}_0^c}^{\top} Z_{\mathcal{G}_0^c} + 2n\lambda_3 M_0)^{-1}\hat{\boldsymbol{\delta}})\|^2 \\
\leq&\|X\boldsymbol{\beta}^0\|^2 + \|\boldsymbol{\epsilon}\|^2 + \frac{\tau^2 n}{16} \min_{K(\mathcal{G}_0^c) \leq K^*} \lambda_{\min}\left(\frac{1}{n} Z_{\mathcal{G}_0^c}^{\top} Z_{\mathcal{G}_0^c}\right). \qquad \text{(A.23)}
\end{aligned}$$

Denote $T_1 = n^{-1}E(\|X\hat{\boldsymbol{\beta}} - X\boldsymbol{\beta}^0\|^2 I_{\{G\}})$, and $T_2 = n^{-1}E(\|X\hat{\boldsymbol{\beta}} - X\boldsymbol{\beta}^0\|^2 I_{\{G^c\}})$, where $G = \{n^{-1}\|X\hat{\boldsymbol{\beta}} - X\boldsymbol{\beta}^0\|^2 \geq D\}$. By the definition, we have

$$\frac{1}{n}E(\|X\hat{\boldsymbol{\beta}} - X\boldsymbol{\beta}^0\|^2) = T_1 + T_2.$$

Next, we work on T_1, T_2. Let

$$D = \frac{3}{n}\|X\boldsymbol{\beta}_0\|^2 + 10\sigma^2 + \frac{3\tau^2}{16}\min_{K(\mathcal{G}_0^c)\leq K^*}\lambda_{\min}\left(\frac{1}{n}Z_{\mathcal{G}_0^c}^\top Z_{\mathcal{G}_0^c}\right). \tag{A.24}$$

For T_1, it follows that

$$\begin{aligned}\int_D^\infty P\left(n^{-1}\|X\hat{\boldsymbol{\beta}} - X\boldsymbol{\beta}^0\|^2 \geq x\right)dx &\leq \int_{10\sigma^2}^\infty P\left(3n^{-1}\|\boldsymbol{\epsilon}\|^2 \geq x\right)dx \\ &\leq \int_{10\sigma^2}^\infty E\left\{\exp\left(\frac{t\|\boldsymbol{\epsilon}\|^2}{\sigma^2}\right)\exp\left(-\frac{ntx}{3\sigma^2}\right)\right\}dx \\ &\leq \int_{10\sigma^2}^\infty \exp\left(-\frac{n}{9\sigma^2}(x-9\sigma^2)\right)dx \\ &\leq \frac{9\sigma^2}{n}\exp\left(-\frac{n}{9}\right) = o\left(\frac{K^0\sigma^2}{n}\right). \end{aligned}\tag{A.25}$$

By (A.23) and (A.24), thus the first '$\leq$' follows. In view of the moment generating function for Chi-squared distribution, taking $t = 1/3$, the third '$\leq$' holds. For T_2,

$$T_2 = E\left(\frac{1}{n}\left\|X\hat{\boldsymbol{\beta}} - X\boldsymbol{\beta}^0\right\|^2 I_{\{G^c\}}I_{\{\hat{\boldsymbol{\beta}}\neq\hat{\boldsymbol{\beta}}^{ols}\}}\right) + E\left(\frac{1}{n}\left\|X\hat{\boldsymbol{\beta}} - X\boldsymbol{\beta}^0\right\|^2\left(1 - I_{\{G\}}\right)I_{\{\hat{\boldsymbol{\beta}}=\hat{\boldsymbol{\beta}}^{ols}\}}\right). \tag{A.26}$$

For the first term in (A.26), if $D = o\{K^0(\log n)^{1/2}\}$, then

$$\begin{aligned}
&E\left(\frac{1}{n}\left\|X\hat{\boldsymbol{\beta}} - X\boldsymbol{\beta}^0\right\|^2 I_{\{G^c\}} I_{\{\hat{\boldsymbol{\beta}} \neq \hat{\boldsymbol{\beta}}^{ols}\}}\right) \\
&\leq DP\left(\hat{\boldsymbol{\beta}} \neq \hat{\boldsymbol{\beta}}^{ols}\right) \leq DP\left(\hat{\boldsymbol{\beta}} \neq \hat{\boldsymbol{\beta}}^{ora}\right) + DP\left(\hat{\boldsymbol{\beta}}^{ora} \neq \hat{\boldsymbol{\beta}}^{ols}\right) \\
&= DO\left(\frac{1}{n(\log n)^{1/2}}\right) = o\left(\frac{K^0\sigma^2}{n}\right). \qquad \text{(A.27)}
\end{aligned}$$

For the second term in (A.26),

$$E\left(\frac{1}{n}\left\|X\hat{\boldsymbol{\beta}} - X\boldsymbol{\beta}^0\right\|^2 I_{\{G\}} I_{\{\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{ols}\}}\right) \leq E\left(\frac{1}{n}\left\|X\hat{\boldsymbol{\beta}} - X\boldsymbol{\beta}^0\right\|^2 I_{\{G\}}\right) = o\left(\frac{K^0\sigma^2}{n}\right), \qquad \text{(A.28)}$$

and

$$E\left(\frac{1}{n}\left\|X\hat{\boldsymbol{\beta}} - X\boldsymbol{\beta}^0\right\|^2 I_{\{\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{ols}\}}\right) \leq E\left(\frac{1}{n}\left\|X\hat{\boldsymbol{\beta}}^{ols} - X\boldsymbol{\beta}^0\right\|^2\right) = \frac{K^0\sigma^2}{n}. \qquad \text{(A.29)}$$

By (A.25), (A.26)-(A.29),

$$\frac{1}{n}E\left(\left\|X\hat{\boldsymbol{\beta}} - X\boldsymbol{\beta}^0\right\|^2\right) = T_1 + T_2 = \frac{K^0\sigma^2}{n}(1 + o(1)).$$

B Appendix

B.1 Proofs of these propositions and the main results

This part contains the propositions and proofs of those theorems. If not specified, we present the results below under the assumptions (A1)-(A5) in Chapter 3. We remark that these proofs are mainly in light of El Karoui (2013, 2018). However, our work overcomes the proof challenges under the vector framework of $\boldsymbol{y}_i, \boldsymbol{\psi}$ and the matrix framework of $X_i, \nabla\boldsymbol{\psi}$, which are not straightforward extensions of El Karoui (2013, 2018). We also remark that if a proof can be given by simply following a proof in El Karoui (2013, 2018), it will be omitted. These proofs are all under those assumptions given in Chapter 3.

Propositions B.1-B.2 are needed in the proof of Theorem 3.2.

Proposition B.1 *We have*

$$\|\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}_i\| \leq \frac{1}{\tau}\|\boldsymbol{\xi}_i\|, \tag{B.1}$$

and

$$\|\boldsymbol{\xi}_i\| = O_{L_k}\left(\frac{polylog(n)}{n}\right), \tag{B.2}$$

where $\boldsymbol{\xi}_i = n^{-1}\sum_{j\neq i} X_j^\top(\nabla\boldsymbol{\psi}(\boldsymbol{r}^*_{1,j}) - \nabla\boldsymbol{\psi}(\tilde{\boldsymbol{r}}_{j,[-i]}))X_j\boldsymbol{\eta}_i$*,* $\boldsymbol{r}^*_{1,j} \in (\boldsymbol{e}_j + X_j\boldsymbol{\beta}_0 - X_j\hat{\boldsymbol{\beta}}_{(i)}, \boldsymbol{e}_j + X_j\boldsymbol{\beta}_0 - X_j(\hat{\boldsymbol{\beta}}_{(i)} + \boldsymbol{\eta}_i))$*,* $\boldsymbol{\eta}_i = n^{-1}(S_i + \tau I_p)^{-1}X_i^\top\boldsymbol{\psi}(\tilde{\boldsymbol{g}}^{-1}(\tilde{\boldsymbol{r}}_{i,[-i]}))$*.*

Proof of Proposition B.1 By Lemma B.1, $\|\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}_i\| \leq \tau^{-1}\|\boldsymbol{\phi}(\tilde{\boldsymbol{\beta}}_i)\|$. It is easy to see that (B.1) and (B.2) follow, if we can prove that $\boldsymbol{\phi}(\tilde{\boldsymbol{\beta}}_i) = \boldsymbol{\xi}_i$, and $\|\boldsymbol{\xi}_i\| = O_{L_k}((n)/n)$.

Note that $\boldsymbol{y}_i = \boldsymbol{e}_i + X_i\boldsymbol{\beta}_0$, and $\tilde{\boldsymbol{\beta}}_i = \hat{\boldsymbol{\beta}}_{(i)} + \boldsymbol{\eta}_i$. Then

$$\begin{aligned}\boldsymbol{\phi}(\tilde{\boldsymbol{\beta}}_i) =& -\frac{1}{n}X_i^\top\boldsymbol{\psi}(\boldsymbol{y}_i - X_i\tilde{\boldsymbol{\beta}}_i) + \frac{1}{n}\sum_{j\neq i} X_j^\top\left[\boldsymbol{\psi}(\boldsymbol{y}_j - X_j\hat{\boldsymbol{\beta}}_{(i)}) - \boldsymbol{\psi}(\boldsymbol{y}_j - X_j(\hat{\boldsymbol{\beta}}_{(i)} + \boldsymbol{\eta}_i))\right] + \tau\boldsymbol{\eta}_i \\ =& -\frac{1}{n}X_i^\top\boldsymbol{\psi}(\boldsymbol{y}_i - X_i\tilde{\boldsymbol{\beta}}_i) + \frac{1}{n}\sum_{j\neq i} X_j^\top\nabla\boldsymbol{\psi}(\tilde{\boldsymbol{r}}_{j,[-i]})X_j\boldsymbol{\eta}_i + \tau\boldsymbol{\eta}_i + \boldsymbol{\xi}_i, \qquad \text{(B.3)}\end{aligned}$$

where $\boldsymbol{r}^*_{1,j} \in (\boldsymbol{y}_j - X_j\hat{\boldsymbol{\beta}}_{(i)}, \boldsymbol{y}_j - X_j(\hat{\boldsymbol{\beta}}_{(i)} + \boldsymbol{\eta}_i))$. Since $(n^{-1}\sum_{j\neq i} X_j^\top\nabla\boldsymbol{\psi}(\tilde{\boldsymbol{r}}_{j,[-i]})X_j + \tau I_p)\boldsymbol{\eta}_i = n^{-1}X_i^\top\boldsymbol{\psi}(\tilde{\boldsymbol{g}}^{-1}(\tilde{\boldsymbol{r}}_{i,[-i]}))$, by the definition of $\tilde{\boldsymbol{g}}$, we have

$$\boldsymbol{y}_i - X_i\tilde{\boldsymbol{\beta}}_i = \boldsymbol{y}_i - X_i\hat{\boldsymbol{\beta}}_{(i)} - X_i\boldsymbol{\eta}_i = \tilde{\boldsymbol{r}}_{i,[-i]} - C_i\boldsymbol{\psi}(\tilde{\boldsymbol{g}}^{-1}(\tilde{\boldsymbol{r}}_{i,[-i]})) = \tilde{\boldsymbol{g}}^{-1}(\tilde{\boldsymbol{r}}_{i,[-i]}).$$

Therefore,

$$-\frac{1}{n}X_i^\top\boldsymbol{\psi}(\boldsymbol{y}_i - X_i\tilde{\boldsymbol{\beta}}_i) + \frac{1}{n}\sum_{j\neq i} X_j^\top\nabla\boldsymbol{\psi}(\tilde{\boldsymbol{r}}_{j,[-i]})X_j\boldsymbol{\eta}_i + \tau\boldsymbol{\eta}_i = \mathbf{0},$$

which, together with (B.3), implies that $\boldsymbol{\phi}(\tilde{\boldsymbol{\beta}}_i) = \boldsymbol{\xi}_i$.

It remains to show that $\|\boldsymbol{\xi}_i\| = O_{L_k}(\text{polyLog}(n)/n)$, the conclusion of which is obviously implied by the following two results,

$$\|\boldsymbol{\eta}_i\| \leq O_{L_k}\left(\frac{1}{\sqrt{n}}\right), \tag{B.4}$$

and

$$\|n^{-1}\sum_{j\neq i} X_j^\top(\nabla\boldsymbol{\psi}(\boldsymbol{r}^*_{1,j}) - \nabla\boldsymbol{\psi}(\tilde{\boldsymbol{r}}_{j,[-i]}))X_j\|_{\max} \leq O_{L_k}\left(\frac{\text{polyLog}(n)}{\sqrt{n}}\right). \tag{B.5}$$

For $\boldsymbol{\eta}_i = n^{-1}(S_i + \tau I_p)^{-1}X_i^\top\boldsymbol{\psi}(\tilde{\boldsymbol{g}}^{-1}(\tilde{\boldsymbol{r}}_{i,[-i]}))$, we have

$$\begin{aligned}
\|\boldsymbol{\eta}_i\| =& \|\frac{1}{n}(S_i + \tau I_p)^{-1}X_i^\top\boldsymbol{\psi}(\tilde{\boldsymbol{g}}^{-1}(\tilde{\boldsymbol{r}}_{i,[-i]}))\| \\
\leq& \frac{1}{n}\lambda_{\max}((S_i + \tau I_p)^{-1})\|X_i^\top\|_{\max}\|\boldsymbol{\psi}(\tilde{\boldsymbol{g}}^{-1}(\tilde{\boldsymbol{r}}_{i,[-i]}))\| \\
\leq& \frac{1}{n\tau}\sqrt{\text{tr}(X_iX_i^\top)}\sup\|\boldsymbol{\psi}\| \\
=& O_{L_k}\left(\frac{\sqrt{p}}{n}\right) = O_{L_k}\left(\frac{1}{\sqrt{n}}\right).
\end{aligned}$$

On the other hand,

$$\begin{aligned}
&\|n^{-1}\sum_{j\neq i} X_j^\top(\nabla\boldsymbol{\psi}(\boldsymbol{r}^*_{1,j}) - \nabla\boldsymbol{\psi}(\tilde{\boldsymbol{r}}_{j,[-i]}))X_j\|_{\max} \\
\leq& \lambda_{\max}(n^{-1}\sum_i X_i^\top X_i)\sup_{j\neq i}\|\nabla\boldsymbol{\psi}(\boldsymbol{r}^*_{1,j}) - \nabla\boldsymbol{\psi}(\tilde{\boldsymbol{r}}_{j,[-i]})\|_{\max} \\
\leq& c\lambda_{\max}(n^{-1}\sum_i X_i^\top X_i)\sup_{j\neq i}\|\boldsymbol{r}^*_{1,j} - \tilde{\boldsymbol{r}}_{j,[-i]}\| \\
\leq& O_{L_k}(\text{polyLog}(n))\sup_{j\neq i}\|X_j\boldsymbol{\eta}_i\|.
\end{aligned} \tag{B.6}$$

The second inequality follows by the assumption (A3). The third inequality holds because $\lambda_{\max}(n^{-1}\sum_i X_i^\top X_i) = O_{L_k}(\text{polyLog}(n))$. Now, we work on that $X_j\boldsymbol{\eta}_i = n^{-1}X_j(S_i+\tau I_p)^{-1}X_i^\top\boldsymbol{\psi}(\tilde{\boldsymbol{g}}^{-1}(\tilde{\boldsymbol{r}}_{i,[-i]}))$, and we have

$$\begin{aligned}\sup_{j\neq i}\|X_j\boldsymbol{\eta}_i\| \leq& \sup_{j\neq i}\lambda_{\max}(n^{-1}X_i(S_i+\tau I_p)^{-1}X_j^\top)\sup\|\boldsymbol{\psi}\|\\ \leq& c\sup_{j\neq i}\text{tr}(n^{-1}X_i(S_i+\tau I_p)^{-1}X_j^\top). \qquad \text{(B.7)}\end{aligned}$$

Denote $X_i = (\boldsymbol{x}_i(1),\ldots,\boldsymbol{x}_i(m))^\top$, $X_j = (\boldsymbol{x}_j(1),\ldots,\boldsymbol{x}_j(m))^\top$, where X_i are independent of S_i and X_j. Thus we have

$$\sup_{j\neq i}\text{tr}(n^{-1}X_i(S_i+\tau I_p)^{-1}X_j^\top) \leq \sum_{t=1}^m \sup_{j\neq i} n^{-1}\boldsymbol{x}_i^\top(t)(S_i+\tau I_p)^{-1}\boldsymbol{x}_j(t).$$

Define $X_{(-i)} = \{X_1,\ldots,X_{i-1},X_{i+1},\ldots,X_n\}$, $\boldsymbol{v}_{j,(i)} = (S_i+\tau I_p)^{-1}\boldsymbol{x}_j(t)$, and $f_1(\boldsymbol{x}_i(t)) = \boldsymbol{x}_i^\top(t)(S_i+\tau I_p)^{-1}\boldsymbol{x}_j(t) = \boldsymbol{x}_i^\top(t)\boldsymbol{v}_{j,(i)}$, with the Lipschitz constant satisfying that $\|\boldsymbol{v}_{j,(i)}\| = \sqrt{\boldsymbol{x}_j^\top(t)(S_i+\tau I_p)^{-2}\boldsymbol{x}_j(t)} \leq \|\boldsymbol{x}_j(t)\|/\tau$. By Lemma 3.36 in El Karoui (2018), it follows that

$$\frac{1}{mn}\sup_{j\neq i,1\leq t\leq m}\boldsymbol{x}_i^\top(t)(S_i+\tau I_p)^{-1}\boldsymbol{x}_j(t) \leq \sup_{j\neq i,1\leq t\leq m}\frac{\|\boldsymbol{x}_j(t)\|}{mn\tau}\sqrt{\text{polyLog}(n)/c_n}+\sup_j|m_{f_1}|.$$

Since $m_{f_1}=0$, $\sup_{j,t}\|\boldsymbol{x}_j(t)\| = O_{L_k}(\sqrt{p})$, $1/c_n = O(\text{polyLog}(n))$ and m is fixed, we have

$$\sup_{j\neq i}\text{tr}(n^{-1}X_i(S_i+\tau I_p)^{-1}X_j^\top) = O_{L_k}(n^{-1/2}\text{polyLog}(n)),$$

which, together with (B.6) and (B.7), completes the proof of (B.5).

Proposition B.2 *(i) Denote $Q_i = C_i - c_i I_m$ and let $q_i^{(s,t)}$ be the (s,t)-th entry of Q_i, $s, t = 1, \ldots, m$. Then,*

$$\sup_{i,s,t} |q_i^{(s,t)}| = O_{L_k}(n^{-1/2}\text{polyLog}(n)). \tag{B.8}$$

(ii) Denote $c_\tau = n^{-1}\text{tr}((S + \tau I_p)^{-1})$, $N_i = C_i - c_\tau I_m$. Define the (s,t)-th entry of N_i by $\nu_i^{(s,t)}$, $1 \leq s, t \leq m$. Then,

$$\sup_{i,s,t} |\nu_i^{(s,t)}| = O_{L_k}(n^{-1/2}\text{polyLog}(n)). \tag{B.9}$$

Proof of Proposition B.2 (i) To prove (B.8), we consider the following two cases. Recall that $C_i = n^{-1}X_i(S_i + \tau I_p)^{-1}X_i^\top, c_i = \text{tr}(n^{-1}(S_i + \tau I_p)^{-1})$.

(a) $s \neq t$.

By Lemma 3.36 in El Karoui (2018),

$$\begin{aligned}\sup_{1\leq i\leq n, 1\leq s\neq t\leq m} \frac{1}{n}|\boldsymbol{x}_i^\top(s)(S_i + \tau I_p)^{-1}\boldsymbol{x}_i(t)| =& \frac{1}{\sqrt{n}\tau} \sup \frac{\|\boldsymbol{x}_i(t)\|}{\sqrt{n}}\text{polyLog}(n) \\ =& O_{L_k}(n^{-1/2}\text{polyLog}(n)).\end{aligned}$$

(b) $s = t$.

By Lemma 3.37 in El Karoui (2018),

$$\sup_{1\leq i\leq n, 1\leq s\leq m} |\frac{1}{n}\boldsymbol{x}_i^\top(s)(S_i + \tau I_p)^{-1}\boldsymbol{x}_i(s) - c_i| = O_{L_k}(n^{-1/2}\text{polyLog}(n)).$$

This completes the proof of (B.8).

We now prove (ii). Note that $\nu_i^{(s,t)} = n^{-1}\boldsymbol{x}_i^\top(s)(S_i + \tau I_p)^{-1}\boldsymbol{x}_i(t) - n^{-1}\mathrm{tr}((S + \tau I_p)^{-1})$. Since (B.48) and (B.49) still hold true with $\boldsymbol{x}_i^\top(s), S_i, I_p, \boldsymbol{x}_i^\top(t), S$ respectively replacing $\boldsymbol{x}_{i,-p}^\top(s), \Delta_p(i), I_{p-1}, \boldsymbol{x}_{i,-p}(t), \Lambda_{i,p}$, we have the desired result. One can mimic the proofs of (B.48) and (B.49) in Lemma B.5 to show (B.9). We thus omit the details.

Proof of Theorem 3.2

(i) The results can be obtained by simply following the proof of Theorem 3.9 given by El Karoui (2018).

(ii) By Proposition B.1 and the assumption (A4), it is easy to obtain that

$$\begin{aligned}
\sup_{j\neq i}\|\tilde{\boldsymbol{r}}_{j,[-i]} - \boldsymbol{r}_j\| =& \sup_{j\neq i}\|X_j(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})\| \\
\leq& \sup_{j\neq i}\|X_j(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}_i)\| + \sup_{j\neq i}\|X_j(\tilde{\boldsymbol{\beta}}_i - \hat{\boldsymbol{\beta}}_{(i)})\| \\
\leq& \sup_{j\neq i}\frac{m\|\boldsymbol{x}_j(1)\|}{\sqrt{n}}\sqrt{n}\|(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}_i)\| + \sup_{j\neq i}\|X_j\boldsymbol{\eta}_i\| \\
=& O_{L_k}\left(n^{-1/2}\mathrm{polyLog}(n)\right) + O_{L_k}\left(n^{-1/2}\mathrm{polyLog}(n)\right) \\
=& O_{L_k}\left(n^{-1/2}\mathrm{polyLog}(n)\right).
\end{aligned}$$

The second '=' holds in view of the bounded supports of $\sup_{1\leq i\leq n}\|\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}_i\|, \sup_{j\neq i}\|X_j\boldsymbol{\eta}_i\|$ (see the details on the proof of Proposition B.1), and $\|\boldsymbol{x}_j(1)\| = O_{L_k}(\sqrt{p})$.

Next, we prove that $\sup_i\|\boldsymbol{r}_i - \tilde{\boldsymbol{g}}^{-1}(\tilde{\boldsymbol{r}}_{i,[-i]})\| = O_{L_k}\left(n^{-1/2}\mathrm{polyLog}(n)\right)$. Since $\tilde{\boldsymbol{\beta}}_i - \hat{\boldsymbol{\beta}}_{(i)} = \boldsymbol{\eta}_i$, $\boldsymbol{\eta}_i = n^{-1}(S_i + \tau I_p)^{-1}X_i^\top\boldsymbol{\psi}(\tilde{\boldsymbol{g}}^{-1}(\tilde{\boldsymbol{r}}_{i,[-i]}))$, $C_i = n^{-1}X_i(S_i + \tau I_p)^{-1}X_i^\top$, we have

$$X_i\tilde{\boldsymbol{\beta}}_i = X_i\hat{\boldsymbol{\beta}}_{(i)} + C_i\boldsymbol{\psi}(\tilde{\boldsymbol{g}}^{-1}(\tilde{\boldsymbol{r}}_{i,[-i]})),$$

which, together with the definition of proximal mapping function $\tilde{\boldsymbol{g}}^{-1}$, yields that

$$\boldsymbol{e}_i + X_i\boldsymbol{\beta}_0 - X_i\tilde{\boldsymbol{\beta}}_i = \tilde{\boldsymbol{r}}_{i,[-i]} - C_i\boldsymbol{\psi}(\tilde{\boldsymbol{g}}^{-1}(\tilde{\boldsymbol{r}}_{i,[-i]})) = \tilde{\boldsymbol{g}}^{-1}(\tilde{\boldsymbol{r}}_{i,[-i]}). \tag{B.10}$$

Note that

$$\boldsymbol{r}_i = \boldsymbol{e}_i + X_i\boldsymbol{\beta}_0 - X_i\hat{\boldsymbol{\beta}} = \boldsymbol{e}_i + X_i\boldsymbol{\beta}_0 - X_i\tilde{\boldsymbol{\beta}}_i - X_i(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}_i). \tag{B.11}$$

By (B.10) and (B.11), it follows that $\boldsymbol{r}_i - \tilde{\boldsymbol{g}}^{-1}(\tilde{\boldsymbol{r}}_{i,[-i]}) = -X_i(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}_i)$, and

$$\sup_i \|\boldsymbol{r}_i - \tilde{\boldsymbol{g}}^{-1}(\tilde{\boldsymbol{r}}_{i,[-i]})\| = \sup_i \|X_i(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}_i)\| = O_{L_k}\left(n^{-1/2}\text{polyLog}(n)\right). \tag{B.12}$$

Finally, we show that $\sup_i \|\boldsymbol{r}_i - \text{prox}_{c_i}(\rho)(\tilde{\boldsymbol{r}}_{i,[-i]})\| = O_{L_k}(n^{-1/2}\text{polyLog}(n))$. By (B.10), $\tilde{\boldsymbol{r}}_{i,[-i]} = \tilde{\boldsymbol{g}}^{-1}(\tilde{\boldsymbol{r}}_{i,[-i]}) + C_i\boldsymbol{\psi}(\tilde{\boldsymbol{g}}^{-1}(\tilde{\boldsymbol{r}}_{i,[-i]}))$. Using the definition of prox function, $\tilde{\boldsymbol{r}}_{i,[-i]} = \text{prox}_{c_i}(\rho)(\tilde{\boldsymbol{r}}_{i,[-i]}) + c_i\boldsymbol{\psi}(\text{prox}_{c_i}(\rho)(\tilde{\boldsymbol{r}}_{i,[-i]}))$. One can easily obtain that

$$\begin{aligned}
&\tilde{\boldsymbol{g}}^{-1}(\tilde{\boldsymbol{r}}_{i,[-i]}) - \text{prox}_{c_i}(\rho)(\tilde{\boldsymbol{r}}_{i,[-i]})\\
=&c_i\boldsymbol{\psi}(\text{prox}_{c_i}(\rho)(\tilde{\boldsymbol{r}}_{i,[-i]})) - C_i\boldsymbol{\psi}(\tilde{\boldsymbol{g}}^{-1}(\tilde{\boldsymbol{r}}_{i,[-i]}))\\
=&c_i[\boldsymbol{\psi}(\text{prox}_{c_i}(\rho)(\tilde{\boldsymbol{r}}_{i,[-i]})) - \boldsymbol{\psi}(\tilde{\boldsymbol{g}}^{-1}(\tilde{\boldsymbol{r}}_{i,[-i]}))] + (c_iI_m - C_i)\boldsymbol{\psi}(\tilde{\boldsymbol{g}}^{-1}(\tilde{\boldsymbol{r}}_{i,[-i]})).
\end{aligned} \tag{B.13}$$

Because $\boldsymbol{\psi}(\tilde{\boldsymbol{g}}^{-1}(\tilde{\boldsymbol{r}}_{i,[-i]})) - \boldsymbol{\psi}(\text{prox}_{c_i}(\rho)(\tilde{\boldsymbol{r}}_{i,[-i]})) = \nabla\boldsymbol{\psi}(\boldsymbol{r}^*_{g,p})(\tilde{\boldsymbol{g}}^{-1}(\tilde{\boldsymbol{r}}_{i,[-i]}) - \text{prox}_{c_i}(\rho)(\tilde{\boldsymbol{r}}_{i,[-i]}))$, where $\boldsymbol{r}^*_{g,p} \in (\tilde{\boldsymbol{g}}^{-1}(\tilde{\boldsymbol{r}}_{i,[-i]}), \text{prox}_{c_i}(\rho)(\tilde{\boldsymbol{r}}_{i,[-i]}))$, (B.13) becomes

$$(I_m + c_i\nabla\boldsymbol{\psi}(\boldsymbol{r}^*_{g,p}))(\tilde{\boldsymbol{g}}^{-1}(\tilde{\boldsymbol{r}}_{i,[-i]}) - \text{prox}_{c_i}(\rho)(\tilde{\boldsymbol{r}}_{i,[-i]})) = (c_iI_m - C_i)\boldsymbol{\psi}(\tilde{\boldsymbol{g}}^{-1}(\tilde{\boldsymbol{r}}_{i,[-i]})).$$

By (B.8) and the assumption (A3),

$$\|\tilde{\boldsymbol{g}}^{-1}(\tilde{\boldsymbol{r}}_{i,[-i]}) - \text{prox}_{c_i}(\rho)(\tilde{\boldsymbol{r}}_{i,[-i]})\| = O_{L_k}(n^{-1/2}\text{polyLog}(n)). \tag{B.14}$$

By (B.12) and (B.14), it holds that

$$\sup_i \|\boldsymbol{r}_i - \text{prox}_{c_i}(\rho)(\tilde{\boldsymbol{r}}_{i,[-i]})\| = O_{L_k}(n^{-1/2}\text{polyLog}(n)).$$

(iii) The proof is analogous to the proof strategy of Proposition 3.10 given in El Karoui (2018). However, we give partial details under our model framework.

By Efron-Stein inequality, $\text{var}(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2) = O(n^{-1}\text{polyLog}(n))$ follows, if we can show that

$$E(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2 - \|\tilde{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_0\|^2)^2 = O(n^{-2}\text{polyLog}(n)), \tag{B.15}$$

$$E(\|\tilde{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_0\|^2 - \|\hat{\boldsymbol{\beta}}_{(i)} - \boldsymbol{\beta}_0\|^2)^2 = O(n^{-2}\text{polyLog}(n)). \tag{B.16}$$

By following the proof of Proposition 3.10 in El Karoui (2018), it is easy to show (B.15). For (B.16), we have

$$\begin{aligned}
&\|\tilde{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_0\|^2 - \|\hat{\boldsymbol{\beta}}_{(i)} - \boldsymbol{\beta}_0\|^2 \\
=&\|\tilde{\boldsymbol{\beta}}_i - \hat{\boldsymbol{\beta}}_{(i)} + \hat{\boldsymbol{\beta}}_{(i)} - \boldsymbol{\beta}_0\|^2 - \|\hat{\boldsymbol{\beta}}_{(i)} - \boldsymbol{\beta}_0\|^2 \\
=&\frac{2}{n}(\hat{\boldsymbol{\beta}}_{(i)} - \boldsymbol{\beta}_0)^\top (S_i + \tau I_p)^{-1} X_i^\top \boldsymbol{\psi}(\tilde{\boldsymbol{g}}^{-1}(\tilde{\boldsymbol{r}}_{i,[-i]})) \\
&+ \frac{1}{n^2}\boldsymbol{\psi}^\top(\tilde{\boldsymbol{g}}^{-1}(\tilde{\boldsymbol{r}}_{i,[-i]}))X_i(S_i + \tau I_p)^{-2} X_i^\top \boldsymbol{\psi}(\tilde{\boldsymbol{g}}^{-1}(\tilde{\boldsymbol{r}}_{i,[-i]})).
\end{aligned}$$

Since $\hat{\boldsymbol{\beta}}_{(i)} - \boldsymbol{\beta}_0$ and S_i are independent of X_i, one can show that $\|(\hat{\boldsymbol{\beta}}_{(i)} - \boldsymbol{\beta}_0)^\top (S_i + \tau I_p)^{-1} X_i^\top\| = O_{L_2}(\|\hat{\boldsymbol{\beta}}_{(i)} - \boldsymbol{\beta}_0\| \text{polyLog}(n))$ by using the same technique as in the proof of Proposition B.1. For the second term, $\lambda_{\max}(n^{-1} X_i (S_i + \tau I_p)^{-2} X_i^\top) \leq \text{tr}(n^{-1} X_i (S_i + \tau I_p)^{-2} X_i^\top) = O_{L_2}(mp/n)$. Based on the above discussions, (B.16) thus follows. This completes the proof.

Propositions B.3-B.5 are needed in the proof of Theorem 3.3.

Proposition B.3 *We have*

$$\|\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{b}}\| \leq \frac{1}{\tau} \|\frac{1}{n} \sum_i X_i^\top (\nabla \boldsymbol{\psi}(\boldsymbol{r}^*_{2,i}) - \nabla \boldsymbol{\psi}(\check{\boldsymbol{r}}_{i,-p})) X_i (\hat{\gamma}_{est} - \tilde{\boldsymbol{b}})\|, \tag{B.17}$$

where $\boldsymbol{r}^*_{2,i} \in (\boldsymbol{e}_i + X_i \boldsymbol{\beta}_0 - X_i \tilde{\boldsymbol{b}}, \boldsymbol{e}_i + X_i \boldsymbol{\beta}_0 - X_i \hat{\boldsymbol{\gamma}}_{est})$, $\hat{\boldsymbol{\gamma}}_{est} - \tilde{\boldsymbol{b}} = (\tilde{b}_p - \beta_{0,p})(((\Delta_p + \tau I_{p-1})^{-1} \boldsymbol{u}_p)^\top, -1)^\top$, *and*

$$\|(\Delta_p + \tau I_{p-1})^{-1} \boldsymbol{u}_p\|^2 \leq \frac{1}{n\tau} \sum_{i=1}^n \boldsymbol{x}_{i,p}^\top \nabla \boldsymbol{\psi}(\check{\boldsymbol{r}}_{i,-p}) \boldsymbol{x}_{i,p} = O_{L_k}(1). \tag{B.18}$$

Proof of Proposition B.3 By (B.39), taking $\boldsymbol{\beta}_2 = \hat{\boldsymbol{\beta}}$, $\boldsymbol{\beta}_1 = \tilde{\boldsymbol{b}}$, we obtain

$$\|\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{b}}\| \leq \frac{1}{\tau} \|\boldsymbol{\phi}(\tilde{\boldsymbol{b}})\|.$$

(B.17) thus follows if we can show that

$$\boldsymbol{\phi}(\tilde{\boldsymbol{b}}) = -\frac{1}{n} \sum_i X_i^\top (\nabla \boldsymbol{\psi}(\boldsymbol{r}^*_{2,i}) - \nabla \boldsymbol{\psi}(\check{\boldsymbol{r}}_{i,-p})) X_i (\hat{\boldsymbol{\gamma}}_{est} - \tilde{\boldsymbol{b}}). \tag{B.19}$$

Denote $\boldsymbol{\phi}(\tilde{\boldsymbol{b}}) = (\boldsymbol{\phi}_{-p}(\tilde{\boldsymbol{b}})^\top, \phi_p(\tilde{\boldsymbol{b}}))^\top$, and $\hat{\boldsymbol{\gamma}}_{est} = (\hat{\boldsymbol{\gamma}}^\top, \beta_{0,p})^\top$. Since $\tilde{\boldsymbol{b}} = ((\hat{\boldsymbol{\gamma}} - (\tilde{b}_p - \beta_{0,p})(\Delta_p + \tau I_{p-1})^{-1} \boldsymbol{u}_p)^\top, \tilde{b}_p)^\top = (\tilde{\boldsymbol{b}}_{-p}^\top, \tilde{b}_p)^\top$, $\Delta_p = n^{-1} \sum_i X_{i,-p}^\top \nabla \boldsymbol{\psi}(\check{\boldsymbol{r}}_{i,-p}) X_{i,-p}$, and

$\boldsymbol{u}_p = n^{-1}\sum_i X_{i,-p}^\top \nabla\boldsymbol{\psi}(\check{\boldsymbol{r}}_{i,-p})\boldsymbol{x}_{i,p}$, we have $\tilde{\boldsymbol{b}}_{-p} - \hat{\boldsymbol{\gamma}} = -(\tilde{b}_p - \boldsymbol{\beta}_{0,p})(\Delta_p + \tau I_{p-1})^{-1}\boldsymbol{u}_p$, and $\hat{\boldsymbol{\gamma}}_{est} - \tilde{\boldsymbol{b}} = (\tilde{b}_p - \beta_{0,p})(((\Delta_p + \tau I_{p-1})^{-1}\boldsymbol{u}_p)^\top, -1)^\top$. We first prove (B.20) and (B.21) which are needed to show (B.19). One can easily verify that

$$\begin{aligned}
&-\frac{1}{n}\sum_i X_{i,-p}^\top \nabla\boldsymbol{\psi}(\check{\boldsymbol{r}}_{i,-p})X_i(\hat{\boldsymbol{\gamma}}_{est} - \tilde{\boldsymbol{b}}) + \tau(\tilde{\boldsymbol{b}}_{-p} - \hat{\boldsymbol{\gamma}})\\
=&-(\tilde{b}_p - \beta_{0,p})\frac{1}{n}\sum_i X_{i,-p}^\top \nabla\boldsymbol{\psi}(\check{\boldsymbol{r}}_{i,-p})(X_{i,-p}, \boldsymbol{x}_{i,p})(((\Delta_p + \tau I_{p-1})^{-1}\boldsymbol{u}_p)^\top, -1)^\top\\
&-\tau(\tilde{b}_p - \beta_{0,p})(\Delta_p + \tau I_{p-1})^{-1}\boldsymbol{u}_p\\
=&(\tilde{b}_p - \beta_{0,p})(-\Delta_p(\Delta_p + \tau I_{p-1})^{-1}\boldsymbol{u}_p + \boldsymbol{u}_p - \tau(\Delta_p + \tau I_{p-1})^{-1}\boldsymbol{u}_p) = \boldsymbol{0}_{p-1}. \qquad \text{(B.20)}
\end{aligned}$$

On the other hand, since

$$\tilde{b}_p = \beta_{0,p}\frac{\xi_n}{\tau + \xi_n} + \frac{n^{-1}\sum_i \boldsymbol{x}_{i,p}^\top \boldsymbol{\psi}(\check{\boldsymbol{r}}_{i,-p})}{\tau + \xi_n},$$

we have

$$\begin{aligned}
&-\frac{1}{n}\sum_i \boldsymbol{x}_{i,p}^\top\left[\boldsymbol{\psi}(\check{\boldsymbol{r}}_{i,-p}) + \nabla\boldsymbol{\psi}(\check{\boldsymbol{r}}_{i,-p})X_i(\hat{\boldsymbol{\gamma}}_{est} - \tilde{\boldsymbol{b}})\right] + \tau\tilde{b}_p\\
=&-\frac{1}{n}\sum_i \boldsymbol{x}_{i,p}^\top\left[\boldsymbol{\psi}(\check{\boldsymbol{r}}_{i,-p}) + \nabla\boldsymbol{\psi}(\check{\boldsymbol{r}}_{i,-p})(X_{i,-p}, \boldsymbol{x}_{i,p})(\tilde{b}_p - \beta_{0,p})(((\Delta_p + \tau I_{p-1})^{-1}\boldsymbol{u}_p)^\top, -1)^\top\right] + \tau\tilde{b}_p\\
=&-\frac{1}{n}\sum_i \boldsymbol{x}_{i,p}^\top\boldsymbol{\psi}(\check{\boldsymbol{r}}_{i,-p}) + \tau\tilde{b}_p + (\tilde{b}_p - \beta_{0,p})\left[-\boldsymbol{u}_p^\top(\Delta_p + \tau I_{p-1})^{-1}\boldsymbol{u}_p + \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_{i,p}^\top\nabla\boldsymbol{\psi}(\check{\boldsymbol{r}}_{i,-p})\boldsymbol{x}_{i,p}\right]\\
=&0. \qquad \text{(B.21)}
\end{aligned}$$

For the first $p-1$ coordinates of $\boldsymbol{\phi}(\tilde{\boldsymbol{b}})$, i.e., $\boldsymbol{\phi}_{-p}(\tilde{\boldsymbol{b}})$, we have

$$\begin{aligned}
\boldsymbol{\phi}_{-p}(\tilde{\boldsymbol{b}}) =& -\frac{1}{n}\sum_i X_{i,-p}^\top \Big[\boldsymbol{\psi}(\boldsymbol{e}_i + X_i\boldsymbol{\beta}_0 - X_i\tilde{\boldsymbol{b}}) - \boldsymbol{\psi}(\boldsymbol{e}_i + X_{i,-p}\boldsymbol{\beta}_{-p} - X_{i,-p}\hat{\boldsymbol{\gamma}})\Big] + \tau(\tilde{\boldsymbol{b}}_{-p} - \hat{\boldsymbol{\gamma}}) \\
=& -\frac{1}{n}\sum_i X_{i,-p}^\top \Big[\boldsymbol{\psi}(\boldsymbol{e}_i + X_i\boldsymbol{\beta}_0 - X_i\tilde{\boldsymbol{b}}) - \boldsymbol{\psi}(\boldsymbol{e}_i + X_i\boldsymbol{\beta}_0 - X_i\hat{\boldsymbol{\gamma}}_{est})\Big] + \tau(\tilde{\boldsymbol{b}}_{-p} - \hat{\boldsymbol{\gamma}}) \\
=& -\frac{1}{n}\sum_i X_{i,-p}^\top \nabla\boldsymbol{\psi}(\boldsymbol{r}_{2,i}^*)X_i(\hat{\boldsymbol{\gamma}}_{est} - \tilde{\boldsymbol{b}}) + \tau(\tilde{\boldsymbol{b}}_{-p} - \hat{\boldsymbol{\gamma}}) \\
=& -\frac{1}{n}\sum_i X_{i,-p}^\top \nabla\boldsymbol{\psi}(\check{\boldsymbol{r}}_{i,-p})X_i(\hat{\boldsymbol{\gamma}}_{est} - \tilde{\boldsymbol{b}}) + \tau(\tilde{\boldsymbol{b}}_{-p} - \hat{\boldsymbol{\gamma}}) \\
& -\frac{1}{n}\sum_i X_{i,-p}^\top [\nabla\boldsymbol{\psi}(\boldsymbol{r}_{2,i}^*) - \nabla\boldsymbol{\psi}(\check{\boldsymbol{r}}_{i,-p}))]X_i(\hat{\boldsymbol{\gamma}}_{est} - \tilde{\boldsymbol{b}}) \\
=& -\frac{1}{n}\sum_i X_{i,-p}^\top [\nabla\boldsymbol{\psi}(\boldsymbol{r}_{2,i}^*) - \nabla\boldsymbol{\psi}(\check{\boldsymbol{r}}_{i,-p})]X_i(\hat{\boldsymbol{\gamma}}_{est} - \tilde{\boldsymbol{b}}), \qquad \text{(B.22)}
\end{aligned}$$

where $\boldsymbol{r}_{2,i}^* \in (\boldsymbol{e}_i + X_i\boldsymbol{\beta}_0 - X_i\tilde{\boldsymbol{b}}, \boldsymbol{e}_i + X_i\boldsymbol{\beta}_0 - X_i\hat{\boldsymbol{\gamma}}_{est})$. The last equality holds true in view of (B.20).

For the last coordinate of $\boldsymbol{\phi}(\tilde{\boldsymbol{b}})$, i.e., $\phi_p(\tilde{\boldsymbol{b}})$, we have

$$\begin{aligned}
\phi_p(\tilde{\boldsymbol{b}}) =& -\frac{1}{n}\sum_i \boldsymbol{x}_{i,p}^\top \boldsymbol{\psi}(\boldsymbol{e}_i + X_i\boldsymbol{\beta}_0 - X_i\tilde{\boldsymbol{b}}) + \tau\tilde{b}_p \\
=& -\frac{1}{n}\sum_i \boldsymbol{x}_{i,p}^\top \nabla\boldsymbol{\psi}(\boldsymbol{r}_{2,i}^*)X_i(\hat{\boldsymbol{\gamma}}_{est} - \tilde{\boldsymbol{b}}) + \tau\tilde{b}_p \\
=& -\frac{1}{n}\sum_i \boldsymbol{x}_{i,p}^\top \Big[\boldsymbol{\psi}(\check{\boldsymbol{r}}_{i,-p}) + \nabla\boldsymbol{\psi}(\check{\boldsymbol{r}}_{i,-p})X_i(\hat{\boldsymbol{\gamma}}_{est} - \tilde{\boldsymbol{b}})\Big] + \tau\tilde{b}_p \\
& -\frac{1}{n}\sum_i \boldsymbol{x}_{i,p}^\top [\nabla\boldsymbol{\psi}(\boldsymbol{r}_{2,i}^*) - \nabla\boldsymbol{\psi}(\check{\boldsymbol{r}}_{i,-p})]X_i(\hat{\boldsymbol{\gamma}}_{est} - \tilde{\boldsymbol{b}}) \\
=& -\frac{1}{n}\sum_i \boldsymbol{x}_{i,p}^\top [\nabla\boldsymbol{\psi}(\boldsymbol{r}_{2,i}^*) - \nabla\boldsymbol{\psi}(\check{\boldsymbol{r}}_{i,-p}))]X_i(\hat{\boldsymbol{\gamma}}_{est} - \tilde{\boldsymbol{b}}). \qquad \text{(B.23)}
\end{aligned}$$

The last equality follows by (B.21). Putting (B.22) and (B.23) together yields (B.19).

For (B.18), we have

$$\|(\Delta_p + \tau I_{p-1})^{-1}\boldsymbol{u}_p\|^2 = \frac{\boldsymbol{x}_p^\top D^{1/2}}{\sqrt{n}} \frac{D^{1/2}X_{-p}}{\sqrt{n}} \left(\frac{X_{-p}^\top D X_{-p}}{n} + \tau I_{p-1}\right)^{-2} \frac{X_{-p}^\top D^{1/2}}{\sqrt{n}} \frac{D^{1/2}\boldsymbol{x}_p}{\sqrt{n}}.$$

Since

$$\left(\frac{X_{-p}^\top D X_{-p}}{n} + \tau I_{p-1}\right)^{-1} \preceq \frac{1}{\tau} I_{p-1},$$

and

$$\frac{D^{1/2}X_{-p}}{\sqrt{n}} \left(\frac{X_{-p}^\top D X_{-p}}{n} + \tau I_{p-1}\right)^{-1} \frac{X_{-p}^\top D^{1/2}}{\sqrt{n}} \preceq I_{mn},$$

we obtain that

$$\|(\Delta_p + \tau I_{p-1})^{-1}\boldsymbol{u}_p\|^2 \leq \frac{1}{n\tau} \boldsymbol{x}_p^\top D \boldsymbol{x}_p = \frac{1}{n\tau} \sum_{i=1}^{n} \boldsymbol{x}_{i,p}^\top \nabla\boldsymbol{\psi}(\check{\boldsymbol{r}}_{i,-p})\boldsymbol{x}_{i,p}.$$

Because $\boldsymbol{x}_{i,p}$ is independent of $\check{\boldsymbol{r}}_{i,-p}$, and $\sup_{\boldsymbol{x}} \lambda_{\max}(\nabla\boldsymbol{\psi}(\boldsymbol{x})) \leq c$, $\|(\Delta_p + \tau I_{p-1})^{-1}\boldsymbol{u}_p\|^2 = O_{L_k}(1)$.

Proposition B.4 *We have*

$$|\tilde{b}_p - \beta_{0,p}| \leq \frac{1}{\tau}|\zeta_p| + |\beta_{0,p}| = O_{L_k}(n^{-1/2} + n^{-\alpha}), \tag{B.24}$$

where $\zeta_p = n^{-1} \sum_i \boldsymbol{x}_{i,p}^\top \boldsymbol{\psi}(\check{\boldsymbol{r}}_{i,-p})$.

Proposition B.5 *We have*

$$\sup_i \|\nabla\boldsymbol{\psi}(\boldsymbol{r}_{2,i}^*) - \nabla\boldsymbol{\psi}(\check{\boldsymbol{r}}_{i,-p})\|_{\max} = O_{L_k}\left(\frac{\text{polyLog}(n)}{\min\{n^{1/2}, n^{\alpha}\}}\right).$$

Proof of proposition B.5 By the assumption (A3),

$$\sup_i \|\nabla \boldsymbol{\psi}(\boldsymbol{r}^*_{2,i}) - \nabla \boldsymbol{\psi}(\check{\boldsymbol{r}}_{i,-p})\|_{\max} \leq c \sup_i \|X_i(\hat{\boldsymbol{\gamma}}_{est} - \tilde{\boldsymbol{b}})\|.$$

It remains to show that

$$\sup_i \|X_i(\hat{\boldsymbol{\gamma}}_{est} - \tilde{\boldsymbol{b}})\| = O_{L_k}\left(\frac{\text{polyLog}(n)}{\min\{n^{1/2}, n^{\alpha}\}}\right).$$

Since $\hat{\boldsymbol{\gamma}}_{est} - \tilde{\boldsymbol{b}} = (\tilde{b}_p - \beta_{0,p})(((\Delta_p + \tau I_{p-1})^{-1}\boldsymbol{u}_p)^\top, -1)^\top$ and $\boldsymbol{u}_p = n^{-1}X_{-p}^\top D\boldsymbol{x}_p$, we have

$$X_i(\hat{\boldsymbol{\gamma}}_{est} - \tilde{\boldsymbol{b}}) = (\tilde{b}_p - \beta_{0,p})[n^{-1}X_{i,-p}(\Delta_p + \tau I_{p-1})^{-1}X_{-p}^\top D\boldsymbol{x}_p - \boldsymbol{x}_{i,p}]. \qquad \text{(B.25)}$$

By Lemma 3.36 in El Karoui (2018), for $s = 1, \ldots, m$, it follows that

$$\begin{aligned}&\sup_i |n^{-1}\boldsymbol{x}_{i,-p}^\top(s)(\Delta_p + \tau I_{p-1})^{-1}X_{-p}^\top D\boldsymbol{x}_p| \\ &= O_{L_k}(\text{polyLog}(n) \sup_i \|n^{-1}\boldsymbol{x}_{i,-p}^\top(s)(\Delta_p + \tau I_{p-1})^{-1}X_{-p}^\top D\|). \qquad \text{(B.26)}\end{aligned}$$

Since

$$\begin{aligned}&\|n^{-1}\boldsymbol{x}_{i,-p}^\top(s)(\Delta_p + \tau I_{p-1})^{-1}X_{-p}^\top D\|^2 \\ =&\frac{1}{n}\boldsymbol{x}_{i,-p}^\top(s)(\Delta_p + \tau I_{p-1})^{-1}\frac{X_{-p}^\top D^2 X_{-p}}{n}(\Delta_p + \tau I_{p-1})^{-1}\boldsymbol{x}_{i,-p}(s) \\ \leq&\frac{1}{n}\lambda_{\max}(D)\boldsymbol{x}_{i,-p}^\top(s)(\Delta_p + \tau I_{p-1})^{-1}\left(\frac{X_{-p}^\top D X_{-p}}{n}(\Delta_p + \tau I_{p-1})^{-1}\right)\boldsymbol{x}_{i,-p}(s) \\ \leq&\frac{\|\boldsymbol{x}_{i,-p}(s)\|^2}{n\tau}\lambda_{\max}(D) \leq \frac{\|\boldsymbol{x}_{i,-p}(s)\|^2}{n\tau}\sup_{1\leq i\leq n}\lambda_{\max}(\nabla\boldsymbol{\psi}(\check{\boldsymbol{r}}_{i,-p})) = O_{L_k}(1),\end{aligned}$$

(B.26) becomes $\sup_i |n^{-1}\boldsymbol{x}_{i,-p}^\top(s)(\Delta_p + \tau I_{p-1})^{-1}X_{-p}^\top D\boldsymbol{x}_p| = O_{L_k}(\text{polyLog}(n))$. We thus have

$$\sup_i \|n^{-1}X_{i,-p}(\Delta_p + \tau I_{p-1})^{-1}D\boldsymbol{x}_p\| = O_{L_k}(\text{polyLog}(n)),$$

which, jointly with (B.25), $\|\boldsymbol{x}_{i,p}\| = O_{L_k}(1)$ and $|\tilde{b}_p - \beta_{0,p}| = O_{L_k}(n^{-1/2} + n^{-\alpha})$ (see Proposition B.4), implies that

$$\sup_i \|X_i(\hat{\boldsymbol{\gamma}}_{est} - \tilde{\boldsymbol{b}})\| = O_{L_k}\left(\frac{\text{polyLog}(n)}{\min\{n^{1/2}, n^\alpha\}}\right).$$

The proof is completed.

Proof of Theorem 3.3 (i) By Propositions B.3-B.5, we have

$$\begin{aligned}\|\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{b}}\| \leq & \lambda_{\max}\left(\frac{1}{n}\sum_i X_i^\top X_i\right)\sup_i \|\nabla\boldsymbol{\psi}(\boldsymbol{r}_{2,i}^*) - \nabla\boldsymbol{\psi}(\check{\boldsymbol{r}}_{i,-p}))\|_{\max}|\tilde{b}_p - \beta_{0,p}| \\ & \times \sqrt{\|(\Delta_p + \tau I_{p-1})^{-1}\boldsymbol{u}_p\|^2 + 1} \\ = & O_{L_k}\left(\text{polyLog}(n)\frac{1}{\min\{n^{1/2}, n^\alpha\}}\frac{\text{polyLog}(n)}{\min\{n^{1/2}, n^\alpha\}}\right) \\ = & O_{L_k}\left(\frac{\text{polyLog}(n)}{(\min\{n^{1/2}, n^\alpha\})^2}\right). \qquad \text{(B.27)}\end{aligned}$$

Furthermore, $\sqrt{n}|\hat{\beta}_p - \tilde{b}_p| \leq \sqrt{n}\|\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{b}}\| = O_{L_k}\left(\frac{n^{1/2}\text{polyLog}(n)}{(\min\{n^{1/2}, n^\alpha\})^2}\right)$.

(ii) By (B.27),

$$\sup_i \|X_i(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{b}})\| \leq \sum_{s=1}^m \sup_i |\boldsymbol{x}_i^\top(s)(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{b}})| = O_{L_k}\left(\frac{n^{1/2}\text{polyLog}(n)}{(\min\{n^{1/2}, n^\alpha\})^2}\right).$$

(iii) Since $\boldsymbol{r}_i - \check{\boldsymbol{r}}_{i,-p} = X_i(\hat{\boldsymbol{\gamma}}_{est} - \hat{\boldsymbol{\beta}}) = X_i(\hat{\boldsymbol{\gamma}}_{est} - \tilde{\boldsymbol{b}}) + X_i(\tilde{\boldsymbol{b}} - \hat{\boldsymbol{\beta}})$, we obtain that

$$\begin{aligned}\sup_i \|\boldsymbol{r}_i - \check{\boldsymbol{r}}_{i,-p}\| \leq & \sup_i \|X_i(\hat{\boldsymbol{\gamma}}_{est} - \tilde{\boldsymbol{b}})\| + \sup_i \|X_i(\tilde{\boldsymbol{b}} - \hat{\boldsymbol{\beta}})\| \\ = & O_{L_k}\left(\frac{\text{polyLog}(n)}{\min\{n^{1/2}, n^{\alpha}\}}\right) + O_{L_k}\left(\frac{n^{1/2}\text{polyLog}(n)}{(\min\{n^{1/2}, n^{\alpha}\})^2}\right) \\ = & O_{L_k}\left(\frac{n^{1/2}\text{polyLog}(n)}{\min\{n, n^{2\alpha}\}}\right).\end{aligned}$$

Propositions B.6-B.8 are needed in the proof of Theorem 3.1.

Proposition B.6 *Define* $c_{\tau,t} = n^{-1}\text{tr}((\Delta_t + \tau I_{p-1})^{-1})$, *where* $\Delta_t = n^{-1}\sum_i X_{i,-t}^{\top}\nabla\boldsymbol{\psi}(\check{\boldsymbol{r}}_{i,-t})X_{i,-t}$, $1 \leq t \leq p$. *We have*

(i)

$$|c_\tau - c_{\tau,t}| = O_{L_k}\left(\frac{n^{1/2}\text{polyLog}(n)}{\min\{n, n^{2\alpha}\}}\right). \tag{B.28}$$

(ii)

$$\left(\frac{p}{n}\right)^2 E\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2 = \frac{p}{n}\frac{1}{n}\sum_{i=1}^{n} E\|\tilde{\boldsymbol{r}}_{i,[-i]} - \text{prox}_{c_\tau}(\rho)(\tilde{\boldsymbol{r}}_{i,[-i]})\|^2 + \tau^2\|\boldsymbol{\beta}_0\|^2 E(c_\tau^2) + o(1). \tag{B.29}$$

Proposition B.7 *We have*

$$\tilde{\boldsymbol{r}}_{i,[-i]} \xrightarrow{\mathcal{D}} \boldsymbol{e}_i + \sqrt{E\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2}\tilde{\boldsymbol{z}}_i, \tag{B.30}$$

as $n, p \to \infty$ *with* $p/n \to \kappa$, *where* $\tilde{\boldsymbol{z}}_i \sim N(\boldsymbol{0}, I_m)$, *and* $\tilde{\boldsymbol{z}}_i$ *is independent of* $\boldsymbol{e}_i$. *Furthermore,* $\tilde{\boldsymbol{r}}_{i,[-i]}$ *and* $\tilde{\boldsymbol{r}}_{j,[-j]}$ *are asymptotically independent for* $i \neq j$.

Proof of Proposition B.7 The proof in Theorem 3.2(iii) imply that $E(\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|^2 - \|\hat{\boldsymbol{\beta}}_{(i)} - \boldsymbol{\beta}_0\|^2)^2 = O(n^{-2}\text{polyLog}(n))$, which reduces to

$$E|\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2 - \|\hat{\boldsymbol{\beta}}_{(i)} - \boldsymbol{\beta}_0\|^2| = O(n^{-1}\text{polyLog}(n)). \tag{B.31}$$

Since $E\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2$ is uniformly bounded (taking $k = 1$ in (B.42)), by (B.31), we obtain that $E\|\hat{\boldsymbol{\beta}}_{(i)} - \boldsymbol{\beta}_0\|^2$ is uniformly bounded. Without loss of generality, we assume that $\|\hat{\boldsymbol{\beta}}_{(i)} - \boldsymbol{\beta}_0\|$ is bounded away from zero. Otherwise, $E\|X_i(\hat{\boldsymbol{\beta}}_{(i)} - \boldsymbol{\beta}_0)\|^2 = mE\|\hat{\boldsymbol{\beta}}_{(i)} - \boldsymbol{\beta}_0\|^2 \to 0$. Hence $X_i(\hat{\boldsymbol{\beta}}_{(i)} - \boldsymbol{\beta}_0) \xrightarrow{\mathcal{D}} \mathbf{0}$. By Theorem 3.3(i) and (B.24), it is easy to see that

$$\begin{aligned}|\hat{\beta}_p - \beta_{0,p}| \leq & |\hat{\beta}_p - \tilde{b}_p| + |\tilde{b}_p - \beta_{0,p}| \\ = & O_{L_k}\left(\frac{\text{polyLog}(n)}{\min\{n, n^{2\alpha}\}}\right) + O_{L_k}\left(\frac{1}{\min\{n^{1/2}, n^{\alpha}\}}\right) = O_{L_k}\left(\frac{1}{\min\{n^{1/2}, n^{\alpha}\}}\right).\end{aligned}$$

Denote $\hat{\boldsymbol{\beta}}_{(i)} = (\hat{\beta}_{(i),1}, \ldots, \hat{\beta}_{(i),p})^\top$. By Theorem 3.2(i),

$$|\hat{\beta}_{(i),p} - \hat{\beta}_p| \leq \|\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}}\| = O_{L_k}(n^{-1/2}).$$

We thus have

$$|\hat{\beta}_{(i),p} - \beta_{0,p}| = O_{L_k}\left(\frac{1}{\min\{n^{1/2}, n^{\alpha}\}}\right),$$

and

$$E(|\hat{\beta}_{(i),p} - \beta_{0,p}|^3) = O\left(\frac{1}{(\min\{n^{1/2}, n^{\alpha}\})^3}\right).$$

It follows that, for $\alpha > 1/3$,

$$E\left(\sum_{k=1}^{p} |\hat{\beta}_{(i),k} - \beta_{0,k}|^3\right) = O\left(\frac{n}{(\min\{n^{1/2}, n^{\alpha}\})^3}\right) = o(1).$$

In light of the proof of Lemma 3.23 in El Karoui (2018), by the assumption (A4),

$$X_i(\hat{\boldsymbol{\beta}}_{(i)} - \boldsymbol{\beta}_0) = \sum_{k=1}^{p} \boldsymbol{x}_{i,k}(\hat{\beta}_{(i),k} - \beta_{0,k}) \xrightarrow{\mathcal{D}} \|\hat{\boldsymbol{\beta}}_{(i)} - \boldsymbol{\beta}_0\|\tilde{\boldsymbol{z}}_i, \tag{B.32}$$

where $\tilde{\boldsymbol{z}}_i \sim N(\mathbf{0}, I_m)$. In view that $\mathrm{var}(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2) = o(1)$ (see Theorem 2.2(iii)), we can similarly obtain that $\mathrm{var}(\|\hat{\boldsymbol{\beta}}_{(i)} - \boldsymbol{\beta}_0\|^2) = o(1)$, which yields that $\|\hat{\boldsymbol{\beta}}_{(i)} - \boldsymbol{\beta}_0\|^2 \xrightarrow{p} E\|\hat{\boldsymbol{\beta}}_{(i)} - \boldsymbol{\beta}_0\|^2$. By Slutsky's theorem, (B.32) and (B.31), $X_i(\hat{\boldsymbol{\beta}}_{(i)} - \boldsymbol{\beta}_0) \xrightarrow{\mathcal{D}} \sqrt{E\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2}\tilde{\boldsymbol{z}}_i$.

For the asymptotic independence of $\tilde{\boldsymbol{r}}_{i,[-i]}$ and $\tilde{\boldsymbol{r}}_{j,[-j]}$, its proof mimics the proof of Lemma 3.23 (second part) given in El Karoui (2018).

Proposition B.8 *Denote the random function* $\delta_n(x) = p/n - \tau x - m + n^{-1}\sum_{i=1}^{n} \mathrm{tr}([I_m + x\nabla\boldsymbol{\psi}(\mathrm{prox}_x(\rho)(\tilde{\boldsymbol{r}}_{i,[-i]}))]^{-1})$. *Under the assumptions (A1)-(A5) and (F1), we have,*

$$\delta_n(c_\tau) = o_{L_k}(1), \tag{B.33}$$

and c_τ *is asymptotically deterministic.*

Proof of Proposition B.8 Recall that

$$P_{ii} = I_m - \left(I_m + n^{-1}\nabla\boldsymbol{\psi}^{1/2}(\check{\boldsymbol{r}}_{i,-p})X_{i,-p}(\Delta_p(i) + \tau I_{p-1})^{-1}X_{i,-p}^{\top}\nabla\boldsymbol{\psi}^{1/2}(\check{\boldsymbol{r}}_{i,-p})\right)^{-1}.$$

We obtain that

$$\frac{1}{n}\mathrm{tr}(P) = m - \frac{1}{n}\sum_{i=1}^{n}\mathrm{tr}([I_m + n^{-1}\nabla\boldsymbol{\psi}^{1/2}(\check{\boldsymbol{r}}_{i,-p})X_{i,-p}(\Delta_p(i)+\tau I_{p-1})^{-1}X_{i,-p}^{\top}\nabla\boldsymbol{\psi}^{1/2}(\check{\boldsymbol{r}}_{i,-p})]^{-1}). \tag{B.34}$$

It is easy to see that

$$\begin{aligned}
&|\mathrm{tr}([I_m + n^{-1}\nabla\boldsymbol{\psi}^{1/2}(\check{\boldsymbol{r}}_{i,-p})X_{i,-p}(\Delta_p(i)+\tau I_{p-1})^{-1}X_{i,-p}^{\top}\nabla\boldsymbol{\psi}^{1/2}(\check{\boldsymbol{r}}_{i,-p})]^{-1}) \\
&- \mathrm{tr}([I_m + c_{\tau,p}\nabla\boldsymbol{\psi}(\check{\boldsymbol{r}}_{i,-p})]^{-1})| \\
\leq&|\mathrm{tr}(c_{\tau,p}\nabla\boldsymbol{\psi}(\check{\boldsymbol{r}}_{i,-p}) - n^{-1}\nabla\boldsymbol{\psi}^{1/2}(\check{\boldsymbol{r}}_{i,-p})X_{i,-p}(\Delta_p(i)+\tau I_{p-1})^{-1}X_{i,-p}^{\top}\nabla\boldsymbol{\psi}^{1/2}(\check{\boldsymbol{r}}_{i,-p}))| \\
=&|\mathrm{tr}([c_{\tau,p}I_m - \frac{1}{n}X_{i,-p}(\Delta_p(i)+\tau I_{p-1})^{-1}X_{i,-p}^{\top}]\nabla\boldsymbol{\psi}(\check{\boldsymbol{r}}_{i,-p}))| \\
=&|\mathrm{tr}(-\Omega_i\nabla\boldsymbol{\psi}(\check{\boldsymbol{r}}_{i,-p}))| = O_{L_k}\left(\frac{\mathrm{polyLog}(n)}{\sqrt{n}}\right).
\end{aligned} \tag{B.35}$$

The last equality follows by (B.45). (B.34), together with (B.44) and (B.35), becomes

$$\frac{p}{n} - \tau c_{\tau,p} - m + \frac{1}{n}\sum_{i=1}^{n}\mathrm{tr}([I_m + c_{\tau,p}\nabla\boldsymbol{\psi}(\check{\boldsymbol{r}}_{i,-p})]^{-1}) = O_{L_k}\left(\frac{\mathrm{polyLog}(n)}{\sqrt{n}}\right). \tag{B.36}$$

Similarly, (B.36) still holds if we replace $c_{\tau,p}, \check{\boldsymbol{r}}_{i,-p}$ with $c_\tau, \boldsymbol{r}_i$, respectively. That is

$$\frac{p}{n} - \tau c_{\tau} - m + \frac{1}{n}\sum_{i=1}^{n}\mathrm{tr}([I_m + c_{\tau}\nabla\boldsymbol{\psi}(\boldsymbol{r}_i)]^{-1}) = O_{L_k}\left(\frac{\mathrm{polyLog}(n)}{\sqrt{n}}\right). \tag{B.37}$$

By Theorem 3.2(ii), and the assumption (A3), it follows that

$$\begin{aligned}
\sup_i \|\nabla\boldsymbol{\psi}(\boldsymbol{r}_i) - \nabla\boldsymbol{\psi}(\mathrm{prox}_{c_i}(\rho)(\tilde{\boldsymbol{r}}_{i,[-i]}))\|_{\max} \leq& \sup_i c\|\boldsymbol{r}_i - \mathrm{prox}_{c_i}(\rho)(\tilde{\boldsymbol{r}}_{i,[-i]})\| \\
=&O_{L_k}(n^{-1/2}\mathrm{polyLog}(n)).
\end{aligned}$$

Since $\|\text{prox}_{c_i}(\rho)(\tilde{\boldsymbol{r}}_{i,[-i]}) - \text{prox}_{c_\tau}(\rho)(\tilde{\boldsymbol{r}}_{i,[-i]})\| = O_{L_k}(n^{-1/2}\text{polyLog}(n))$ (proved in Proposition B.6(ii)), by the assumption (A3), we have

$$\|\nabla\boldsymbol{\psi}(\text{prox}_{c_i}(\rho)(\tilde{\boldsymbol{r}}_{i,[-i]})) - \nabla\boldsymbol{\psi}(\text{prox}_{c_\tau}(\rho)(\tilde{\boldsymbol{r}}_{i,[-i]}))\|_{\max} = O_{L_k}(n^{-1/2}\text{polyLog}(n)).$$

Therefore,

$$\|\nabla\boldsymbol{\psi}(\boldsymbol{r}_i) - \nabla\boldsymbol{\psi}(\text{prox}_{c_\tau}(\rho)(\tilde{\boldsymbol{r}}_{i,[-i]}))\|_{\max} = O_{L_k}(n^{-1/2}\text{polyLog}(n)). \tag{B.38}$$

By (B.37) and (B.38), (B.33) follows.

Note that c_τ is asymptotically deterministic if one can show that c_τ converges to a constant in probability. By the assumption (F1), for a given $\boldsymbol{u}_0 \in \mathbb{R}^m$, $\text{tr}((I_m + x\nabla\boldsymbol{\psi}(\text{prox}_x(\rho)(\boldsymbol{u}_0)))^{-1})$ is a decreasing function for $x \geq 0$. Thus for any random vector $\boldsymbol{u} \in \mathbb{R}^m$, $E(\text{tr}((I_m + x\nabla\boldsymbol{\psi}(\text{prox}_x(\rho)(\boldsymbol{u})))^{-1}))$ is decreasing for $x \geq 0$. Given $\tau > 0$, $E(\delta_n(x))$ is strictly decreasing for $x \geq 0$. Since $E(\delta_n(0)) = p/n > 0$ and $E(\delta_n(x)) \to -\infty$ as $x \to +\infty$, there exists a unique root of $E(\delta_n(x)) = 0$ for $x > 0$. We denote the unique root by μ.

Define $\boldsymbol{\mathcal{S}} = \{x : \delta_n(x) = o_{L_k}(1)\}$, and $\mathcal{F}_{n,\epsilon} = \{x : |E(\delta_n(x))| \leq \epsilon\}$ for any given ϵ. Note that $\mathcal{F}_{n,\epsilon} \subseteq (0, p/(n\tau) + \epsilon/\tau]$ and $\mathcal{F}_{n,\epsilon}$ is compact. Using the technique as in the proof of Lemma 3.26 in El Karoui (2018), and by (B.33) and the result in Lemma B.8, one can easily obtain that, for any $\epsilon > 0$, c_τ belongs to $\mathcal{F}_{n,\epsilon}$ with high probability. Let μ_1 be the limit of c_τ as $n \to \infty$. Since $\mathcal{F}_{n,\epsilon}$ is compact, we conclude that $\mu_1 \in \mathcal{F}_{n,\epsilon}$ with high probability. On the other hand, if $\epsilon = 0$, then $\mathcal{F}_{n,\epsilon}$ reduces to a single point, μ, as $n \to \infty$. It follows that $\mu = \mu_1$ with high probability. Thus $c_\tau \to \mu$ with high probability as $n \to \infty$. We remark that an event occurs with high probability can be made as close as desired to 1 by making n large enough. This completes the proof.

Proof of Theorem 3.1 Let μ be the limit of c_τ as $n, p \to \infty$ with $p/n \to \kappa(>0)$. By (B.33), c_τ is asymptotically arbitrarily close to the solution of $E(\delta_n(x)) = 0$. Together with $\nabla \text{prox}_\mu(\rho)(\boldsymbol{z}) = (I_m + \mu \nabla \boldsymbol{\psi}(\boldsymbol{z}))^{-1}$, the first equality in (3.7) holds. By (B.29) and (B.31), the second equation in (3.7) follows.

Proof of Corollary 3.1 By simply following the proof of Theorem 6.1 in El Karoui (2013) in the beginning,

$$\|\hat{\boldsymbol{\beta}}_M - \hat{\boldsymbol{\beta}}\| \leq \frac{\sqrt{2\tau}}{c\lambda_{\min}(n^{-1}\sum_{i=1}^n X_i^\top X_i)} \sqrt{\frac{1}{n}\sum_{i=1}^n \rho(\boldsymbol{e}_i + X_i\boldsymbol{\beta}_0)}.$$

By the mean value theorem, it follows that $\rho(\boldsymbol{e}_i + X_i\boldsymbol{\beta}_0) = \rho(\mathbf{0}) + \boldsymbol{\psi}^\top(\boldsymbol{r}_{i,e}^*)(\boldsymbol{e}_i + X_i\boldsymbol{\beta}_0) = \boldsymbol{\psi}^\top(\boldsymbol{r}_{i,e}^*)(\boldsymbol{e}_i + X_i\boldsymbol{\beta}_0)$, where $\boldsymbol{r}_{i,e}^* \in (\mathbf{0}, \boldsymbol{e}_i + X_i\boldsymbol{\beta}_0)$. It is easy to see that $\|X_i\boldsymbol{\beta}_0\| = O_p(1)$ because, by the assumptions (A2) and (A4), $E\|X_i\boldsymbol{\beta}_0\|^2 = m\|\boldsymbol{\beta}_0\|^2 = O(1)$. By the assumptions (A5), we conclude that $\|\boldsymbol{e}_i\| = O_p(1)$. Together with that $\sup\|\boldsymbol{\psi}\| \leq c$, it follows that $\rho(\boldsymbol{e}_i + X_i\boldsymbol{\beta}_0) = \boldsymbol{\psi}^\top(\boldsymbol{r}_{i,e}^*)(\boldsymbol{e}_i + X_i\boldsymbol{\beta}_0) = O_p(1)$. By Theorem 2.16 in Bai (1999), $\lambda_{\min}(n^{-1}\sum_{i=1}^n X_i^\top X_i) \to \left(\sqrt{m} - \sqrt{p/n}\right)^2$ in probability and almost surely. Thus $\|\hat{\boldsymbol{\beta}}_M - \hat{\boldsymbol{\beta}}\| \to 0$ as $\tau \to 0$, and it follows that

$\left| \|\hat{\boldsymbol{\beta}}_M - \boldsymbol{\beta}_0\| - \lim_{\tau \to 0} \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| \right| \leq \lim_{\tau \to 0} \|\hat{\boldsymbol{\beta}}_M - \hat{\boldsymbol{\beta}}\| = 0$. This completes the proof.

B.2 Proofs of lemmas

This part contains the proofs of these lemmas that are used to prove those propositions and theorems. If not specified, we present the results below under the assumptions (A1)-(A5) in Chapter 3.

Lemma B.1 *For any two vectors $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathbb{R}^p$, we have*

$$\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\| \leq \frac{1}{\tau} \|\boldsymbol{\phi}(\boldsymbol{\beta}_1) - \boldsymbol{\phi}(\boldsymbol{\beta}_2)\|. \tag{B.39}$$

The lemma can be easily proved by following the same procedure as the proof of Proposition 2.1 in El Karoui (2013).

Lemma B.2 *Denote $\boldsymbol{w}_n = n^{-1} \sum_i X_i^\top \boldsymbol{\psi}(\boldsymbol{e}_i)$. Then*

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| \leq \|\boldsymbol{\beta}_0\| + \frac{1}{\tau} \|\boldsymbol{w}_n\|, \tag{B.40}$$

and

$$E\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2 \leq 2 \left(\|\boldsymbol{\beta}_0\|^2 + \frac{1}{\tau^2} \frac{mp}{n} \frac{1}{n} \sum_i E\|\boldsymbol{\psi}(\boldsymbol{e}_i)\|^2 \right). \tag{B.41}$$

Furthermore, as $p, n \to \infty$ with p/n tending to a constant, for any positive integer $k \geq 1$, there exists a constant $c > 0$ such that

$$E\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^{2k} \leq c(\|\boldsymbol{\beta}_0\|^{2k} + \tau^{-2k} c) = O(1). \tag{B.42}$$

Proof of Lemma B.2 For the proof of this lemma, one can refer to the proof line of Lemma 2.2 in El Karoui (2013) . $E\|\boldsymbol{w}_n\|^2 \leq n^{-2}mp\sum_{i=1}^n E\|\boldsymbol{\psi}(\boldsymbol{e}_i)\|^2$. Because $\boldsymbol{e}_i$ and X_i are independent, $EX_i = \mathbf{0}$ (see the assumption (A4)) and $\sup_{\boldsymbol{x}} \|\boldsymbol{\psi}(\boldsymbol{x})\| < \infty$ (see the assumption (A3)), and thus

$$\begin{aligned} E\|\boldsymbol{w}_n\|^2 \leq & \frac{1}{n^2}\sum_{i=1}^n E\lambda_{\max}(X_iX_i^\top)E\|\boldsymbol{\psi}(\boldsymbol{e}_i)\|^2 \\ \leq & \frac{1}{n^2}\sum_{i=1}^n E\mathrm{tr}(X_iX_i^\top)E\|\boldsymbol{\psi}(\boldsymbol{e}_i)\|^2 \\ = & \frac{mp}{n^2}\sum_{i=1}^n E\|\boldsymbol{\psi}(\boldsymbol{e}_i)\|^2. \end{aligned} \tag{B.43}$$

The '=' follows by the fact that $E\mathrm{tr}(X_iX_i^\top) = \mathrm{tr}(E(X_iX_i^\top)) = mp$. Hence (B.41) holds.

Lemma B.3 *We have $\xi_n \geq 0$, and*

$$|\xi_n - n^{-1}\mathrm{tr}(D^{1/2}(I_{mn} - P)D^{1/2})| = O_{L_k}\left(\frac{\mathrm{polyLog}(n)}{\sqrt{n}}\right).$$

The lemma can be easily proved by following the proof of Lemma 3.13 in El Karoui (2018).

Lemma B.4 *Denote $c_{\tau,p} = n^{-1}\mathrm{tr}((\Delta_p+\tau I_{p-1})^{-1})$, $\Omega_i = n^{-1}X_{i,-p}(\Delta_p(i)+\tau I_{p-1})^{-1}X_{i,-p}^\top - c_{\tau,p}I_m$. Therefore, we have*

$$\left|\frac{1}{n}\mathrm{tr}(P) - \frac{1}{n}\mathrm{tr}(D^{1/2}(I_{mn} - P)D^{1/2})c_{\tau,p}\right| \leq \frac{1}{n}\sum_i \mathrm{tr}(\nabla\boldsymbol{\psi}(\check{\boldsymbol{r}}_{i,-p})\Omega_i),$$

and

$$\frac{1}{n}\mathrm{tr}(P) = \frac{p-1}{n} - \tau c_{\tau,p}. \tag{B.44}$$

Proof of Lemma B.4 Using the fact that $I_m = A^{-1}A$ with $A = I_m + \nabla\boldsymbol{\psi}^{1/2}(\check{\boldsymbol{r}}_{i,-p})(c_{\tau,p}I_m + \Omega_i)\nabla\boldsymbol{\psi}^{1/2}(\check{\boldsymbol{r}}_{i,-p})$, we have

$$\begin{aligned} P_{ii} &= (I_m - P_{ii})(\nabla\boldsymbol{\psi}^{1/2}(\check{\boldsymbol{r}}_{i,-p})(c_{\tau,p}I_m + \Omega_i)\nabla\boldsymbol{\psi}^{1/2}(\check{\boldsymbol{r}}_{i,-p})) \\ &= (I_m - P_{ii})\nabla\boldsymbol{\psi}(\check{\boldsymbol{r}}_{i,-p})c_{\tau,p} + (I_m - P_{ii})\nabla\boldsymbol{\psi}^{1/2}(\check{\boldsymbol{r}}_{i,-p})\Omega_i\nabla\boldsymbol{\psi}^{1/2}(\check{\boldsymbol{r}}_{i,-p}), \end{aligned}$$

which entails that

$$\begin{aligned} &\left|\frac{1}{n}\mathrm{tr}(P) - \frac{1}{n}\mathrm{tr}(D^{1/2}(I_{mn} - P)D^{1/2})c_{\tau,p}\right| \\ =&\frac{1}{n}\sum_i \mathrm{tr}((I_m - P_{ii})\nabla\boldsymbol{\psi}^{1/2}(\check{\boldsymbol{r}}_{i,-p})\Omega_i\nabla\boldsymbol{\psi}^{1/2}(\check{\boldsymbol{r}}_{i,-p})) \\ \leq&\frac{1}{n}\sum_i \mathrm{tr}(\nabla\boldsymbol{\psi}(\check{\boldsymbol{r}}_{i,-p})\Omega_i). \end{aligned}$$

Since $B = n^{-1/2}D^{1/2}X_{-p}$, $P = B(B^\top B + \tau I_{p-1})^{-1}B^\top$, $\Delta_p = B^\top B$, we have

$$\mathrm{tr}(P) = \mathrm{tr}((\Delta_p + \tau I_{p-1})^{-1}\Delta_p) = \mathrm{tr}(I_{p-1}) - \tau\mathrm{tr}((\Delta_p + \tau I_{p-1})^{-1}) = p - 1 - n\tau c_{\tau,p}.$$

Lemma B.5 *Denote the* (s,t)*-th entry of* Ω_i *by* $\omega_i^{(s,t)}$*. Then*

$$\sup_{i,s,t}|\omega_i^{(s,t)}| = O_{L_k}\left(\frac{\mathrm{polyLog}(n)}{\sqrt{n}}\right). \tag{B.45}$$

Proof of Lemma B.5 Let $\check{\boldsymbol{r}}_{j,-p}^{(i)}$ be the residuals obtained by leaving the p-th predictor and i-th observation out. By Theorem 3.2 (Lemma B.5 is not required for the proof of Theorem 3.2), we have

$$\sup_{j\neq i}|\check{\boldsymbol{r}}_{j,-p}^{(i)}-\check{\boldsymbol{r}}_{i,-p}|=O_{L_k}\left(\frac{\text{polyLog}(n)}{\sqrt{n}}\right),$$

which, jointly with the assumption (A3), implies that

$$\sup_i\sup_{j\neq i}\|\nabla\boldsymbol{\psi}(\check{\boldsymbol{r}}_{j,-p}^{(i)})-\nabla\boldsymbol{\psi}(\check{\boldsymbol{r}}_{i,-p})\|_{\max}=O_{L_k}\left(\frac{\text{polyLog}(n)}{\sqrt{n}}\right). \tag{B.46}$$

Define $\Lambda_{i,p}=n^{-1}\sum_{j\neq i}X_{j,-p}^\top\nabla\boldsymbol{\psi}(\check{\boldsymbol{r}}_{j,-p}^{(i)})X_{j,-p}$. We have

$$\begin{aligned}
&\|(\Delta_p(i)+\tau I_{p-1})^{-1}-(\Lambda_{i,p}+\tau I_{p-1})^{-1}\|_{\max}\\
=&\|(\Delta_p(i)+\tau I_{p-1})^{-1}(\Lambda_{i,p}-\Delta_p(i))(\Lambda_{i,p}+\tau I_{p-1})^{-1})\|_{\max}\\
\leq&\|n^{-1}\sum_{j\neq i}X_{j,-p}^\top\nabla\boldsymbol{\psi}(\check{\boldsymbol{r}}_{j,-p}^{(i)})X_{j,-p}-n^{-1}\sum_{j\neq i}X_{j,-p}^\top\nabla\boldsymbol{\psi}(\check{\boldsymbol{r}}_{i,-p})X_{j,-p}\|_{\max}\\
\leq&\frac{1}{\tau^2}\lambda_{\max}(n^{-1}\sum_i X_i^\top X_i)\sup_i\sup_{j\neq i}\|\nabla\boldsymbol{\psi}(\check{\boldsymbol{r}}_{j,-p}^{(i)})-\nabla\boldsymbol{\psi}(\check{\boldsymbol{r}}_{i,-p})\|_{\max}\\
=&O_{L_k}\left(\frac{\text{polyLog}(n)}{\sqrt{n}}\right).
\end{aligned}$$

The first equality follows by the fact that $A^{-1}-B^{-1}=A^{-1}(B-A)B^{-1}$ for invertible matrix A and B, and the second equality holds in view of (B.46) and because $\lambda_{\max}(n^{-1}\sum_i X_i^\top X_i)=O_{L_k}(\text{polyLog}(n))$, a straightforward result of Lemma B-5 in

El Karoui (2013). Similarly, we have,

$$\begin{aligned}
&\left|\frac{1}{n}\mathrm{tr}((\Delta_p(i)+\tau I_{p-1})^{-1})-\frac{1}{n}\mathrm{tr}((\Lambda_{i,p}+\tau I_{p-1})^{-1})\right|\\
\leq&\frac{1}{n}\sup_i\sup_{j\neq i}\|\nabla\boldsymbol{\psi}(\check{\boldsymbol{r}}_{j,-p}^{(i)})-\nabla\boldsymbol{\psi}(\check{\boldsymbol{r}}_{i,-p})\|_{\max}\mathrm{tr}(n^{-1}\sum_i X_i^\top X_i)\\
=&O_{L_k}\left(\frac{p}{n}\frac{\mathrm{polyLog}(n)}{\sqrt{n}}\mathrm{polyLog}(n)\right)=O_{L_k}\left(\frac{\mathrm{polyLog}(n)}{\sqrt{n}}\right).
\end{aligned}\tag{B.47}$$

Next, we show that (B.45) holds true for $s=t$ and $s\neq t$. Firstly, we work on the case where $s=t$. Denote $\omega_i^{(s,s)}=n^{-1}\boldsymbol{x}_{i,-p}^\top(s)(\Delta_p(i)+\tau I_{p-1})^{-1}\boldsymbol{x}_{i,-p}(s)-c_{\tau,p}$, where $\boldsymbol{x}_{i,-p}(s)$ is the s-th row of $X_{i,-p}$. It is easy to see that the conclusion (B.45) is implied by (B.47) and the following results (B.48)-(B.50).

$$\begin{aligned}
&\left|\frac{1}{n}\boldsymbol{x}_{i,-p}^\top(s)(\Delta_p(i)+\tau I_{p-1})^{-1}\boldsymbol{x}_{i,-p}(s)-\frac{1}{n}\boldsymbol{x}_{i,-p}^\top(s)(\Lambda_{i,p}+\tau I_{p-1})^{-1}\boldsymbol{x}_{i,-p}(s)\right|\\
\leq&\frac{\|\boldsymbol{x}_{i,-p}(s)\|^2}{n}O_{L_k}\left(\frac{\mathrm{polyLog}(n)}{\sqrt{n}}\right)=O_{L_k}\left(\frac{\mathrm{polyLog}(n)}{\sqrt{n}}\right),
\end{aligned}\tag{B.48}$$

and in light of Lemma 3.37 in El Karoui (2013),

$$\sup_{i,s}\left|\frac{1}{n}\boldsymbol{x}_{i,-p}^\top(s)(\Lambda_{i,p}+\tau I_{p-1})^{-1}\boldsymbol{x}_{i,-p}(s)-\frac{1}{n}\mathrm{tr}((\Lambda_{i,p}+\tau I_{p-1})^{-1})\right|=O_{L_k}\left(\frac{\mathrm{polyLog}(n)}{\sqrt{n}}\right),\tag{B.49}$$

where $\Lambda_{i,p}$ is independent of $X_{i,-p}$. Since $A^{-1}-B^{-1}=A^{-1}(B-A)B^{-1}$, we obtain that

$$\begin{aligned}
&\frac{1}{n}\mathrm{tr}((\Delta_p(i)+\tau I_{p-1})^{-1}-(\Delta_p+\tau I_{p-1})^{-1})\\
\leq&\frac{1}{n\tau^2}\mathrm{tr}(n^{-1}X_{i,-p}^{\top}\nabla\boldsymbol{\psi}(\check{\boldsymbol{r}}_{i,-p})X_{i,-p})\\
=&\frac{1}{n\tau^2}\frac{1}{n}\sum_{s=1}^{p-1}\boldsymbol{x}_{i,s}^{\top}\nabla\boldsymbol{\psi}(\check{\boldsymbol{r}}_{i,-p})\boldsymbol{x}_{i,s}=O_{L_k}(n^{-1}).
\end{aligned} \tag{B.50}$$

Secondly, we show that (B.45) still holds for $s\neq t$, where $\omega_i^{(s,t)}=n^{-1}\boldsymbol{x}_{i,-p}^{\top}(s)(\Delta_p(i)+\tau I_{p-1})^{-1}\boldsymbol{x}_{i,-p}(t)$. We thus have

$$\begin{aligned}
&\left|\frac{1}{n}\boldsymbol{x}_{i,-p}^{\top}(s)(\Delta_p(i)+\tau I_{p-1})^{-1}\boldsymbol{x}_{i,-p}(t)-\frac{1}{n}\boldsymbol{x}_{i,-p}^{\top}(s)(\Lambda_{i,p}+\tau I_{p-1})^{-1}\boldsymbol{x}_{i,-p}(t)\right|\\
\leq&\frac{\|\boldsymbol{x}_{i,-p}(s)\|\|\boldsymbol{x}_{i,-p}(t)\|}{n}O_{L_k}\left(\frac{\mathrm{polyLog}(n)}{\sqrt{n}}\right)=O_{L_k}\left(\frac{\mathrm{polyLog}(n)}{\sqrt{n}}\right).
\end{aligned} \tag{B.51}$$

On the other hand, since $n^{-1}\boldsymbol{x}_{i,-p}^{\top}(s)(\Lambda_{i,p}+\tau I_{p-1})^{-1}\boldsymbol{x}_{i,-p}(t)=\boldsymbol{x}_{i,-p}^{\top}(s)\boldsymbol{v}$, where $\boldsymbol{v}=n^{-1}(\Lambda_{i,p}+\tau I_{p-1})^{-1}\boldsymbol{x}_{i,-p}(t)$ and $\boldsymbol{v}$ is independent of $\boldsymbol{x}_{i,-p}(s)$, by Lemma 3.36 in El Karoui (2018), we have

$$\begin{aligned}
&\frac{1}{n}\sup_{1\leq i\leq n,1\leq s\neq t\leq m}|\boldsymbol{x}_{i,-p}^{\top}(s)(\Lambda_{i,p}+\tau I_{p-1})^{-1}\boldsymbol{x}_{i,-p}(t)|\\
\leq&\frac{1}{\sqrt{n}}\sup\frac{\|\boldsymbol{x}_{i,-p}(t)\|}{\sqrt{n}}\mathrm{polyLog}(n)=O_{L_k}\left(\frac{\mathrm{polyLog}(n)}{\sqrt{n}}\right).
\end{aligned} \tag{B.52}$$

Then, (B.51)-(B.52) imply that

$$\sup_{1\leq i\leq n,1\leq s\neq t\leq m}|\frac{1}{n}\boldsymbol{x}_{i,-p}^{\top}(s)(\Delta_p(i)+\tau I_{p-1})^{-1}\boldsymbol{x}_{i,-p}(t)|=O_{L_k}\left(\frac{\mathrm{polyLog}(n)}{\sqrt{n}}\right). \tag{B.53}$$

We complete the proof.

Lemma B.6 *We have*

$$\left| c_{\tau,p}(\xi_n + \tau) - \frac{p-1}{n} \right| = O_{L_k}(n^{-1/2}\text{polyLog}(n)), \tag{B.54}$$

and

$$\frac{p^2}{n^2} nE(\tilde{b}_p - \beta_{0,p})^2 = \frac{1}{n}\sum_{i=1}^{n} E\|c_{\tau,p}\boldsymbol{\psi}(\check{\boldsymbol{r}}_{i,-p})\|^2 + n\tau^2\beta_{0,p}^2 E(c_{\tau,p}^2) + o(1). \tag{B.55}$$

The lemma can be easily proved by following the proof of Proposition 3.18 in El Karoui (2018).

Lemma B.7 *Denote the random function* $g_n(x) = n^{-1}\sum_{i=1}^{n} \text{tr}([I_m + x\nabla\boldsymbol{\psi}(\text{prox}_x(\rho)(\tilde{\boldsymbol{r}}_{i,[-i]}))]^{-1})$, $x \geq 0$. *For any* $(x, y) \in \mathbb{R}^2$, *and* $0 \leq x \leq b$, $0 \leq y \leq b$, *where* $0 < b < \infty$ *and* $b \in \mathbb{R}$, *there exists a constant c such that*

$$\sup_{(x,y):|x-y|\leq\eta, 0\leq x,y\leq b} |g_n(x) - g_n(y)| \leq \eta(1+b)c.$$

Proof of Lemma B.7 Define $h_{\boldsymbol{u}}(x) = \text{tr}([I_m + x\nabla\boldsymbol{\psi}(\text{prox}_x(\rho)(\boldsymbol{u}))]^{-1}) = \text{tr}\left(\nabla\text{prox}_x(\rho)(\boldsymbol{u})\right)$, where $x \geq 0$, $\boldsymbol{u} \in \mathbb{R}^m$. We have

$$\begin{aligned} h_{\boldsymbol{u}}(x) - h_{\boldsymbol{u}}(y) =& \text{tr}([I_m + x\nabla\boldsymbol{\psi}(\text{prox}_x(\rho)(\boldsymbol{u}))]^{-1}) - \text{tr}([I_m + y\nabla\boldsymbol{\psi}(\text{prox}_y(\rho)(\boldsymbol{u}))]^{-1}) \\ \leq& \text{tr}(y\nabla\boldsymbol{\psi}(\text{prox}_y(\rho)(\boldsymbol{u})) - x\nabla\boldsymbol{\psi}(\text{prox}_x(\rho)(\boldsymbol{u}))) \qquad \text{(B.56)} \\ =& \text{tr}((y-x)\nabla\boldsymbol{\psi}(\text{prox}_x(\rho)(\boldsymbol{u})) - y[\nabla\boldsymbol{\psi}(\text{prox}_x(\rho)(\boldsymbol{u})) - \nabla\boldsymbol{\psi}(\text{prox}_y(\rho)(\boldsymbol{u}))]) \\ =& (y-x)\text{tr}(\nabla\boldsymbol{\psi}(\text{prox}_x(\rho)(\boldsymbol{u}))) - y\text{tr}(\nabla\boldsymbol{\psi}(\text{prox}_x(\rho)(\boldsymbol{u})) - \nabla\boldsymbol{\psi}(\text{prox}_y(\rho)(\boldsymbol{u}))). \end{aligned}$$

Since $A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}$, $\mathrm{tr(A^{-1} - B^{-1})} \leq \mathrm{tr(B - A)}$ upon taking $A = I_m + x\nabla\boldsymbol{\psi}(\mathrm{prox}_x(\rho)(\boldsymbol{u})), B = I_m + y\nabla\boldsymbol{\psi}(\mathrm{prox}_y(\rho)(\boldsymbol{u}))$. The inequality (B.56) thus follows. By the mean value theorem, $\mathrm{prox}_x(\rho)(\boldsymbol{u}) - \mathrm{prox}_y(\rho)(\boldsymbol{u}) = -(I_m + z\nabla\boldsymbol{\psi}(\mathrm{prox}_z(\rho)(\boldsymbol{u})))^{-1}\boldsymbol{\psi}(\mathrm{prox}_z(\rho)(\boldsymbol{u}))(x - y)$, where $z \in (\min\{x, y\}, \max\{x, y\})$. It thus follows that

$$\begin{aligned}
&\|\nabla\boldsymbol{\psi}(\mathrm{prox}_x(\rho)(\boldsymbol{u})) - \nabla\boldsymbol{\psi}(\mathrm{prox}_y(\rho)(\boldsymbol{u}))\|_{\max} \\
\leq & c\|\mathrm{prox}_x(\rho)(\boldsymbol{u}) - \mathrm{prox}_y(\rho)(\boldsymbol{u})\| \\
= & c\| - [I_m + z\nabla\boldsymbol{\psi}(\mathrm{prox}_z(\rho)(\boldsymbol{u}))]^{-1}\boldsymbol{\psi}(\mathrm{prox}_z(\rho)(\boldsymbol{u}))(x - y)\| \\
\leq & c|x - y|\|\boldsymbol{\psi}(\mathrm{prox}_z(\rho)(\boldsymbol{u}))\|.
\end{aligned}$$

For all $\boldsymbol{u} \in \mathbb{R}^m$, we obtain

$$\begin{aligned}
&\sup_{(x,y):|x-y|\leq\eta, 0\leq x\leq b, 0\leq y\leq b} |h_{\boldsymbol{u}}(x) - h_{\boldsymbol{u}}(y)| \\
\leq & |x - y|\mathrm{tr}(\nabla\boldsymbol{\psi}(\mathrm{prox}_x(\rho)(\boldsymbol{u}))) + y\mathrm{tr}(c|x - y|\|\boldsymbol{\psi}(\mathrm{prox}_x(\rho)(\boldsymbol{u}))\|I_m) \\
\leq & m\eta \sup \lambda_{\max}(\nabla\boldsymbol{\psi}(\mathrm{prox}_x(\rho)(\boldsymbol{u}))) + \eta mbc \sup \|\boldsymbol{\psi}(\mathrm{prox}_z(\rho)(\boldsymbol{u}))\| \\
\leq & m\eta(1 + b)c.
\end{aligned}$$

Therefore, we have

$$\sup_{(x,y):|x-y|\leq\eta, 0\leq x\leq b, 0\leq y\leq b} |g_n(x) - g_n(y)| \leq \frac{1}{n}\sum_{i=1}^{n} \sup_{x,y} |h_{\boldsymbol{u}_i}(x) - h_{\boldsymbol{u}_i}(y)| \leq \eta(1 + b)c.$$

Lemma B.8 *For any given $x_0 \leq b < \infty$, we have, $g_n(x_0) - E(g_n(x_0)) = o_{L_2}(1)$, and $E(\sup_{0 \leq x \leq b} |g_n(x) - E(g_n(x))|) = o(1)$.*

The lemma can be proved by following the proof of Lemma 3.25 in El Karoui (2018). We omit the details.

C Appendix

This part contains the proofs of lemmas and theorems in Chapter 4. We remark that the proofs of Theorem 4.1 and Theorems 4.3-4.4 mimic the proofs of Sgouropoulos et al. (2015). These proofs are all under those assumptions given in Chapter 4.

Proof of Lemma 4.1 We use the mathematical induction method to prove (4.16). When $n = 2$, denote $e_{(i)} = |a_{(i)} - b_{(i)}|$, and $e_i = |a_i - b_i|$, $i = 1, 2$. By the assumption (A1), we need to show that

$$\rho(e_{(1)}) + \rho(e_{(2)}) \leq \rho(e_1) + \rho(e_2). \tag{C.1}$$

Without loss of generality, suppose that $e_1 \geq e_2$. Then $e_1 \geq \max\{e_{(1)}, e_{(2)}\}$, and $e_2 \in [0, \min\{e_{(1)}, e_{(2)}\}]$ or $e_2 \in (\min\{e_{(1)}, e_{(2)}\}, \max\{e_{(1)}, e_{(2)}\}]$ or $e_2 \in (\max\{e_{(1)}, e_{(2)}\}, e_1]$. By the convexity of $\rho(\cdot)$, (C.1) holds for any setting of e_2.

Next, assume that (4.16) is true for all $k < n$. Let

$$\sum_{i=1}^{k} \rho(a_{(i)} - b_{(i)}) \leq \sum_{i=1}^{k} \rho(a_i - b_i). \tag{C.2}$$

Then we need to show that (4.16) still holds for $k+1$,

$$\sum_{i=1}^{k+1} \rho(a_{(i)} - b_{(i)}) \leq \sum_{i=1}^{k+1} \rho(a_i - b_i). \tag{C.3}$$

Without loss of generality, we assume that

$$a_{(k+1)} = a_{k+1} \text{ and } b_{(k+1)} = b_l,\ l = 1, 2, \dots, k+1. \tag{C.4}$$

(1) If $l = k+1$, then

$$\begin{aligned}
\sum_{i=1}^{k+1} \rho(a_i - b_i) &= \sum_{i=1}^{k} \rho(a_i - b_i) + \rho(a_{k+1} - b_{k+1}) \\
&\geq \sum_{i=1}^{k} \rho(a_{(i)} - b_{(i)}) + \rho(a_{k+1} - b_{k+1}) \\
&= \sum_{i=1}^{k+1} \rho(a_{(i)} - b_{(i)}).
\end{aligned}$$

The first inequality can be obtained by (C.2). The second equality is guaranteed by (C.4).

(2) If $l \neq k+1$,

$$\begin{aligned}
\sum_{i=1}^{k+1} \rho(a_i - b_i) &= \sum_{i=1, i\neq l}^{k} \rho(a_i - b_i) + \rho(a_l - b_l) + \rho(a_{k+1} - b_{k+1}) \\
&\geq \sum_{i=1, i\neq l}^{k} \rho(a_i - b_i) + \rho(a_l - b_{k+1}) + \rho(a_{k+1} - b_l) \\
&\geq \sum_{i=1}^{k} \rho(a_{(i)} - b_{(i)}) + \rho(a_{k+1} - b_l) \\
&= \sum_{i=1}^{k+1} \rho(a_{(i)} - b_{(i)}).
\end{aligned}$$

By (C.4), the first inequality can be obtained by following the proof for $n = 2$. The second inequality is guaranteed by (C.2). And the second equality is implied by (C.4). Based on the above discussion, (C.3) holds. Hence the result (4.16) follows.

Proof of Theorem 4.1. Since $S_n(\boldsymbol{\beta}) \geq 0$, by Monotone Convergence Theorem, we only need to show $S_n^{k+1}(\hat{\boldsymbol{\beta}}^{(k+1)}) \leq \; S_n^k(\hat{\boldsymbol{\beta}}^{(k)})$, $k = 1, 2, \ldots$.

$$S_n^k(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \rho(Y_{(i)} - \boldsymbol{\beta}^\top \boldsymbol{X}_{(i)}^{(k-1)}) + \lambda \sum_{j=1}^{p} \omega_j |\beta_j|. \tag{C.5}$$

For $k+1$, it follows from that

$$\begin{aligned} S_n^{k+1}(\hat{\boldsymbol{\beta}}^{(k+1)}) &= \frac{1}{n}\sum_{i=1}^{n} \rho(Y_{(i)} - (\hat{\boldsymbol{\beta}}^{(k+1)})^\top \boldsymbol{X}_{(i)}^{(k)}) + \lambda \sum_{j=1}^{p} \omega_j |\hat{\beta}_j^{(k+1)}| \\ &\leq \frac{1}{n}\sum_{i=1}^{n} \rho(Y_{(i)} - (\hat{\boldsymbol{\beta}}^{(k)})^\top \boldsymbol{X}_{(i)}^{(k)}) + \lambda \sum_{j=1}^{p} \omega_j |\hat{\beta}_j^{(k)}| \\ &\leq \frac{1}{n}\sum_{i=1}^{n} \rho(Y_{(i)} - (\hat{\boldsymbol{\beta}}^{(k)})^\top \boldsymbol{X}_{(i)}^{(k-1)}) + \lambda \sum_{j=1}^{p} \omega_j |\hat{\beta}_j^{(k)}| \\ &= S_n^k(\hat{\boldsymbol{\beta}}^{(k)}), \end{aligned}$$

where $\{\boldsymbol{X}_{(i)}^{(k)}\}$ is a permutation of $\{\boldsymbol{X}_i\}$ at the kth iteration such that $(\hat{\boldsymbol{\beta}}^{(k)})^\top \boldsymbol{X}_{(1)}^{(k)} \leq (\hat{\boldsymbol{\beta}}^{(k)})^\top \boldsymbol{X}_{(2)}^{(k)} \leq \cdots \leq (\hat{\boldsymbol{\beta}}^{(k)})^\top \boldsymbol{X}_{(n)}^{(k)}$. The first inequality follows from the definition of $\hat{\boldsymbol{\beta}}^{(k+1)}$. The second inequality is guaranteed by Lemma 4.1. Hence $\lim_{k\to+\infty} S_n^k(\hat{\boldsymbol{\beta}}^{(k)})$ exists, which leads to the convergence of the algorithm.

Proof of Lemma 4.2 Put $W = \boldsymbol{\beta}^\top \boldsymbol{X}$, $Q_n(\alpha) = Q_{n,Y}(\alpha) - Q_{n,W}(\alpha), Q(\alpha) = Q_Y(\alpha) - Q_W(\alpha)$, $\alpha \in \Omega_n = [a_n, b_n]$. By (A2), $S(\boldsymbol{\beta}) = S(\boldsymbol{\beta}; a_n, b_n) + o_p(n^{-\tau_0})$,

$S_n(\boldsymbol{\beta}) = S_n(\boldsymbol{\beta}; n_1, n_2) + o_p(n^{-\tau_0})$. Then $n^{\tau_0}\{S_n(\boldsymbol{\beta}) - S(\boldsymbol{\beta})\} = n^{\tau_0}\{S_n(\boldsymbol{\beta}; n_1, n_2) - S(\boldsymbol{\beta}; a_n, b_n)\} + o_p(1)$. If the following

$$n^{\tau_0}\left\{\left|\frac{1}{n}\sum_{i=1}^{n}\rho(Q_n(i/n)) - \frac{1}{n}\sum_{i=n_1+1}^{n_2}\rho(Q(i/n))\right|\right\} = o_p(1), \tag{C.6}$$

and

$$\frac{1}{n}\sum_{i=n_1+1}^{n_2}\rho(Q(i/n)) = \int_{a_n}^{b_n}\rho(Q(\alpha))d\alpha + o(1/n), \tag{C.7}$$

hold, then (4.18) follows. Note that a_n, b_n, n_1, n_2 are already defined in Remark 4.1.

First, we show that (C.6) hold under the assumptions.

$$\begin{aligned}
& n^{\tau_0}\left|\frac{1}{n}\sum_{i=n_1+1}^{n_2}\rho(Q_n(i/n)) - \frac{1}{n}\sum_{i=n_1+1}^{n_2}\rho(Q(i/n))\right| \\
\leq & n^{\tau_0} M\frac{1}{n}\sum_{i=n_1+1}^{n_2}\left|\{Q_{n,Y}(i/n) - Q_Y(i/n)\} - \{Q_{n,W}(i/n) - Q_W(i/n)\}\right| \qquad \text{(C.8)} \\
\leq & n^{\tau_0} M\frac{1}{n}\sum_{i=n_1+1}^{n_2}\left(\frac{|F_{n,Y}(Q_Y(i/n)) - i/n|}{f_Y(Q_Y(i/n))} + \frac{|F_{n,W}(Q_W(i/n)) - i/n|}{f_W(Q_W(i/n))}\right) \qquad \text{(C.9)} \\
& + n^{\tau_0} M\frac{1}{n}\sum_{i=n_1+1}^{n_2}\left(\frac{R_n(i/n)/\sqrt{n}}{f_Y(Q_Y(i/n))} + \frac{R_n(i/n)/\sqrt{n}}{f_W(Q_W(i/n))}\right). \qquad \text{(C.10)}
\end{aligned}$$

(C.8) follows from the Lipschitz continuity of ρ on Ω_n (see (A1)). By (4.14), we obtain (C.9) and (C.10). By the Dvoretzky-Kiefer-Wolfowitz inequality, for any constant $\epsilon > 0$ and any integer $n > 0$ it holds that

$$P\left\{\sup_{0\leq\alpha\leq 1}|F_{n,Y}(Q_Y(\alpha)) - \alpha| > \epsilon\right\} \leq 2e^{-2n\epsilon^2},$$

$$P\left\{\sup_{0\leq\alpha\leq 1}|F_{n,W}(Q_W(\alpha)) - \alpha| > \epsilon\right\} \leq 2e^{-2n\epsilon^2}.$$

Let $\epsilon = n^{-\tau_2}$ for some $\tau_2 \in (\tau_1, 1/2)$, and put

$$\mathcal{A}_n = \left\{ \sup_{0\le\alpha\le 1} |F_{n,Y}(Q_Y(\alpha)) - \alpha| \le \epsilon \right\} \cap \left\{ \sup_{0\le\alpha\le 1} |F_{n,W}(Q_W(\alpha)) - \alpha| \le \epsilon \right\}.$$

Then $P(\mathcal{A}_n) \ge 1 - 4e^{-2n\epsilon^2} \to 1$ as $n \to \infty$. Let $\alpha_i = i/n$. Since $\inf_{Q_\xi(\alpha)\in\Omega_n} f_\xi(Q_\xi(\alpha)) = n^{-(\tau_1-\tau_0)}$ (by (A2)), (C.9) $\xrightarrow{p} 0$, and (C.10) $\xrightarrow{a.s.} 0$. Thus (C.6) follows.

Second, by the assumptions (A1) and (A2), $\rho(\cdot)$ is a continuous convex discrepancy function, and $Q_Y(\alpha)$ and $Q_W(\alpha)$ are continuous on Ω_n, which guarantees that the existence of $\int_{a_n}^{b_n} \rho(Q(\alpha))d\alpha$. By Taylor expansion, $|Q_\xi(\alpha) - Q_\xi(i/n)| = n^{-1}\{f_\xi(Q_\xi(i/n))\}^{-1}(1 + o(1))$ for any $|\alpha - i/n| \le 1/n$. Hence, (C.7) holds. This completes the proof.

Proof of Theorem 4.2. Since $\mathcal{B}$ is a compact set, there exists a finite number of points $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_H$ in $\mathcal{B}$ such that

$$b_1(\boldsymbol{\beta}_1, r_1) \cup \cdots \cup b_H(\boldsymbol{\beta}_H, r_H) \supseteq \mathcal{B},$$

where $b_h(\boldsymbol{\beta}_h, r_h)$ denotes the sphere with center $\boldsymbol{\beta}_h$ and radius $r_h,\ h = 1, 2, \ldots, H$. We thus have

$$\sup_{\boldsymbol{\beta}\in\mathcal{B}} |S_n(\boldsymbol{\beta}) - S(\boldsymbol{\beta})| \le \sum_{h=1}^{H} \sup_{\boldsymbol{\beta}\in R_h} |S_n(\boldsymbol{\beta}) - S(\boldsymbol{\beta})|,$$

where $R_h = b_h(\boldsymbol{\beta}_h, r_h) \cap \mathcal{B}, h = 1, 2, \ldots, H$. Then (4.17) can be obtained from the following result

$$\sup_{\boldsymbol{\beta}\in R_h} |S_n(\boldsymbol{\beta}) - S(\boldsymbol{\beta})| \xrightarrow{p} 0, \text{ as } n \to \infty, \tag{C.11}$$

for any h such that $r_h \to 0$. As $r_h \to 0$, $\boldsymbol{\beta} \to \boldsymbol{\beta}_h$.

Next, we show that (C.11) holds true. For $\forall\ \boldsymbol{\beta} \in R_h$,

$$|S_n(\boldsymbol{\beta}) - S(\boldsymbol{\beta})| \le \underbrace{|S_n(\boldsymbol{\beta}) - S_n(\boldsymbol{\beta}_h)|}_{I} + \underbrace{|S_n(\boldsymbol{\beta}_h) - S(\boldsymbol{\beta}_h)|}_{II} + \underbrace{|S(\boldsymbol{\beta}_h) - S(\boldsymbol{\beta})|}_{III}.$$

We first consider (III). Since $S(\boldsymbol{\beta})$ is continuous, by the continuous mapping theorem, it follows that

$$|S(\boldsymbol{\beta}_h) - S(\boldsymbol{\beta})| \to 0 \ \text{ as } r_h \to 0. \tag{C.12}$$

We next evaluate (II). By Lemma 4.2 with $\tau = 0$, we have

$$|S_n(\boldsymbol{\beta}_h) - S(\boldsymbol{\beta}_h)| \xrightarrow{p} 0 \text{ , as } \ n \to \infty. \tag{C.13}$$

To study (I), we put $W = \boldsymbol{\beta}^\top \boldsymbol{X},\ W_h = \boldsymbol{\beta}_h^\top \boldsymbol{X}$. Then we have

$$\begin{aligned}
&|S_n(\boldsymbol{\beta}) - S_n(\boldsymbol{\beta}_h)| \\
&= \left| \frac{1}{n} \sum_{i=n_1+1}^{n_2} \{\rho(Q_{n,Y}(i/n) - Q_{n,W}(i/n)) - \rho(Q_{n,Y}(i/n) - Q_{n,W_h}(i/n))\} + \lambda(\|\boldsymbol{\omega}^\top \boldsymbol{\beta}\|_1 - \|\boldsymbol{\omega}^\top \boldsymbol{\beta}_h\|_1) \right| \\
&\le \left| \frac{1}{n} \sum_{i=n_1+1}^{n_2} \{\rho(Q_Y(i/n) - Q_W(i/n)) - \rho(Q_Y(i/n) - Q_{W_h}(i/n))\} \right| + o_p(1) + \lambda \|\boldsymbol{\omega}\|_1 \|\boldsymbol{\beta} - \boldsymbol{\beta}_h\|_1 \\
&\le \left| \int_{a_n}^{b_n} \{\rho(Q_Y(\alpha) - Q_W(\alpha)) - \rho(Q_Y(\alpha) - Q_{W_h}(\alpha))\}\, d\alpha \right| + o_p(1) + r_n + \lambda \|\boldsymbol{\omega}\|_1 \|\boldsymbol{\beta} - \boldsymbol{\beta}_h\|_1 \\
&\le M \int_{a_n}^{b_n} |Q_W(\alpha) - Q_{W_h}(\alpha)| d\alpha + \lambda \|\boldsymbol{\omega}\|_1 \|\boldsymbol{\beta} - \boldsymbol{\beta}_h\|_1 + o_p(1) + r_n \\
&\le M\epsilon + \lambda \|\boldsymbol{\omega}\|_1 \epsilon + o_p(1) + r_n.
\end{aligned}$$

By (C.6), there exists a set $\mathcal{A}_n$ with $P(\mathcal{A}_n) \to 1$ such that the first inequality follows. The second inequality can be obtained by (C.7), where $r_n \to 0$. In view of the

Lipschitz continuity of ρ, the third inequality holds. Since $\boldsymbol{\beta} \to \boldsymbol{\beta}_h$, the distribution of W approximates to the distribution of W_h. For any $\epsilon > 0$, $|Q_W(\alpha) - Q_{W_h}(\alpha)| \leq \epsilon$, which guarantees the last inequality. We thus obtain that

$$|S_n(\boldsymbol{\beta}) - S_n(\boldsymbol{\beta}_h)| \xrightarrow{p} 0, \text{ as } n \to \infty. \tag{C.14}$$

By (C.12), (C.13) and (C.14), (C.11) holds. This completes the proof.

Proof of Theorem 4.3. As $\boldsymbol{X}, Y$ are bounded, then the matching quantiles M-estimate $\hat{\boldsymbol{\beta}}_n$ is also bounded. Let $\mathcal{B}$ be a compact set that contains $\boldsymbol{\beta}^0$ and $\hat{\boldsymbol{\beta}}_n$ with probability 1. Then it follows that

$$S_n(\boldsymbol{\beta}^0) - S(\boldsymbol{\beta}^0) \geq S_n(\hat{\boldsymbol{\beta}}_n) - S(\boldsymbol{\beta}^0) \geq S_n(\hat{\boldsymbol{\beta}}_n) - S(\hat{\boldsymbol{\beta}}_n).$$

The first and second inequalities follow from the definition of $\hat{\boldsymbol{\beta}}_n$ and $\boldsymbol{\beta}^0$, respectively. By Lemma 4.2, both $S_n(\boldsymbol{\beta}^0) - S(\boldsymbol{\beta}^0)$ and $S_n(\hat{\boldsymbol{\beta}}_n) - S(\hat{\boldsymbol{\beta}}_n)$ converge to 0 in probability. Hence $S_n(\hat{\boldsymbol{\beta}}_n) - S(\boldsymbol{\beta}^0) \xrightarrow{p} 0$.

Proof of Theorem 4.4. We use the contraction method to prove the theorem. Suppose there exists an $\varepsilon_0 > 0$, such that

$$\limsup_{n \to \infty} P\{d(\hat{\boldsymbol{\beta}}_n, \mathcal{B}_0) \geq \varepsilon_0\} > 0.$$

Then $\exists\ \delta_1 > 0$ and an integer subsequence $\{n_k\} \subset \{n\}$ such that $\lim_{k \to \infty} P(\mathcal{A}_k) = \delta_1 > 0$, where $\mathcal{A}_k = \{\hat{\boldsymbol{\beta}}_{n_k} : d(\hat{\boldsymbol{\beta}}_{n_k}, \mathcal{B}_0) \geq \varepsilon_0\}$. Define a subset $\mathcal{D}_1$ which is ε_0-distance from $\mathcal{B}_0$. Let $\mathcal{D}_1 = \{\boldsymbol{\beta} \in \mathcal{B} : d(\boldsymbol{\beta}, \mathcal{B}_0) \geq \varepsilon_0\}$. Then $\mathcal{D}_1$ is also a compact set, and $\mathcal{A}_k \subset \mathcal{D}_1$.

By the definition of $\mathcal{B}_0$, for $\forall\, \boldsymbol{\beta}^0 \in \mathcal{B}_0$, there exists a $\delta_2 > 0$ such that $\inf_{\boldsymbol{\beta} \in \mathcal{D}_1} S(\boldsymbol{\beta}) = \delta_2 + S(\boldsymbol{\beta}^0)$. Next define $\mathcal{B}_k = \{\hat{\boldsymbol{\beta}}_{n_k} : |S_{n_k}(\hat{\boldsymbol{\beta}}_{n_k}) - S(\hat{\boldsymbol{\beta}}_{n_k})| < \delta_2/2\}$, then $P(\mathcal{B}_k) \to 1$ (by Lemma 4.2).

Then on the set $\mathcal{A}_k \cap \mathcal{B}_k (\neq \emptyset)$, we have

$$S_{n_k}(\hat{\boldsymbol{\beta}}_{n_k}) \geq S(\hat{\boldsymbol{\beta}}_{n_k}) - \delta_2/2 \geq \inf_{\boldsymbol{\beta} \in \mathcal{D}_1} S(\boldsymbol{\beta}) - \delta_2/2 = S(\boldsymbol{\beta}^0) + \delta_2/2 > S(\boldsymbol{\beta}^0).$$

It contradicts to the fact that $S_n(\hat{\boldsymbol{\beta}}_n) \to S(\boldsymbol{\beta}^0)$ in probability (by Theorem 4.3). We complete the proof.