DISSERTATION

SOME TOPICS IN HIGH-DIMENSIONAL ROBUST INFERENCE AND GRAPHICAL MODELING

Submitted by

Youngseok Song

Department of Statistics

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Summer 2021

Doctoral Committee:

Advisor: Wen Zhou

Jay Breidt
Dan Cooley
Kim Hoke

ABSTRACT

SOME TOPICS IN HIGH-DIMENSIONAL ROBUST INFERENCE AND GRAPHICAL MODELING

In this dissertation, we focus on large-scale robust inference and high-dimensional graphical modeling. Especially, we study three problems: a large-scale inference method by a tail-robust regression, model specification tests for dependence structure of Gaussian Markov random fields, and a robust Gaussian graph estimation.

First of all, we consider the problem of simultaneously testing a large number of general linear hypotheses, encompassing covariate-effect analysis, analysis of variance, and model comparisons. The new challenge that comes along with the overwhelmingly large number of tests is the ubiquitous presence of heavy-tailed and/or highly skewed measurement noise, which is the main reason for the failure of conventional least squares based methods. The new testing procedure is built on data-adaptive Huber regression, and a new covariance estimator of the regression estimate. Under mild conditions, we show that the proposed methods produce consistent estimates of the false discovery proportion. Extensive numerical experiments, along with an empirical study on quantitative linguistics, demonstrate the advantage of our proposal compared to many state-of-the-art methods when the data are generated from heavy-tailed and/or skewed distributions.

In the next chapter, we focus on the Gaussian Markov random fields (GMRFs) and, by utilizing the connection between GMRFs and precision matrices, we propose an easily implemented procedure to assess the spatial structures modeled by GMRFs based on spatio-temporal observations. The new procedure is flexible to assess a variety of structures including the isotropic and directional dependence as well as the Matérn class. A comprehensive simulation study has been conducted to demonstrate the finite sample performance of the procedure. Motivated from the ef-

forts on modeling flu spread across the United States, we also apply our method to the Google Flu Trend data and report some very interesting epidemiological findings.

Finally, we propose a high-dimensional precision matrix estimation method via nodewise distributionally robust regressions. The distributionally robust regression with an ambiguity set defined by Wasserstein-2 ball has a computationally tractable dual formulation, which is linked to square-root regressions. We propose an iterative algorithm that has a substantial advantage in terms of computation time. Extensive numerical experiments study the performance of the proposed method under various precision matrix structures and contamination models.

ACKNOWLEDGEMENTS

First of all, I would like to thank my advisor, Professor Wen Zhou, for his support and patience over the past years. His endless passion for excellent research, great ideas from his insight, and research advice from his expertise have inspired me to pursue interesting research topics and helped me to overcome challenges and difficulties in research. I always want to be a statistician like him, who has great strength in theory and application sides. He is not just an academic advisor but also a great mentor.

Besides, I would like to thank my committee members: Prof. Jay Breidt, Prof. Dan Cooley, and Prof. Kim Hoke, for their invaluable insight and expertise to improve my dissertation. I have learned much from discussions with Prof. Breidt and Prof. Cooley and their courses. Prof. Hoke has provided me generous support to work on her research project.

I would like to thank my collaborators, Prof. Jinyuan Chang, Prof. Yumou Qiu, Prof. Wen-Xin Zhou, for their support and patience. I also thank Prof. Eva K. Fischer and Prof. Kimberly A. Hughes for the project opportunity to answer biological questions in evolution.

Thanks to my friends and colleagues both inside and outside of CSU statistics department. Our discussions have been great motivations and helped me to think outside of the box when I faced challenges. Especially, thanks to Dr. Joshua Hewitt for his introduction to `R` parallel computing, which allow me to save an enormous amount of computation time. Thanks to Dr. Kyungtae Kim and Dr. Myungjoo Shin for priceless discussions, which remind me about the importance of applications.

I thank my parents who have dedicated their lives to raising me. No words can express my gratitude to them. I also thank my parents-in-law for their supports. I also thank to my sister and her family.

Last but not least, special thanks to my spouse, Mihyun Kim. She is not just my life companion but also a great researcher. Without her support and love, I did not start my Ph.D., and I could not have finished this long journey. I dedicate this work to her.

DEDICATION

I would like to dedicate this thesis to my life partner, Mihyun.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

In this chapter, we review the latest development on robust inference and Gaussian graphical modeling, which serve as the cornerstone of this dissertation.

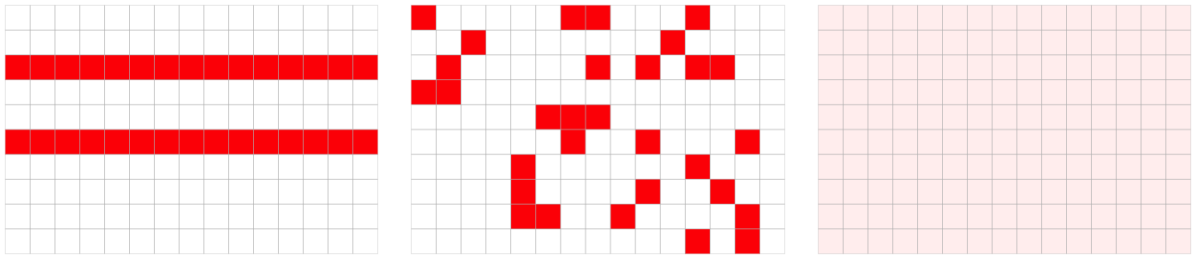
1.1 Robust Statistics

All statistical procedures rely on assumptions about data generation distributions such as moments, symmetry, as well as independence, homogeneity, and so on. When the assumptions are violated, the performance of statistical procedures are usually impaired and result in spurious discoveries and false conclusions. From the seminal work by Huber [1], statisticians have been exploring *robustness* of statistical methods for more than a half century, where the robustness is defined as the insensitivity of statistical methods against deviations between the model assumptions and data generation mechanisms [2]. The *distributional robustness*, which assumes that the true underlying data generation distribution deviates slightly from the assumptions of statistical models or methods, has been a primary focus in traditional robust statistics [3]. A few different notions of robustness have been employed in statistics, optimization, and machine learning.

Huber's ϵ-contamination model (or contamination model) is of the primary interest in the traditional robust statistics [1]. Consider the class of data generation distributions

$$\mathcal{F}_\epsilon = (1-\epsilon)F + \epsilon G \tag{1.1.1}$$

where F and G are both unknown distributions, and $\epsilon \in (0,1)$ models the contamination fraction. The observations are assumed to be random samples from F_ϵ. Under model (1.1.1), the robustness is measured by the *breakdown point* [4] or the *influence function* [5]. Intuitively, the breakdown point of an estimator is the proportion of contamination that the estimator can tolerate yet still provides a reasonably accurate estimate of the underlying true distribution, F in (1.1.1), and its



(a) Rowwise contamination **(b)** Cellwise contamination **(c)** Distributional perturbation

Figure 1.1: Demonstrations of each type of contaminations. Each row represents a multivariate sample. The blank cells are samples from the true underlying distribution while the red ones are those from the arbitrary contaminated distribution. The light pink cells in panel (c) are samples from distributions that are slighlty different to the true underlying distribution, and compared to panels (a) and (b), all data in (c) are contaminated.

functional. For example, the breakdown point of the median is 0.5, whereas the breakdown point of the sample mean is 0. The breakdown point cannot exceed 0.5 otherwise it is not possible to distinguish the underlying true and contamination distributions. In contrast, the influence function quantifies a limiting influence of a new observation with value x on a statistical procedure. Let $T(F)$ be a parameter of interest for the underlying true distribution F. The influence function with respect to $T(F)$ is defined by

$$IC(x, F, T) := \lim_{t \to 0+} \frac{T\{(1-t)F + t\delta_x\} - T(F)}{t}, \tag{1.1.2}$$

where δ_x is the probability measure with $\mathbb{P}(X = x) = 1$. The gross error sensitivity, $\gamma^* := \sup_x |IC(x, F, T)|$, is then defined for quantifying the robustness [5]. If γ^* is unbounded, the statistical procedure will fail in the presence of an arbitrarily large outlier. For example, consider the location estimation problem in $\mathbb{R}$. The influence function of the sample mean is $IC(x, F, T) = x - \mu$ where μ is the mean of F, which diverges as $|x| \to \infty$. On the other hand, the influence function of the median is $\text{sgn}\{x - F^{-1}(0.5)\}/[2f_X\{F^{-1}(0.5)\}]$. Both the breakdown point and gross error sensitivity (or the influence function) therefore characterize sort of robustness of statistical methods against outliers. For the finite sample version of (1.1.2) such as the sensitivity curve and the jackknife, we refer to [2].

Cell-wise contamination model is a multivariate generalization of the traditional contamination model [6]. Let $\mathbf{X}_i \in \mathbb{R}^p$ be an observed vector, $\mathbf{Y}_i$ be a random sample from the unobserved true distribution, and $\mathbf{Z}_i$ be an outlier generated from the contamination distribution. In addition, denote $\mathbf{B}_i = diag(B_{i1}, \ldots, B_{ip})$ a diagonal matrix whose entries are Bernoulli random variables with $\mathbb{P}(B_{ij} = 1) = \epsilon_j$ for $j = 1, \ldots, p$. Assume $\mathbf{Y}_i$, $\mathbf{Z}_i$, and $\mathbf{B}_i$ are independent. The cell-wise contamination model is defined by

$$\mathbf{X}_i = (\mathbf{I} - \mathbf{B}_i)\mathbf{Y}_i + \mathbf{B}_i\mathbf{Z}_i, \ \ \forall i = 1, \ldots, n. \tag{1.1.3}$$

Model (1.1.3) is called the *fully independent contamination model* when $B_{i1}, \ldots, B_{ip}$ are independent and the *fully dependent contamination model* when $\mathbb{P}(B_{i1} = \cdots = B_{ip}) = 1$. Alternatively, they are also referred as the *cellwise contamination model* and the *rowwise contamination model*, respectively. Figure 1.1 (a)–(b) illustrate the two contamination settings when there are 20% of contaminated samples or cells.

Tail-robustness is a lately developed notion to study the insensitivity of estimators or statistical procedures against the heavy-tailedness and/or skewness of the observations. The performance of majority of statistical estimators or procedures heavily relies on distributional assumptions of data such as the normality/sub-Gaussianity, the symmetry of underlying error distributions, etc. However, when these assumptions are violated, which is not uncommon in practice [7], the performance of most widely-used statistical methods, such as least squares-based methods, would be impaired severely by observations that are rarely observed in lighter-tail distributions. We call these observations as stochastic outliers. Therefore, we need to develop estimators and procedures that are robust against the stochastic outliers, which have better finite-sample performance than non-robust methods. We call them *tail-robust estimators* and *tail-robust procedures*, respectively. To study finite-sample performance, the theoretical studies on tail-robust estimators focus on nonasymptotic deviation bounds under weak moment assumptions. Following the pioneering work in [8], the estimation theories under heavy-tailed models have been developed for mean estimation [9, 10], covariance estimation [11], and regressions [12]. Also, a few inference procedures

that enjoy tail-robustness for testing hypotheses on high-dimensional mean vectors have been documented in [7, 13, 14].

Finally, we introduce the *distributionally robust optimization*. Most statistical estimators can be formulated by minimizing the empirical risk, that is

$$\widehat{\theta} := \operatorname*{argmin}_{\theta} \mathbb{E}_{\widehat{\mathbb{P}}} \ell_\theta(x) = \operatorname*{argmin}_{\theta} \frac{1}{n} \sum_{i=1}^{n} \ell_\theta(x_i) \tag{1.1.4}$$

for some loss function $\ell(\cdot)$ and empirical distribution $\widehat{\mathbb{P}}$. Formulation in (1.1.4) provides an estimator that performs well on the training data sampled from a common distribution, but it may perform poorly on out-of-sample data. This phenomenon, *overfitting*, is widely observed when one trains statistical learning models, such as least squares-based method or simple tree models. To overcome the overfitting issue, the (data-driven) *distributionally robust optimizations* consider the following problem:

$$\widehat{\theta} := \operatorname*{argmin}_{\theta} \sup_{\mathbb{Q} \in \mathcal{U}} \mathbb{E}_{\mathbb{Q}} \ell_\theta(x), \tag{1.1.5}$$

where $\mathcal{U}$ is a set of distributions that is referred as the ambiguity set, which is specified with respect to the empirical distribution $\widehat{\mathbb{P}}$. Minimizing the worst-case risk, $\sup_{\mathbb{Q} \in \mathcal{U}} \mathbb{E}_{\mathbb{Q}} \ell_\theta(x)$, allows us to obtain an estimator that performs uniformly well over all probability distributions in $\mathcal{U}$. Conceptually, the data generation distribution is therefore allowed to be completely different from the model and the deviation is quantified by set $\mathcal{U}$. See Figure 1.1 (c) for an illustration. The formulation in (1.1.5) has been employed in machine learning, including the traditional regression models, the support vector machines, and the generative adversarial networks [15–21].

Directly solving (1.1.5) is often computationally intractable, so that the distributionally robust optimization focuses on developing equivalent and computationally tractable optimization problems using the duality argument. The dual formulation depends on the choice of the ambiguity set. Hence, a natural question from (1.1.5) is how to specify the ambiguity set $\mathcal{U}$. Two discrepancy measures, the f-divergence [22–24] and the Wasserstein distance [25], have been primarily used for defining the ambiguity set in the optimization literature. The *f-divergence* is defined by

$$D_f(\mathbb{Q}\|\mathbb{P}) := \int f\left(\frac{d\mathbb{Q}}{d\mathbb{P}}\right) d\mathbb{P} \tag{1.1.6}$$

for probability measures $\mathbb{P}$ and $\mathbb{Q}$, where $f : \mathbb{R} \to \mathbb{R}_+ \cup \{+\infty\}$ is a convex function satisfying $f(1) = 0$ and $f(t) = +\infty$ for any $t < 0$. It includes a large number of well-known information metrics, such as the Kullback-Leibler divergence, the Rényi divergence, the χ^2-divergence, the Hellinger distance, and so on. The f-divergence has been used to define the ambiguity set and the corresponding distributionally robust optimization problems in machine learning [19, 26–29]. On the other hand, the *Wasserstein distance* is defined by

$$W_p(\mathbb{P}, \mathbb{Q}) := \left[\inf_{\pi \in \Pi(\mathbb{P},\mathbb{Q})} \mathbb{E}_{(X,Y)\sim\pi}\{d(X, Y)\}^p\right]^{1/p}, \quad p \in [1, \infty) \tag{1.1.7}$$

for some metric $d(\cdot, \cdot)$ and $p \in [1, \infty)$ [25]. The Wasserstein distance is motivated from the optimal transport theories back in the 18th century, and the Wasserstein-distributionally robust optimization problems have been studied in both statistics and machine learning [16, 17, 20, 30–33]. We refer to [34] for more choices of ambiguity set in the distributionally robust optimization such as the moment based ambiguity sets, the shape-preserving models, and the Kernel-based models.

1.2 Gaussian Graphical Models

For a random vector $\mathbf{X} = (X_1, \ldots, X_p)^{\mathrm{T}} \in \mathbb{R}^p$, the probabilistic graphical models aim to capture the conditional dependence between (X_i, X_j) given all other entries of $\mathbf{X}$, denoted by $\mathbf{X}_{-(i,j)}$. If there exists conditional dependence between X_i and X_j given $\mathbf{X}_{-(i,j)}$, nodes i and j are connected by an edge. Assume that $\mathbf{X}$ is a multivariate normal random vector with covariance matrix $\mathbf{\Sigma}$, then its precision matrix $\mathbf{\Omega} := \mathbf{\Sigma}^{-1} = \{\omega_{ij}\}_{1\le i,j\le p}$ recovers the corresponding undirected graph of nodes $\{i : 1 \le i \le p\}$ using the support of ω_{ij}'s. Specifically, it holds that $\omega_{ij} \neq 0$ if and only if X_i and X_j are independent given $\mathbf{X}_{-(i,j)}$. Therefore, estimating the Gaussian graph is equivalent to identifying nonzero entries of corresponding $\mathbf{\Omega}$.

For the last decades, the Gaussian graphical model has been extensively studied and applied in practice. Given n independent and identically distributed (*i.i.d.*) observations from the Gaussian graphical model, in low-dimensional setting where $n \gg p$, the sample covariance matrix is a consistent estimator of covariance $\mathbf{\Sigma}$ so that the inverse of sample covariance matrix $\widehat{\mathbf{\Sigma}}^{-1}$ naturally estimates the precision matrix and recovers the underlying graph. However, in the high-dimensional regime where $p \gg n$, the sample covariance is not consistent, neither is it invertible. Additional structural assumptions on the graph, together with new estimators, are required to estimate the precision matrix and recover the underlying graph. For instance, the sparsity assumption, which requires the number of edges connecting nodes to be small or slowly growing in the number of nodes, is a natural and flexible structural condition to be imposed. By imposing this condition, estimators encouraging sparsity of $\mathbf{\Omega}$ provide consistent and sometimes optimal estimate to the precision matrix and therefore accurately recover the underlying graph. Along this line, three categories of estimations are prevalent in the literature and widely used in practice, including the neighborhood-based approach (or nodewise regression), the penalized likelihood estimations, and the constraint optimization.

The nodewise regression was first proposed by [35]. It is a neighborhood set estimation procedure using Lasso [36]. For each $j = 1, \ldots, p$, consider regression of each variable against all other variables

$$\widehat{\boldsymbol{\beta}}_j = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\operatorname{argmin}} \left(\frac{1}{2n} \|X_j - \mathbf{X}_{-j}\boldsymbol{\beta}\|_2^2 + \lambda_j \|\boldsymbol{\beta}\|_1 \right) \tag{1.2.1}$$

where $\mathbf{X}_{-j} = \{X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_p\}$. If $\widehat{\beta}_{j,k} \neq 0$, the kth node is in the neighborhood set of the jth node. Note that the estimated regression coefficients do not have to be symmetric, that is $\widehat{\beta}_{j,k} \neq 0$ does not imply $\widehat{\beta}_{k,j} \neq 0$. Although the original proposal by [35] does not provide the precision matrix estimate from the nodewise regression estimator, we can reconstruct it by using the blockwise matrix inversion formula. Define the jth column of estimator by

$$\widetilde{\mathbf{\Omega}}_j = \widehat{\mathbf{\Gamma}}_j / \widehat{\sigma}_j^2. \tag{1.2.2}$$

where $\widehat{\mathbf{\Gamma}}_j := (-\widehat{\beta}_{j,1}, \ldots, -\widehat{\beta}_{j,j-1}, 1, -\widehat{\beta}_{j,j+1}, \ldots, -\widehat{\beta}_{j,p})^T$ and $\widehat{\sigma}_j^2 = n^{-1}\|X_j - \mathbf{X}_{-j}\widehat{\boldsymbol{\beta}}_j\|_2^2$.

The graphical Lasso (Glasso), which is a penalized likelihood estimation [37–39], and the CLIME estimator [40], both estimate $\mathbf{\Omega}$ directly using optimization procedures involving sample covariance matrix $\widehat{\mathbf{\Sigma}}$. Motivated from the penalized log-likelihood over non-negative definite matrix, Glasso solves the following optimization problem

$$\widetilde{\mathbf{\Omega}} := \underset{\mathbf{\Omega}}{\operatorname{argmin}}[-\log\det\mathbf{\Omega} + \operatorname{Tr}(\widehat{\mathbf{\Sigma}}\mathbf{\Omega}) + \lambda\|\mathbf{\Omega}\|_1], \tag{1.2.3}$$

where λ is the tuning parameter, and (1.2.3) can be solved efficiently using the dual form, which can be further reduced into vector minimization problems [39]. A fast algorithm proposed by [41] is implemented in `glasso` R package, and [42, 43] presented `glassoFast` R package to improve the computational efficiency and resolve the non-termination issue in `glasso` package. In contrast, CLIME is based on the following Dantzig-type optimization problem

$$\widetilde{\mathbf{\Omega}} := \underset{\mathbf{\Omega}}{\operatorname{argmin}} \|\mathbf{\Omega}\|_1 \text{ subject to } \|\widehat{\mathbf{\Sigma}}\mathbf{\Omega} - \mathbf{I}\|_\infty \leq \lambda, \tag{1.2.4}$$

where λ is a tuning parameter. It is known that (1.2.4) is equivalent to p coordinate-wise optimization problems,

$$\widehat{\boldsymbol{\beta}}_j := \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \|\boldsymbol{\beta}\|_1 \text{ subject to } \|\widehat{\mathbf{\Sigma}}\boldsymbol{\beta} - \mathbf{e}_j\|_\infty \leq \lambda,$$

where $\mathbf{e}_j$ is j-th basis vector and $\widetilde{\mathbf{\Omega}} = \begin{bmatrix} \widehat{\boldsymbol{\beta}}_1 & \cdots & \widehat{\boldsymbol{\beta}}_p \end{bmatrix}$. Compared to Glasso, CLIME tends to recover more sparse graph. CLIME has been efficiently implemented in both `clime` and `fastclime` R packages [44].

We review two refinements to make the preliminary estimate $\widetilde{\mathbf{\Omega}}$ from (1.2.1)–(1.2.4) positive (semi-)definite. The one is replacing the off-diagonal entries by $\widehat{\omega}_{jk} = \widehat{\omega}_{kj} = \widetilde{\omega}_{jk}\mathbb{I}(|\widetilde{\omega}_{jk}| \leq |\widetilde{\omega}_{kj}|) + \widetilde{\omega}_{kj}\mathbb{I}(|\widetilde{\omega}_{jk}| > |\widetilde{\omega}_{kj}|)$ from [40], which makes $\widehat{\mathbf{\Omega}}$ be positive definite with high probability. The other method is to find nearest approximation to a set of positive definite matrices by solving an optimization problem $\operatorname{argmin}_{\mathbf{\Omega} \succ \mathbf{0}} \|\widetilde{\mathbf{\Omega}} - \mathbf{\Omega}\|$ where $\|\cdot\|$ is some appropriate matrix norm [45–47].

Inferential methods for high-dimensional Gaussian graphs have been developed over the past years. To recover the underlying graph and the connectivity among nodes, inference on the precision matrix has been considered such as the multiple testing problem with $p(p-1)/2$ hypotheses

$$H_{0,jk} : \omega_{j,k} = 0 \text{ v.s. } H_{1,jk} : \omega_{j,k} \neq 0 \tag{1.2.5}$$

for $1 \leq j < k \leq p$. [48] proposed a test statistic using the bias correction of the sample covariance of residuals from the nodewise regression estimator such as the Dantzig selector [49] or Lasso. Using the debiased testing statistics, the author also proposed a multiple testing procedure to control the false discovery rate for $p(p-1)/2$ hypotheses. [50] and [51] used (1.2.2) and (1.2.3) as preliminary estimators, respectively, and obtained a de-sparsified and de-biased estimator, $\widehat{\mathbf{\Omega}} = \widetilde{\mathbf{\Omega}} - \widetilde{\mathbf{\Omega}}^{\mathrm{T}}(\widehat{\mathbf{\Sigma}}\widetilde{\mathbf{\Omega}} - \mathbf{I})$, for inference. [52] investigated the asymptotic normality and minimax optimality for estimating the Gaussian graphical model. They proposed bivariate nodewise regressions based on the scaled Lasso estimator [53]. The aforementioned inference methods for Gaussian graphical models have also been implemented in `SILGGM` R package [54].

Being widely used in spatio-temporal data analysis, ecology, computer vision, and genomics, Gaussian Markov random fields (GMRFs) are closely related to the Gaussian graphical models [55]. GMRF is a Markov random field when the underlying distribution is multivariate Gaussian distribution. By the definition of GMRF, there is conditional dependence between two vertices given all other vertices if and only if $\omega_{ij} \neq 0$. Hence, there is a link between the precision matrix estimation and recovering the Gaussian Markov random field. Many GMRF-type models impose additional dependence assumptions such as the isotropic, the directional, and the distance structures. These assumptions help to reduce the number of parameters to be estimated, and offers both theoretical and computational convenience. However, the validity of these assumptions rely on the prior knowledge of dependence structures. Though a few goodness-of-fit type tests for spatial Markov random fields have been proposed [56, 57], they focus on special classes of structures. GMRF is a Markov random field when the underlying distribution is multivariate Gaussian distribution. Hence, by definition, the conditional dependence between two vertices given other vertices

on a Markov random field is specified by the conditional dependence parameters, which is directly related to the precision matrix under the Gaussianity. Making use of such a link between the precision matrix and the dependence structure of Markov random field leads to a unified framework of model specification tests on the dependence structures of a GMRF model.

The aforementioned estimators mostly rely on assumptions of normality and *i.i.d.* data. However, these assumptions are usually not satisfied or impossible to verify in practice, which will lead to spurious discovery. In literature, efforts on tackling this challenge are scattered. To gain robustness against the violation of Gaussianity assumption, estimators using the copula models have been explored by [58], where, by applying the Glasso to transformed variables, the estimated precision matrix of transformed variables has been shown to recover the graph structure of the original variables accurately under certain sparsity and regularity assumptions. In addition, taking the advantage of robustness of rank statistics, [46] employed the (Spearman's or Kendall's) rank correlation instead of the sample covariance to estimate the underlying graph. On the other hand, robust estimations to the precision matrix under the cellwise contamination model (1.1.3) have been studied in [59–61], where some robust covariance estimate is first obtained, then the sparse precision matrix is estimated using the Glasso or CLIME with the robust covariance estimate as an input instead of the sample covariance matrix. Particularly, the breakdown point of the estimator for the cellwise contamination model was studied [61].

1.3 Outline

The rest of dissertation proceeds as follows. In Chapter 2, we study a large-scale inference method by employing a novel robust regression that is robust against tail-behavior or asymmetric distributions. We prove that the proposed procedure provide consistent estimate of the false discovery proportion under mild assumptions. We also obtain the non-asymptotic tail bound for a new covariate estimator of the regression estimate. In Chapter 3, we focus on the model specification tests of the Gaussian Markov random fields based on temporally dependent data. The proposed method is flexible to test a large number of linear and nonlinear dependence structures of GMRF.

In Chapter 4, we consider the robust estimation of Gaussian graph models via the nodewise regression method. We employ a distributional robust linear regression with an ambiguity set defined by Wasserstein-2 distance. We document the properties of the proposed methods and study the performance of proposed method by numerical simulations under a variety of distributional perturbations. We conclude the dissertation in Chapter 5 with discussion of future works. We provide proofs from theoretical studies, extra results from numerical studies, and additional results from real data analysis in Appendix A–B.

Chapter 2

Large-scale Inference of Multivariate Regression for Heavy-tailed and Asymmetric Data

2.1 Introduction

Multivariate regression is a fundamental statistical tool for data analysis in various fields ranging from biology, financial economics, linguistics, psychology, to social science. By modeling thousands or tens of thousands of responses and covariates or experimental factors, it provides statistical decisions on the individual levels by simultaneously testing many general linear hypotheses, including covariate-effect analysis, analysis of variance, model comparisons, etc. For example, multivariate regression has become a standard tool in the differential expression analysis in genomics [62], and has also been commonly used in corpus linguistics for the word usage comparison [63]. We refer to [64] for a more comprehensive review on relevant applications.

To simultaneously test many general linear hypotheses, a conventional practice is to compute individual p-values based on F-tests or likelihood ratio tests, and then employ multiple testing procedures to control the false discovery rate [65–67]. This standard approach and its theoretical validity, however, often rely on strong distributional assumptions, such as the normality/sub-Gaussianity or symmetry condition on the error distribution. Its effectiveness in terms of false discovery rate control and power may be compromised when dealing with heavy-tailed and/or skewed data with large scales, such as the microarray data [68], the functional magnetic resonance imaging data [69], and text data [70].

To overcome the above challenge, a procedure that is robust against heavy-tailed and/or skewed error distribution is desired. Heavy tailedness increases the chance of observing data that are more extreme than the majority. We refer to these outlying data points as stochastic outliers. A procedure that is robust against such outliers, evidenced by its better finite sample performance than a non-

robust method is called a *tail-robust procedure* [11]. Different from the conventional robustness under Huber's ϵ-contamination model [1], the notion of tail-robustness focuses on the challenge that methods minimizing the empirical risk perform poorly as the empirical risk is not uniformly close to the population risk given heavy-tailed and/or skewed errors [71]. Lately, a variety of new methods and estimation theory under heavy-tailed models have been developed [8–10, 12, 72], while less progress has been documented in terms of inference, especially in a large-scale setting [7, 13, 14].

Building on the idea of *adaptive Huber regression*, we develop a joint robust multiple testing procedure to test many general linear hypotheses in the presence of heavy-tailed and/or skewed errors. This general framework includes the large-scale simultaneous mean testing problem as a special case. First, we employ the adaptive Huber regression to estimate the multivariate regression model parameters, based on which we construct a robust test statistic and compute the approximated p-values to estimate the false discovery proportion (FDP). Next, we apply Storey's false discovery rate controlling procedure [66] to determine a threshold so that hypotheses with p-values below this threshold are rejected. By allowing the robustification parameter to diverge with the sample size, the adaptive Huber regression estimator admits tight non-asymptotic deviation bound and is asymptotically efficient [12]. Theoretically, the non-asymptotic Bahadur representation is a crucial step for establishing the limiting distribution of the estimator or its functionals. Practically, the proposed method can be fully data-driven [73], and therefore is computationally attractive and applicable to real large-scale problems.

The main contributions of this paper are as follows. Methodologically, we develop a tail-robust multiple testing procedure to simultaneously draw inference on large scale multivariate regressions in the presence of heavy-tailed and/or skewed errors. Compared to the traditional practice in multivariate and high-dimensional statistics, our method imposes mild moment conditions on the data, while the dimension p is allowed to grow exponentially fast with the sample size n. These features make our method particularly advantageous and appealing for conducting inference on large-scale multivariate regression models with heavy-tailed and/or asymmetric errors, which is corroborated

by the comprehensive simulation studies. Also, motivated by [74], we propose a novel covariance estimator of the adaptive Huber regression estimate, and we derive a new exponential-type deviation bound for the covariance that is of independent interest for studying about the controlling of false discovery proportion. For the theoretical analysis of the new procedure, we explore and develop a couple of interesting new technical results, by which we show that the proposed method controls the false discovery proportion asymptotically under mild moment and correlation conditions on the error vector. From computational perspective, the proposed method takes the advantages of the computational efficiency of data-adaptive Huber regression [73]. In addition to numerical experiments, we apply the proposed method to analyze the text data from the Standardized Gutenberg Project Corpus [75]. We identify the genre representative words in works of William Shakespeare, and also investigate the differences among works of Lewis Carroll, Charles Dickens, and Arthur Conan Doyle. This empirical study demonstrates that our method is a useful addition to the existing toolkit for modeling and analyzing text data in quantitative linguistics.

The rest of the paper proceeds as follows. In Section 2.2, we revisit testing general linear hypotheses based on the multivariate regression, and introduce our multiple testing procedure based on the adaptive Huber regression. Particularly, we introduce a novel Huber-type estimator of the covariance matrix of the regression coefficient in Section 2.2.2. We establish the statistical guarantees of our procedure in Section 2.3, and characterize the nonasymptotic performance of the proposed Huber-type estimator of the covariance matrix of the regression coefficient. Section 2.4 is devoted to simulation studies. In Section 2.5, we apply our method to the well-known quantitative linguistic data set, the Gutenberg Project. Extensions of the proposed method are discussed in Section 2.6. All the proofs and additional numerical and empirical analysis are provided in the supplemental material.

2.2 Model and Methodology

Throughout the paper, we write $\|\mathbf{u}\| = (\sum_{i=1}^{d} u_i^2)^{1/2}$ as the ℓ_2-norm of vector $\mathbf{u} = (u_1, ..., u_d)^{\mathrm{T}} \in \mathbb{R}^d$. Let $\langle \mathbf{u}, \mathbf{w} \rangle$ be the inner product of vectors $\mathbf{u}$ and $\mathbf{w}$ and $\|\mathbf{u}\|^2 = \langle \mathbf{u}, \mathbf{u} \rangle$. Denote $\mathbb{S}^{d-1} = \{\mathbf{u} \in$

$\mathbb{R}^d : \|\mathbf{u}\| = 1\}$ the unit sphere in $\mathbb{R}^d$. For matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$, denote $\|\mathbf{A}\| = \sup_{\mathbf{u}\in\mathbb{S}^{d-1}} \|\mathbf{A}\mathbf{u}\|$, $\lambda_{\max}(\mathbf{A})$, and $\lambda_{\min}(\mathbf{A})$ the spectral norm, the maximum eigenvalue, and the minimum eigenvalue, respectively. Let $\Phi(z) := \mathbb{P}(U < z)$ with $U \sim N(0,1)$ be the cumulative distribution function of standard normal. Denote $\mathbb{I}(\cdot)$ the indicator function.

Suppose we observe independent data vectors $\{(\mathbf{Y}_i, \mathbf{X}_i)\}_{i=1}^n$, where $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{ip})^{\mathrm{T}}$, $\mathbf{X}_i = (X_{i1}, \ldots, X_{id})^{\mathrm{T}}$ with $d \geq 1$ and $d/n \to 0$ as $n \to \infty$. For each $j = 1, \ldots, p$, the conditional expectation of Y_{ij} given the explanatory variables $\mathbf{X}_i$ is modeled through $\mathbb{E}(Y_{ij}|\mathbf{X}_i) = \mu_j + \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_j$. Define data matrices $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_n)^{\mathrm{T}} \in \mathbb{R}^{n\times p}$ and $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)^{\mathrm{T}} \in \mathbb{R}^{n\times d}$, so that the multivariate regression model of interest has the matrix form

$$\mathbf{Y} = \mathbf{1}_n\boldsymbol{\mu}^{\mathrm{T}} + \mathbf{X}\mathbf{B} + \boldsymbol{\Xi}, \tag{2.2.1}$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^{\mathrm{T}}$ is the intercept vector, $\mathbf{1}_n = (1, \ldots, 1)^{\mathrm{T}} \in \mathbb{R}^n$, $\mathbf{B} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p) \in \mathbb{R}^{d\times p}$ consists of the slope coefficients, and $\boldsymbol{\Xi} = (\boldsymbol{\epsilon}_1^{\mathrm{T}}, \ldots, \boldsymbol{\epsilon}_n^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{n\times p}$ with $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{ip})^{\mathrm{T}}$. Independent of $\mathbf{X}_i$'s, the p-dimensional residual errors $\boldsymbol{\epsilon}_i$'s are independent and identically distributed with mean zero and covariance matrix $\boldsymbol{\Sigma}_\epsilon = (\sigma_{\epsilon,jk})_{1\leq j,k\leq p}$. To ease the notation, let $\boldsymbol{\theta}_j = (\mu_j, \boldsymbol{\beta}_j^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{d+1}$ and $\mathbf{Z}_i = (1, \mathbf{X}_i^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{d+1}$, and define the parameter and design matrices as $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p) \in \mathbb{R}^{(d+1)\times p}$ and $\mathbf{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_n)^{\mathrm{T}}$, so that (2.2.1) reduces to $\mathbf{Y} = \mathbf{Z}\boldsymbol{\Theta} + \boldsymbol{\Xi}$. Based on model (2.2.1), we are interested in making simultaneous inference on the p general linear hypotheses:

$$H_{0j} : \mathbf{C}\boldsymbol{\theta}_j = \mathbf{c}_{0j} \ \text{ versus } \ H_{1j} : \mathbf{C}\boldsymbol{\theta}_j \neq \mathbf{c}_{0j} \ \text{ for } j = 1, \ldots, p, \tag{2.2.2}$$

where matrix $\mathbf{C} \in \mathbb{R}^{q\times(d+1)}$ and vectors $\mathbf{c}_{0j} \in \mathbb{R}^q$ are prescribed, and $\mathrm{rank}(\mathbf{C}) = q \leq d+1$. The linear hypotheses in (2.2.2) encompass a variety of important applications, including the inference on contrasts in the analysis of variance and testing for treatment effects. Likelihood-based or least squares-based methods have been employed under the assumption that the covariates and/or errors follow either normal distributions or some light-tailed symmetric distributions [76–78]. With a large p, the underlying distributions, by chance alone, may have quite different scales and can be

highly skewed and heavy-tailed. Therefore, outliers will occur more frequently, challenging the efficacy of standard inference methods. Throughout this paper, we will not make any parametric distributional assumptions, such as the normality or elliptical symmetry. Instead, we define moment parameters $v_{j,\delta} = \{\mathbb{E}(|\epsilon_{1j}|^{2+\delta})\}^{1/(2+\delta)}$ for $j = 1, \ldots, p$ and $\delta > 0$. Specifically, set $v_j = v_{j,2}$.

To test the linear hypotheses in (2.2.2), we first estimate the model parameters robustly in the presence of heavy-tailed skewed errors. For $j = 1, \ldots, p$, define Huber-type M-estimators $\widehat{\boldsymbol{\theta}}_j$ as

$$\widehat{\boldsymbol{\theta}}_j := (\widehat{\mu}_j, \widehat{\boldsymbol{\beta}}_j^{\mathrm{T}})^{\mathrm{T}} = \underset{\mu \in \mathbb{R}, \boldsymbol{\beta} \in \mathbb{R}^d}{\operatorname{argmin}} \sum_{i=1}^{n} \ell_{\tau_j}(Y_{ij} - \mu - \mathbf{X}_i^{\mathrm{T}} \boldsymbol{\beta}), \tag{2.2.3}$$

where $\ell_\tau(x) = (x^2/2)\mathbb{I}(|x| \le \tau) + (\tau|x| - \tau^2/2)\mathbb{I}(|x| > \tau)$ is the Huber loss [1] parameterized by $\tau > 0$. Our theoretical analysis suggests that, with $\tau_j \asymp n^{1/(2+\delta)}\{\log(np) + d\}^{-1/(2+\delta)}$ for some $\delta > 0$, the estimators $\widehat{\boldsymbol{\theta}}_j$ are close to $\boldsymbol{\theta}_j$ uniformly over $j = 1, \ldots, p$ with high probability even when p grows exponentially fast with n. Here, the divergence of τ_j guarantees $\widehat{\theta}_j$ to be sub-Gaussian even the error only admits $(2+\delta)$th finite moment, and more importantly, the order of τ_j grants the desired approximation error of Bahadur representation to $\widehat{\boldsymbol{\theta}}_j$ (Theorem 2.3.1) as well as the uniform non-asymptotic bounds of the estimated covariance of $\widehat{\boldsymbol{\theta}}_j$ (Theorem 2.3.2). As noticed in the literature [7–9, 12, 73], the divergent τ_j is necessary to balance the bias and robustness in the presence of heavy tailed and/or skewed errors. On the other hand, the order of τ_j in our setting is different from the earlier studies on the adaptive Huber regressions. For example, with the finite $(1+\epsilon)$th moment of error, [12] focused on the estimation for adaptive Huber regression that corresponds to $p = 1$ in our setting and considered $\tau_j = O(n^{\max\{1/(1+\epsilon),1/2\}}(d+\log n)^{-\max\{1/(1+\epsilon),1/2\}})$, while [7] used $\tau_j = O(n^{1/2}\{\log(np)\}^{-1/2})$ for testing p-dimensional mean vectors under the assumption of finite fourth moment of errors, which corresponds to $d = 1$ in our setting. In practice, τ_j can be chosen by either the cross-validation or the recent data-driven method by [73]. The latter avoids a grid search for each j, and hence is computationally appealing, especially when p is large. Using these robust estimates $\widehat{\boldsymbol{\theta}}_j$'s, we then construct test statistics whose approximate p-values for (2.2.2) are obtained under the null. Partnered with the Benjamini-Hochberg method [65] or

its variants, e.g., [66], we develop a robust procedure to simultaneously test the p hypotheses in (2.2.2).

2.2.1 Test procedure for general linear hypotheses

We are in position to detail our test procedure for (2.2.2). Given estimators $\widehat{\boldsymbol{\theta}}_j$ obtained from (2.2.3) with $\tau_j = \tau_{0j} n^{1/(2+\delta)} \{\log(np) + d\}^{-1/(2+\delta)}$ for $\tau_{0j} \geq v_{j,\delta}$ and $\delta \in (0, 2]$, we consider the following test statistic

$$V_j = n(\mathbf{C}\widehat{\boldsymbol{\theta}}_j - \mathbf{c}_{0j})^{\mathrm{T}} (\mathbf{C}\widehat{\boldsymbol{\Sigma}}_j \mathbf{C}^{\mathrm{T}})^{-1} (\mathbf{C}\widehat{\boldsymbol{\theta}}_j - \mathbf{c}_{0j}) \tag{2.2.4}$$

for $j = 1, \ldots, p$, where $\widehat{\boldsymbol{\Sigma}}_j$ is an estimate of $\boldsymbol{\Sigma}_j := \mathrm{cov}(n^{1/2}\widehat{\boldsymbol{\theta}}_j)$ as we will discuss in Section 2.3.2. The null hypothesis H_{0j} in (2.2.2) will be rejected whenever V_j is large. As we will show, V_j's are asymptotically χ^2_q-distributed under H_{0j} uniformly in j. Leveraging this, we can estimate the false discovery proportion, so as to determine the rejection threshold that makes the estimated false discovery proportion below a prespecified level $\alpha \in (0, 1)$.

Let $\mathcal{H}_0 = \{j : 1 \leq j \leq p, H_{0j} \text{ is true}\}$ and $p_0 := |\mathcal{H}_0|$. Denote the number of discoveries and false discoveries by $R(z) = \sum_{j=1}^{p} \mathbb{I}(V_j \geq z)$ and $V(z) = \sum_{j \in \mathcal{H}_0} \mathbb{I}(V_j \geq z)$, respectively, for threshold $z > 0$. The false discovery proportion is defined as $\mathrm{FDP}(z) = V(z)/R(z)$ with the convention $0/0 = 0$. According to the law of large numbers, $V(z)$ should be close to $p_0 \mathbb{P}(\chi^2_q > z)$ while the number of nulls p_0 is not accessible in general. When both p and p_0 are large and $p_1 = p - p_0 = o(p)$ is small, which is known as the sparse setting in the high-dimensional regime, the approximated false discovery proportion $\mathrm{AFDP}(z) = \widehat{V}(z)/R(z)$ with $\widehat{V}(z) = p\,\mathbb{P}(\chi^2_q > z)$ is a reasonable and slightly conservative surrogate for the asymptotic approximation $p_0 \mathbb{P}(\chi^2_q > z)/R(z)$ and $\mathrm{FDP}(z)$. Using $\mathrm{AFDP}(z)$, we can determine the threshold $\widehat{z}_\alpha = \inf\{z \geq 0 : \mathrm{AFDP}(z) \leq \alpha\}$ for the nominal level α. For $j = 1, \ldots, p$, H_{0j} will be rejected whenever $V_j \geq \widehat{z}_\alpha$.

Of note, if $\pi_0 = p_0/p$ is bounded away from 1 as $p \to \infty$, $\mathrm{AFDP}(z)$ may overestimate $\mathrm{FDP}(z)$. To improve the power, we may combine existing estimations of π_0 in the literature

with our procedure to calibrate the threshold of rejection in a more adaptive fashion. For example, [66] estimates $V(z)$ by $p\widehat{\pi}_0(\eta)\,\mathbb{P}(\chi_q^2 > z)$ for a predetermined $\eta \in [0,1)$, where $\widehat{\pi}_0(\eta) = \{(1-\eta)p\}^{-1}\sum_{j=1}^{p}\mathbb{I}(P_j > \eta)$ estimates $\pi_0 = p_0/p$, and P_j is the p-value associated with the jth test statistic. Among a few studies regarding the selection of η, [67] suggest $\eta = 0.5$, and [79] recommend $\eta = \alpha$ for dependent hypotheses. Using this estimate of $V(z)$, our threshold of rejection can be refined accordingly by $\widehat{z}_\alpha^{\eta} = \inf\{z \geq 0 : p\widehat{\pi}_0(\eta)\,\mathbb{P}(\chi_q^2 > z)/R(z) \leq \alpha\}$.

2.2.2 A refined Huber-type estimator of $\boldsymbol{\Sigma}_j$

A naive estimator of $\boldsymbol{\Sigma}_j = \mathrm{cov}(n^{1/2}\widehat{\boldsymbol{\theta}}_j)$ for conducting the test is $\widetilde{\sigma}_{\epsilon,jj}\widehat{\boldsymbol{\Sigma}}_Z^{-1}$, where $\widetilde{\sigma}_{\epsilon,jj}$ is an estimate of $\sigma_{\epsilon,jj}$, and $\widehat{\boldsymbol{\Sigma}}_Z = n^{-1}\sum_{i=1}^{n}\mathbf{Z}_i\mathbf{Z}_i^{\mathrm{T}}$. When $d = 1$, [7] propose a U-statistic-based variance estimator, and an adaptive Huber-type estimator of the second moment which, combined with mean estimator, is used to estimate variance. The computational complexity of calculating a U-statistic-based estimator is $O(n^2 d)$, and hence grows fast with d. For the latter, because the square data is severely right-skewed, the Huber-type truncation will inevitably lead to underestimation of the second moment and therefore the variance. Motivated by the classical theory of Huber regression, we propose an alternative estimator $\widehat{\boldsymbol{\Sigma}}_j$ based on the asymptotic covariance of the conventional Huber regression estimator; see Section 7.6 of [2].

Given $\tau > 0$, the classical Huber regression estimator $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ admits that $n^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ converges to $N(0, \boldsymbol{\Sigma}_\tau)$ in distribution, where $\boldsymbol{\Sigma}_\tau = \{\mathbb{P}(|\epsilon| < \tau)\}^{-2}\mathbb{E}\{\ell_\tau'(\epsilon)^2\}\boldsymbol{\Sigma}_Z^{-1}$ and $\boldsymbol{\Sigma}_Z = \mathbb{E}(\mathbf{Z}\mathbf{Z}^{\mathrm{T}}) \in \mathbb{R}^{(d+1)\times(d+1)}$ [74]. Resembling $\boldsymbol{\Sigma}_\tau$, our estimator $\widehat{\boldsymbol{\Sigma}}_j$ consists of three Huber-type estimates and makes use of the tapering function [80]

$$\mathbb{I}_\tau^*(x) = \mathbb{I}(|x| \leq \tau) + h_n^{-1}(\tau + h_n - |x|)\mathbb{I}(\tau < |x| \leq \tau + h_n), \tag{2.2.5}$$

which is h_n^{-1}-Lipschitz continuous with $h_n > 0$ denoting a smoothing parameter. To avoid notational clutter, the dependence of $\mathbb{I}_\tau^*$ on h_n will be assumed without displaying. Given a robustification parameter $\tau_j > 0$ and the corresponding estimate $\widehat{\boldsymbol{\theta}}_j$ from (2.2.3), define $\mathbf{W}_j = n^{-1}\sum_{i=1}^{n}\mathbb{I}_{\tau_j}^*(e_{ij})\mathbf{Z}_i\mathbf{Z}_i^{\mathrm{T}}$ and $m_j = n^{-1}\sum_{i=1}^{n}\mathbb{I}_{\tau_j}^*(e_{ij})$, where $e_{ij} = Y_{ij} - \mathbf{Z}_i^{\mathrm{T}}\widehat{\boldsymbol{\theta}}_j$. Respectively, $\mathbf{W}_j$

and m_j are estimates of $\mathbb{P}(|\epsilon_{1j}| \le \tau_j)\mathbf{\Sigma}_Z$ and $\mathbb{P}(|\epsilon_{1j}| \le \tau_j)$. Recall that $\widehat{\mathbf{\Sigma}}_Z = n^{-1}\sum_{i=1}^n \mathbf{Z}_i\mathbf{Z}_i^{\mathrm{T}}$. Inspired by (7.83) of [2], we define the covariance estimator $\widehat{\mathbf{\Sigma}}_j$ in (2.2.4) as

$$\widehat{\mathbf{\Sigma}}_j = \frac{\sum_{i=1}^n \{\ell'_{\tau_j}(e_{ij})\}^2}{(n-d-1)K_j}\mathbf{W}_j^{-1}\widehat{\mathbf{\Sigma}}_Z\mathbf{W}_j^{-1}, \tag{2.2.6}$$

where $K_j = 1 + (nm_j)^{-1}(d+1)(1-m_j)$ is a correction factor that benefits finite sample performance. The tapering parameter h_n, for example, can be set as $n^{-1/4}$ in practice.

For the conventional Huber regression with $\tau > 0$ fixed, it can be shown that, with $\mathbb{I}^*_\tau(x)$ replaced by $\mathbb{I}(|x| \le \tau)$, $\widehat{\mathbf{\Sigma}}_j$ converges in probability to $\mathbf{\Sigma}_\tau$ as $n \to \infty$. To legitimize the use of V_j for testing (2.2.2), we will show in Section 2.3.2 that with adaptive τ_j, the covariance estimator $\widehat{\mathbf{\Sigma}}_j$ in (2.2.6) is close to $\mathbf{\Sigma}_j$ uniformly over $j = 1, \ldots, p$ with high probability.

2.3 Statistical Guarantees

In this section, we establish theoretical guarantees of our method by first assuming a known $\mathbf{\Sigma}_j$, and then exploring the closeness between $\mathbf{\Sigma}_j$ and $\widehat{\mathbf{\Sigma}}_j$ in (2.2.6). Hereafter, we focus on $\mathbf{Z}_i$ being random (except for the first coordinate), and leave the results under fixed designs to the supplement.

2.3.1 Approximation of FDP with known Σ_j

First, we assume the covariance matrices $\mathbf{\Sigma}_j = \mathrm{cov}(n^{1/2}\widehat{\boldsymbol{\theta}}_j)$, $j = 1, \ldots, p$, are known. Consider the oracle test statistic $V_j^\circ = n(\mathbf{C}\widehat{\boldsymbol{\theta}}_j - \mathbf{c}_{0j})^{\mathrm{T}}(\mathbf{C}\mathbf{\Sigma}_j\mathbf{C}^{\mathrm{T}})^{-1}(\mathbf{C}\widehat{\boldsymbol{\theta}}_j - \mathbf{c}_{0j})$. Given $z \ge 0$, write $R^\circ(z) = \sum_{j=1}^p \mathbb{I}(V_j^\circ > z)$, $V^\circ(z) = \sum_{j\in\mathcal{H}_0}\mathbb{I}(V_j^\circ > z)$, and the false discovery proportion $\mathrm{FDP}^\circ(z) = V^\circ(z)/R^\circ(z)$. Heuristically, V_j° is approximately χ^2_q-distributed under H_{0j}, so that we can approximate $\mathrm{FDP}^\circ(z)$ by

$$\mathrm{AFDP}^\circ(z) = \frac{p_0\,\mathbb{P}(\chi^2_q > z)}{R^\circ(z)}. \tag{2.3.1}$$

To show that $\mathrm{AFDP}^{\circ}(z)$ provides a valid asymptotic (pointwise) approximation of $\mathrm{FDP}^{\circ}(z)$, we impose the following technical conditions. Denote $\mathbf{R}_{\epsilon} = (r_{\epsilon,jk})_{1\le j,k\le p}$ the correlation matrix of $\boldsymbol{\epsilon}_1 = (\epsilon_{11},\ldots,\epsilon_{1p})^{\mathrm{T}}$, that is, $\mathbf{R}_{\epsilon} = \mathbf{D}_{\epsilon}^{-1}\boldsymbol{\Sigma}_{\epsilon}\mathbf{D}_{\epsilon}^{-1}$ with $\mathbf{D}_{\epsilon}^2 = \mathrm{diag}(\sigma_{\epsilon,11},\ldots,\sigma_{\epsilon,pp})$.

Condition 1. *(i)* $p = p(n) \to \infty$ *and* $\log(p) = o(n^{1/2})$ *as* $n \to \infty$*; (ii) the error vectors* $\epsilon_1,\ldots,\epsilon_n$ *are independent, and satisfy* $\mathbb{E}(\epsilon_{ij}|\mathbf{Z}_i) = 0$, $\mathbb{E}(\epsilon_{ij}^2|\mathbf{Z}_i) = \sigma_{\epsilon,jj}$*; (iii) there exist constants* $\delta \in (0,2]$ *and* $c_{\epsilon}, C_{\epsilon} > 0$ *such that* $c_{\epsilon} \le \min_{1\le j\le p}\sigma_{\epsilon,jj}^{1/2} \le \max_{1\le j\le p} v_{j,\delta} \le C_{\epsilon}$*; and (iv) there exist* $\kappa_0 \in (0,1)$ *and* $\kappa_1 > 0$ *such that* $\max_{1\le j\ne k\le p}|r_{\epsilon,jk}| \le \kappa_0$ *and* $p^{-2}\sum_{1\le j\ne k\le p}|r_{\epsilon,jk}| = O(p^{-\kappa_1})$.

In Condition 1, (i) is a commonly assumed asymptotic regime for (n,p) in high-dimensional statistical inference; (ii) is standard for linear regression models; compared to the traditional settings which presumes the finite fourth or higher order moments of errors, (iii) only assumes the uniform boundedness of the $(2+\delta)$th moments of error variables; and (iv) allows weak dependence among $\epsilon_{11},\ldots,\epsilon_{1p}$. In addition, we impose the following conditions on the (random) predictor $\mathbf{Z}_i$. Denote $\widetilde{\mathbf{Z}}_i = \boldsymbol{\Sigma}_Z^{-1/2}\mathbf{Z}_i$, where $\boldsymbol{\Sigma}_Z = \mathbb{E}(\mathbf{Z}\mathbf{Z}^{\mathrm{T}})$ is assumed to be positive definite.

Condition 2. *The predictors* $\mathbf{Z}_i$ *are independent and identically distributed, and follow a sub-Gaussian distribution: there exists* $A_0 > 0$ *such that for any* $\mathbf{u} \in \mathbb{R}^{d+1}$ *and* $t \ge 0$, $\mathbb{P}(|\langle \mathbf{u}, \widetilde{\mathbf{Z}}_i\rangle| \ge A_0\|\mathbf{u}\|t) \le 2\exp(-t^2)$.

A few examples of the sub-Gaussian random vector in $\mathbb{R}^d$ in Condition 2 include Bernoulli random vectors, Gaussian random vectors, and random vectors uniformly distributed on the sphere centered at 0 with radius $d^{1/2}$. We refer to [81] for an overview of sub-Gaussian vectors. Under Conditions 1 and 2, Theorem 2.3.1 shows that $\mathrm{AFDP}^{\circ}(\cdot)$ in (2.3.1) provides a consistent (pointwise) estimator of $\mathrm{FDP}^{\circ}(\cdot)$. The consistency of false discovery proportion estimation serves as the cornerstone to the validity of the proposed testing procedure.

Theorem 2.3.1. *Assume Conditions 1 and 2 hold, and* $p_0 \ge ap$ *for some* $a \in (0,1)$. *Let* $\tau_j = \tau_{0j}n^{1/(2+\delta)}\{\log(np)+d\}^{-1/(2+\delta)}$ *with* $\tau_{0j} \ge v_{j,\delta}$ *and* $\delta \in (0,2]$. *Then, for any* $z \ge 0$, $|\mathrm{FDP}^{\circ}(z) - \mathrm{AFDP}^{\circ}(z)| = o_{\mathbb{P}}(1)$ *as* $n,p \to \infty$.

We conclude this subsection with two remarks. If we strengthen Condition 1-(iii) to uniformly bounded k-th moments for $k \geq 4$, Theorem 2.3.1 remains valid with $\tau_j = \tau_{0j} n^{1/(2+\delta)} \{\log(np) + d\}^{-1/(2+\delta)}$ and $\delta \in (0, k-2]$. In addition, to prove Theorem 2.3.1, we will show that $|\mathrm{FDP}^\circ(z) - \mathrm{AFDP}^\circ(z)| = O_{\mathbb{P}}\{p^{-\kappa_1} q^{1/2} + q^{7/4} n^{-1/2} + q\{\log(np) + d\}^{\delta/(2+\delta)} n^{-\delta/(2+\delta)}\}$. This explicit rate is non-trivial and reveals how the parameter q, which corresponds to the dimension of the hypothesis, affects the difficulty of testing (2.2.2). We will revisit this via numerical studies in Section 2.4.

2.3.2 Statistical guarantees with estimated covariance input $\widehat{\mathbf{\Sigma}}_j$

Next, we establish the statistical guarantee of the procedure proposed in Section 2.2.1 using estimated covariance matrices $\widehat{\mathbf{\Sigma}}_j$ defined in (2.2.6). To this end, in Proposition 2.3.1, we provide a mild condition on the (uniform) accuracy of estimated covariances, which is required for the approximate false discovery proportion to be consistent. Let $\widetilde{\mathbf{\Sigma}}_j$ be a generic estimator of $\mathbf{\Sigma}_j$ for each j. The corresponding false discovery proportion and its approximation are $\widetilde{\mathrm{FDP}}(z) = \widetilde{V}(z)/\widetilde{R}(z)$ and $\widetilde{\mathrm{AFDP}}(z) = p_0\, \mathbb{P}(\chi^2_q > z)/\widetilde{R}(z)$ for $z \geq 0$, where $\widetilde{V}(z) = \sum_{j \in H_0} \mathbb{I}(\widetilde{V}_j > z)$, $\widetilde{R}(z) = \sum_{j=1}^p \mathbb{I}(\widetilde{V}_j > z)$, and $\widetilde{V}_j = n(\mathbf{C}\widehat{\boldsymbol{\theta}}_j - \mathbf{c}_{0j})^{\mathrm{T}} (\mathbf{C}\widetilde{\mathbf{\Sigma}}_j \mathbf{C}^{\mathrm{T}})^{-1} (\mathbf{C}\widehat{\boldsymbol{\theta}}_j - \mathbf{c}_{0j})$ for $j = 1, \ldots, p$.

Proposition 2.3.1. *Suppose that the conditions of Theorem 2.3.1 hold. As long as the estimated covariances* $\{\widetilde{\mathbf{\Sigma}}_j\}_{j=1}^p$ *satisfy* $\max_{1 \leq j \leq p} \|\widetilde{\mathbf{\Sigma}}_j - \mathbf{\Sigma}_j\| = o_{\mathbb{P}}\{(\log(np) + d)^{-1}\}$*, we have* $|\widetilde{\mathrm{FDP}}(z) - \widetilde{\mathrm{AFDP}}(z)| = o_{\mathbb{P}}(1)$ *for any* $z > 0$ *as* $n, p \to \infty$.

By verifying that $\widehat{\mathbf{\Sigma}}_j$ defined in (2.2.6) satisfy the required accuracy in Proposition 2.3.1, together with Theorem2.3.1, the following theorem establishes the convergence in probability of the approximated false discovery proportion to the false discovery proportion for any $z > 0$ as $n, p \to \infty$.

Theorem 2.3.2. *Suppose that the conditions of Theorem 2.3.1 hold. Let* $\widehat{\mathbf{\Sigma}}_j$ *be the covariance estimators given in* (2.2.6) *with* $\tau_j = \tau_{0j} n^{1/(2+\delta)} \{\log(np) + d\}^{-1/(2+\delta)}$ *and* $\tau_{0j} \geq v_{j,\delta}$ *for* $\delta \in (0, 2]$. *Then, with probability at least* $1 - 16n^{-1}$*,*

$$\max_{1 \leq j \leq p} \left\| \widehat{\mathbf{\Sigma}}_j - \mathbf{\Sigma}_j \right\| \leq C_1 \max \left[\left\{ \frac{\log(np) + d}{n} \right\}^{\delta/(2+\delta)}, \frac{\Delta}{h_n} \right], \tag{2.3.2}$$

where $\Delta = \{d^{1/2} + (2\log n)^{1/2}\}[n^{-1}\{\log(np) + d\}]^{1/2}$ *and* $C_1 > 0$ *only depends on* $\lambda_{\max}(\Sigma_Z)$, A_0, *and* $v_{j,\delta}$.

Theorem 2.3.2 implies that the required accuracy in Proposition 2.3.1, that is, $\max_{1\le j\le p} \|\widehat{\mathbf{\Sigma}}_j - \mathbf{\Sigma}_j\| = o_{\mathbb{P}}\{(\log(np) + d)^{-1}\}$, is met if $\log(p) + d = o(n^{\delta/(2+2\delta)})$ $(0 < \delta \le 2)$ and $\Delta/h_n = o\{(\log(np) + d)^{-1}\}$, such as $h_n = n^{-1/4}$. Thus far we have focused on $\widehat{\mathbf{\Sigma}}_j$ given in (2.2.6). In fact, the conclusion in Theorem 2.3.2 remains valid for some variant of $\widehat{\mathbf{\Sigma}}_j$, such as $\widehat{\mathbf{\Sigma}}_j^{(1)} = \sum_{i=1}^n \{\ell'_{\tau_j}(e_{ij})\}^2\{(n-d-1)m_j\}^{-1}K_j\mathbf{W}_j^{-1}$.

2.4 Simulation Studies

2.4.1 Model settings

To examine the finite sample performance of the proposed procedure, we consider the following methods: (i) the proposed method that employs data-adaptive Huber regression [73]; (ii) the proposed method with τ_j's selected via five-fold cross-validation [12]; (iii) least squares based multiple testing method; (iv) empirical Bayes based multiple testing procedure implemented via `limma` [62]; (v) `limma` with conventional robust regression; and (vi) empirical Bayes based multiple testing procedure for count data implemented via `edgeR` [82]. Specifically, we set $\delta = 2$ in (2.2.3) (i.e., assume the errors have finite fourth moments) and $h_n = n^{-1/4}$ in (2.2.5). For (ii), we set $\tau_j = c\widehat{v}_j n^{1/4}\{\log(np) + d\}^{-1/4}$ with $\widehat{v}_j^4 = n^{-1}\sum_{i=1}^n (Y_{ij} - \bar{Y}_{\cdot j})^4$, and choose c from $\{0.25, 0.5, 0.75, 1, 1.25, 1.5\}$ based on cross-validation that minimizes the mean-squared prediction error. For (i)–(iii), we employed the FDR controlling procedure by [66] to determine the threshold.

Both `limma` and `edgeR` are widely-used softwares to test a large number of regression models, and serve as benchmark methods in genomics study. Based on the linear model, `limma` employs empirical Bayes methods to shrink individual variances towards a common value in the hope of better controlling the false discovery rate. Method (v) is a modified version of `limma` that employs the traditional robust M-estimation instead of the least squares. `edgeR` is widely used to

model count data with large variations via the negative binomial model. To implement `edgeR`, we round each response variable Y_{ij} to its nearest integer.

We generate data from model (2.2.1) for $n = 85, 120, 150$, $p = 1000, 2000$, $p_1 = 50$, and $d = 6, 8$. Entries of $\mathbf{X} \in \mathbb{R}^{n\times d}$ are independently drawn from $N(0, 1)$, and each column is standardized to have zero mean and unit variance. We consider three heavy-tailed error distributions: (a) Pareto distribution with shape parameter 4 and scale parameter 1, (b) log-normal distribution with $\mu = 0$ and $\sigma = 1$, and (c) a mixture of the log-normal distribution in (b) and the t_2 distribution with proportion 0.7 and 0.3 respectively. Setting (c) reflects more challenging scenarios in practice as t_2 distribution does not even have finite second moment. All settings are also highly skewed. Under each setting, we first generate $\mathbf{E} = (\epsilon_{ij})_{1\le i\le n, 1\le j\le p}$ that has independent entries. To incorporate dependency, we set $\mathbf{\Xi} = 100\mathbf{R}_\epsilon^{1/2}\mathbf{E}$, where the correlation matrix $\mathbf{R}_\epsilon$ has one of the following three structures: *Model* 1, $\mathbf{R}_\epsilon$ is the identity matrix; *Model* 2, $\mathbf{R}_\epsilon = (r_{\epsilon,jk})_{1\le j,k\le p}$ is sparse with $r_{\epsilon,jj} = 1$ and $r_{\epsilon,ij} = r_{\epsilon,ji}$ independently drawn from $0.3 \times \text{Bernoulli}(0.1)$ for $i \neq j$; and *Model* 3, $\mathbf{R}_\epsilon = (r_{\epsilon,jk})_{1\le j,k\le p}$ with $r_{\epsilon,jj} = 1$, $r_{\epsilon,j,j+1} = r_{\epsilon,j+1,j} = 0.3$, $r_{\epsilon,j,j+2} = r_{\epsilon,j+2,j} = 0.1$, and $r_{\epsilon,j,j+k} = r_{\epsilon,j+k,j} = 0$ for $k \ge 3$.

For each $j = 1, \ldots, p$, we set $\mu_j = 5000$ and consider two hypotheses: *Hypothesis* 1, $H_{0j} : 1^{\mathrm{T}}\boldsymbol{\beta}_j = 0$ versus $H_{aj} : 1^{\mathrm{T}}\boldsymbol{\beta}_j \neq 0$, where $q = 1$, and *Hypothesis* 2, $H_{0j} : \boldsymbol{\beta}_j = 0 \in \mathbb{R}^d$ versus $H_{aj} : \boldsymbol{\beta}_j \neq 0$ $(j = 1, \ldots, p)$, where $q = d$. For *Hypothesis* 1, we let $\beta_{jk} \sim \text{Unif}(-150, 150)$ for $1 \le j \le p$ and $1 \le k \le d-1$, $\beta_{jd} = -\sum_{k=1}^{d-1}\beta_{jk}$ for $1 \le j \le p - p_1$ so that $1^{\mathrm{T}}\boldsymbol{\beta}_j = 0$, and $\beta_{jd} = \delta d^{1/2}W_j - \sum_{k=1}^{d-1}\beta_{jk}$ for $p - p_1 + 1 \le j \le p$, where W_j are Rademacher random variables. For *Hypothesis* 2, let $\boldsymbol{\beta}_j = 0$ for $1 \le j \le p - p_1$, and $\beta_{jk} = (2d^{-1})^{1/2}\delta W_{jk}$ for $p - p_1 + 1 \le j \le p$ and $1 \le k \le d$, where W_{jk} are Rademacher random variables. We take $\delta = 75\eta$ and $\eta = 0.3$. The signal strength is determined by η and d.

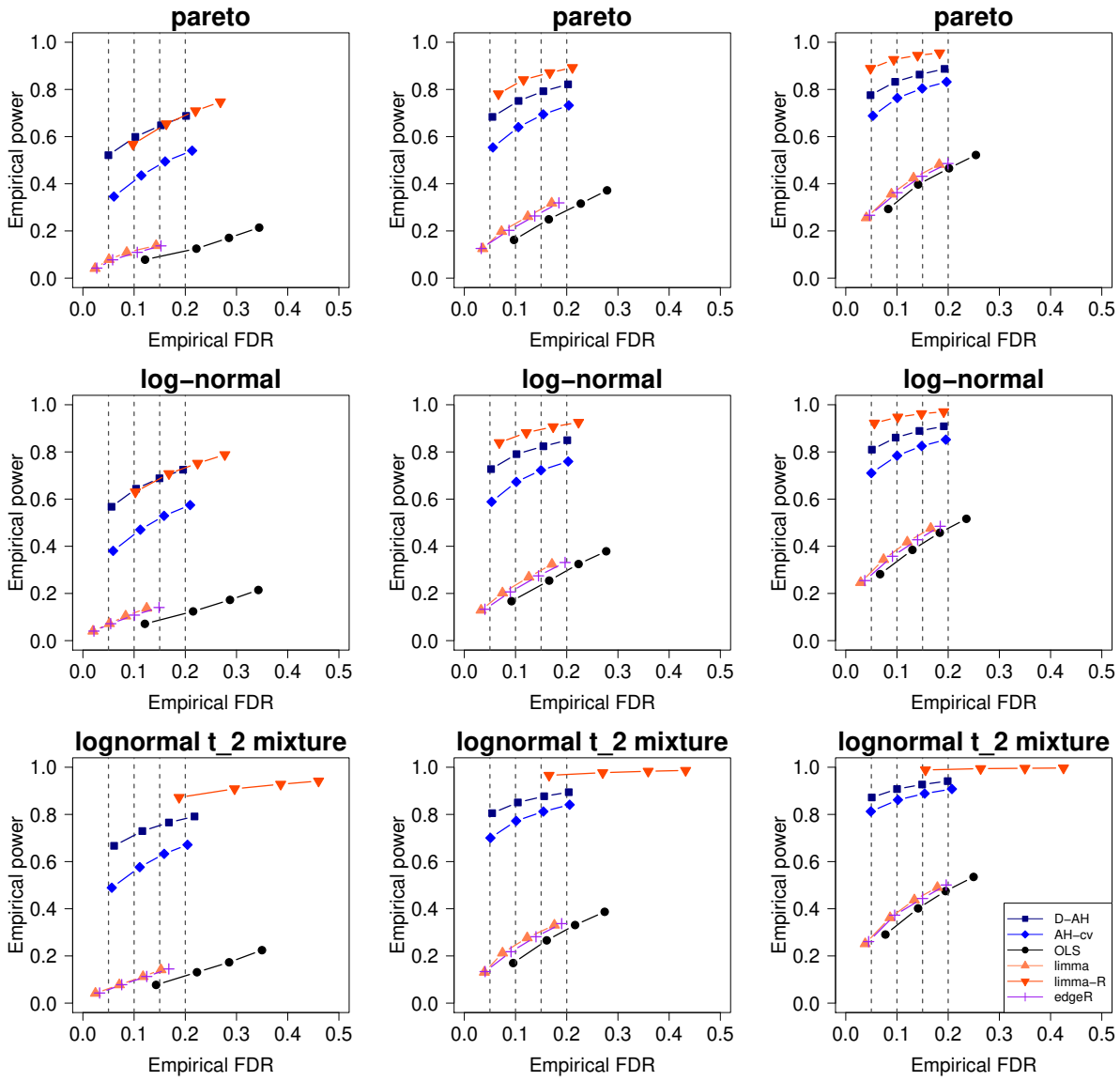
2.4.2 Numerical performance

For each model, we take the nominal level $\alpha = 0.05, 0.1, 0.15, 0.2$, and carry out 250 Monte Carlo simulations at each α. Figures 2.1 and 2.2 report the empirical false discovery rate and

power under *Model* 2 with $p = 1000$ and $d = 6$. The results under other model settings are given in Section A.5 of the supplementary material. Each point corresponds to a nominal level (marked as a vertical gray dashed line) with x-axis and y-axis representing, respectively, the empirical false discovery rate and power. Therefore, the closer the point is to the corresponding vertical line, the more the empirical and nominal false discovery rates coincide.

From Figures 2.1 and 2.2, for different error distributions and hypotheses of interest, the proposed method, with either data-driven Huber regression or cross-validation, is able to control the false discovery rate in general while maintain high power. The competing methods, especially when n is small, are either too conservative with a notable power loss or too liberal to control the false discovery rate. The advantage of our method is more substantial for linear hypotheses with $q > 1$; see Figure 2.2. The numerical evidence favors the use of data-adaptive Huber regression over cross-validation in terms of both statistical accuracy and computational cost. Both `limma` and `edgeR` are fairly conservative, suggesting that researchers should take precautions when using them for heavy-tailed and skewed data. Method (v) is comparable to the proposed methods when n is large, but completely fails to control the false discovery rate in the setting of mixture errors of log-normal and t_2. Overall, the empirical power of all methods increases with n, and drops for larger p as shown in Figures A.1-A.11 in the supplementary material. Since the intrinsic difficulty of the testing problem increases with q, the empirical power of all methods is lower when $q = d = 8$; see Figures A.3 and A.4.

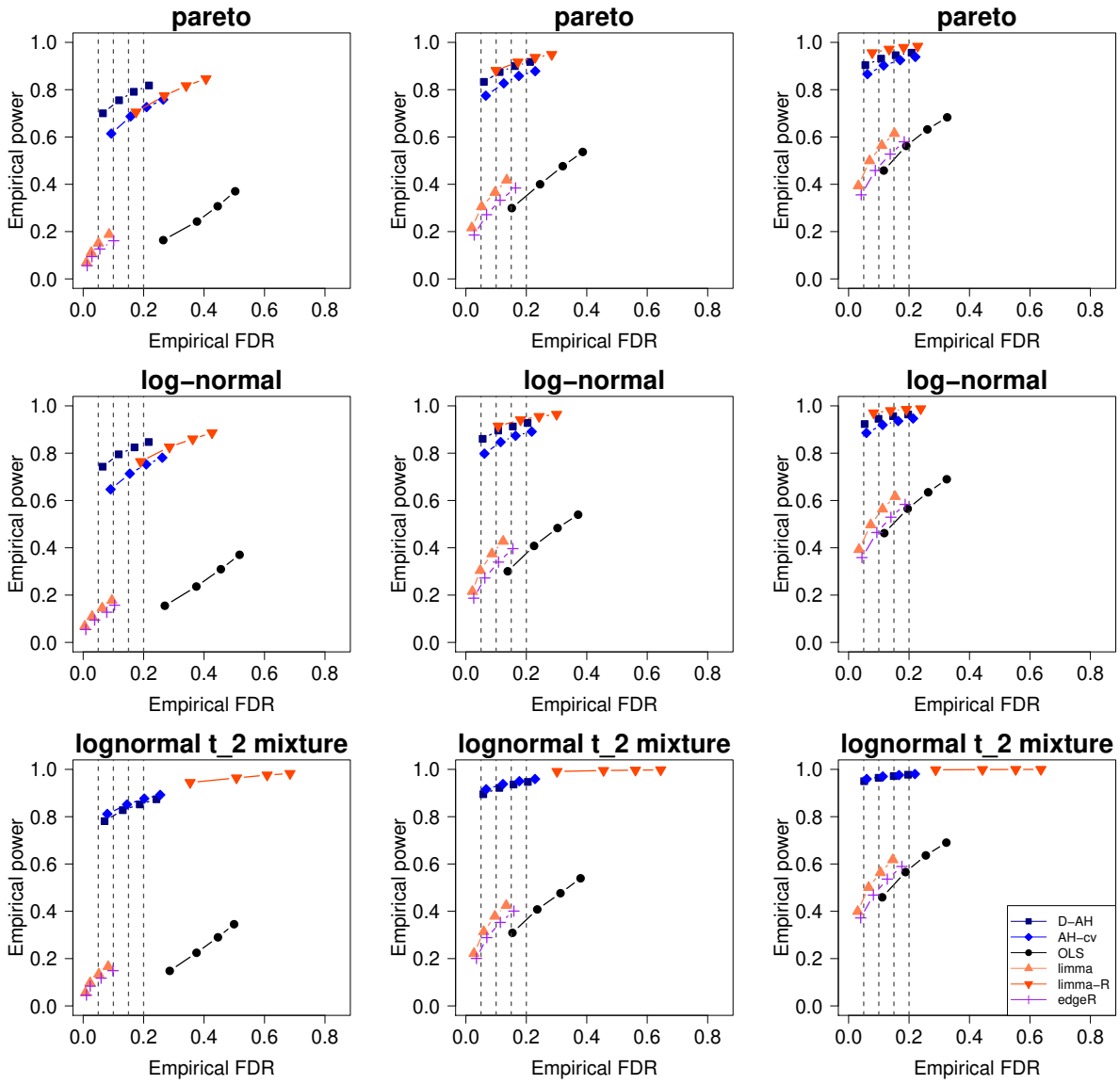
We further examine the power performance with varying signal strengths, determined by η. We exclude methods (iii) and (v) due to their failure on controlling the false discovery rate. In the above data generating process, we take $n = 100$, $p = 1000$, $d = 6$, and choose equally spaced η varying within $[0.3, 0.7]$ for *Hypothesis* 1 and $[0.3, 0.5]$ for *Hypothesis* 2. The results are summarized in Figure 2.3, from which we see that the proposed methods outperform the competitors under all three error settings. The gains in power are considerable when the error distribution is both heavy-tailed and skewed. Again, for our method, the data-adaptive approach is slightly more powerful



(a) $n = 85$ **(b)** $n = 120$ **(c)** $n = 150$

Figure 2.1: Plots of empirical false discovery rate (FDR) and power for testing *Hypothesis* 1 under *Model* 2 with $p = 1000$ and $d = 6$ by six methods: the proposed method with data-adaptive Huber regression (D-AH, ■); the proposed method with five-fold cross-validation (AH-cv, ♦); the least squares method (OLS, ●); `limma` (▲); `limma` with robust regression (`limma`-R, ▼); and `edgeR` (+). Each point corresponds to a nominal level (marked as a vertical gray dashed line) with x-axis representing the empirical false discovery rate and y-axis denoting the power.

than cross-validation. As the error dependence becomes stronger (Model 3) or the tail gets heavier (mixture error of log-normal and t_2), the power slightly drops for all methods.

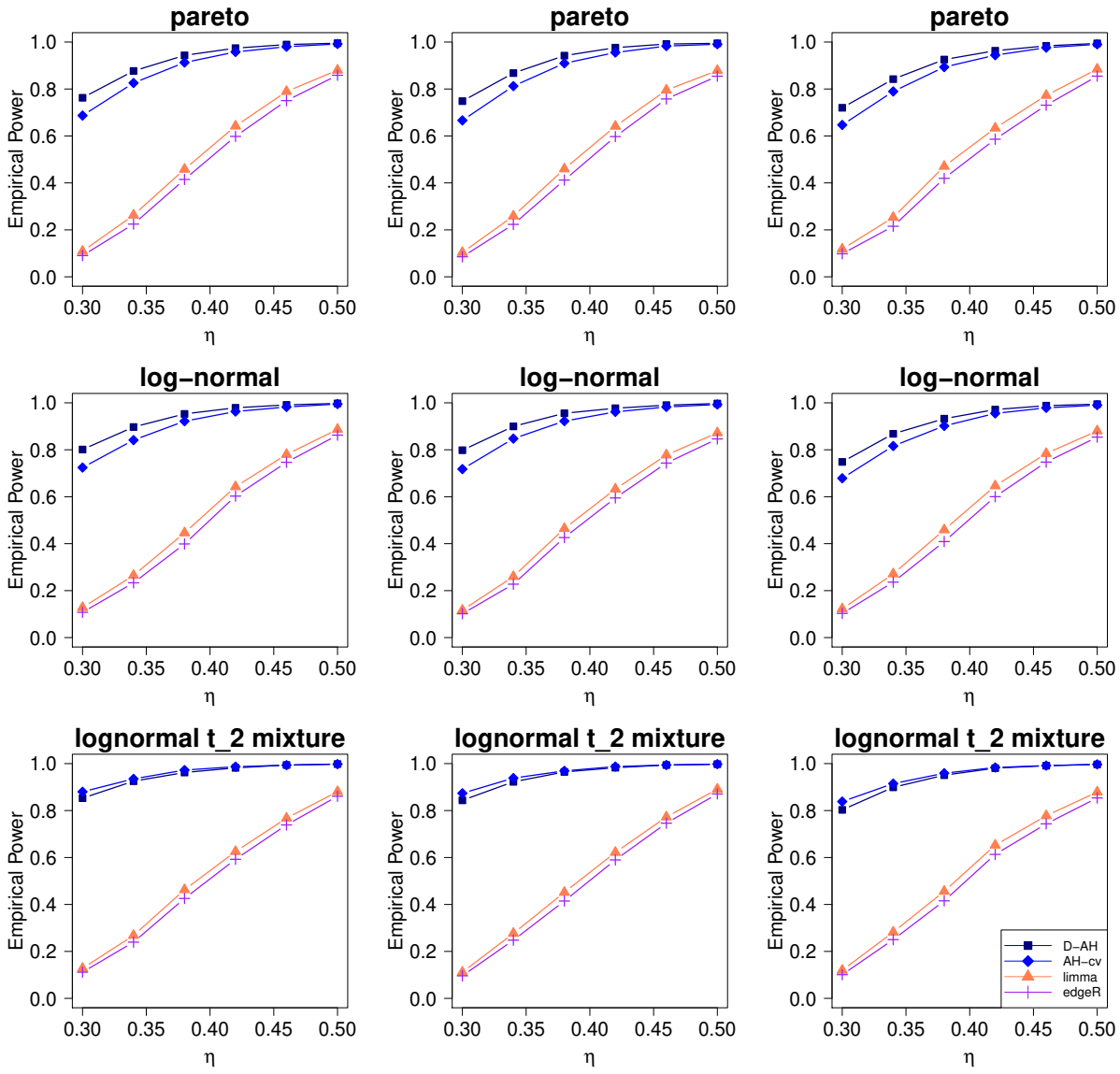


(a) $n = 85$ **(b)** $n = 120$ **(c)** $n = 150$

Figure 2.2: Plots of empirical false discovery rate (FDR) and power for testing *Hypothesis* 2 under *Model* 2 with $p = 1000$ and $d = 6$ by six methods: the proposed method with data-adaptive Huber regression (D-AH, ■); the proposed method with cross-validation (AH-cv, ⧫); the least squares method (OLS, ●); `limma` (▲); `limma` with robust regression (`limma`-R, ▼); and `edgeR` (+). Each point corresponds to a nominal level (marked as a vertical gray dashed line) with x-axis representing the empirical false discovery rate and y-axis denoting the power.

2.5 Real Data Analysis: The Gutenberg Project

Large-scale text data have been collected and used in many applications from linguistics to natural language processing. Corpus linguistics has arisen along with the technological advancement to access, store, and process vast amount of text in a short time [83]. Statistical inference on text



(a) Model 1 **(b)** Model 2 **(c)** Model 3

Figure 2.3: Plots of empirical powers for testing *Hypothesis* 2 with $n = 100$, $p = 1000$, $d = 6$, and $\eta \in \{0.30, 0.34, \ldots, 0.46, 0.5\}$ by four methods: the proposed method with data-adaptive Huber regression (D-AH, ■); the proposed method with cross-validation (AH-cv, ♦); `limma` (▲); and `edgeR` (+).

data from literary publications has drawn great attention in order to provide novel and revealing linguistic discoveries. As a well-known public accessible digital library to literary publications, the Project Gutenberg is founded in 1971, and offers 60156 e-books as of September 03, 2019 with various formats. The Standardized Project Gutenberg Corpus (SPGC, [75]) is a text corpus of Project Gutenberg, and provides a static version of the corpus (https://doi.org/10.5281/zenodo.2422560).

It consists of three data types: raw text, sequences of word-tokens, and word counts. In addition, SPGC also contains metadata about books, such as the author information (name, years of birth and death), language, subject categories, and type of books (text, sound, etc.), whereas the latter provides collections such as “science fiction" or “western".

In this section, we apply our method to word counts from SPGC to identify “differentially represented" words for different hypotheses of interest. That is, to find the idiosyncratic words to represent an author or a category of publications. Specifically, we consider two problems: a comparison of works of Lewis Carroll, Charles Dickens, and Arthur Conan Doyle, and the study of works of William Shakespeare. Table A.1 in the supplementary material displays a snapshot of the raw data, which is highly skewed. From the histograms of empirical kurtosis of word counts in Figure A.14, the data is heavy-tailed in both book-wise and word-wise. For data pre-processing, we first merge word counts from different books, and then remove the words whose total count is less than half the number of books or those only appear in less than 20% of the books under consideration. Finally, we normalize the filtered word counts by the total counts [84]. More details are deferred to the supplementary files.

For the first problem, the three British authors are from the mid 19th to early 20th century, and have similar vocabulary usage. On the other hand, we also observe separations and clusters of their 167 works based on the word usage in Figure A.14 in the supplementary file. To identify differentially represented words in their works, we use model (2.2.1) with

$$\mathbf{X}_i = \begin{cases} (1,1,0)^{\mathrm{T}} & \text{the } i\text{th book is written by Lewis Carroll} \\ (1,0,1)^{\mathrm{T}} & \text{the } i\text{h book is written by Charles Dickens} \\ (1,-1,-1)^{\mathrm{T}} & \text{the } i\text{th book is written by Arthur Conan Doyle} \end{cases}$$

for $i = 1, \ldots, 167$ books and $\boldsymbol{\beta}_j = (\mu_j, \alpha_{1j}, \alpha_{2j})^{\mathrm{T}}$ for $j = 1, \ldots, 6839$ words. We consider the following linear hypotheses:

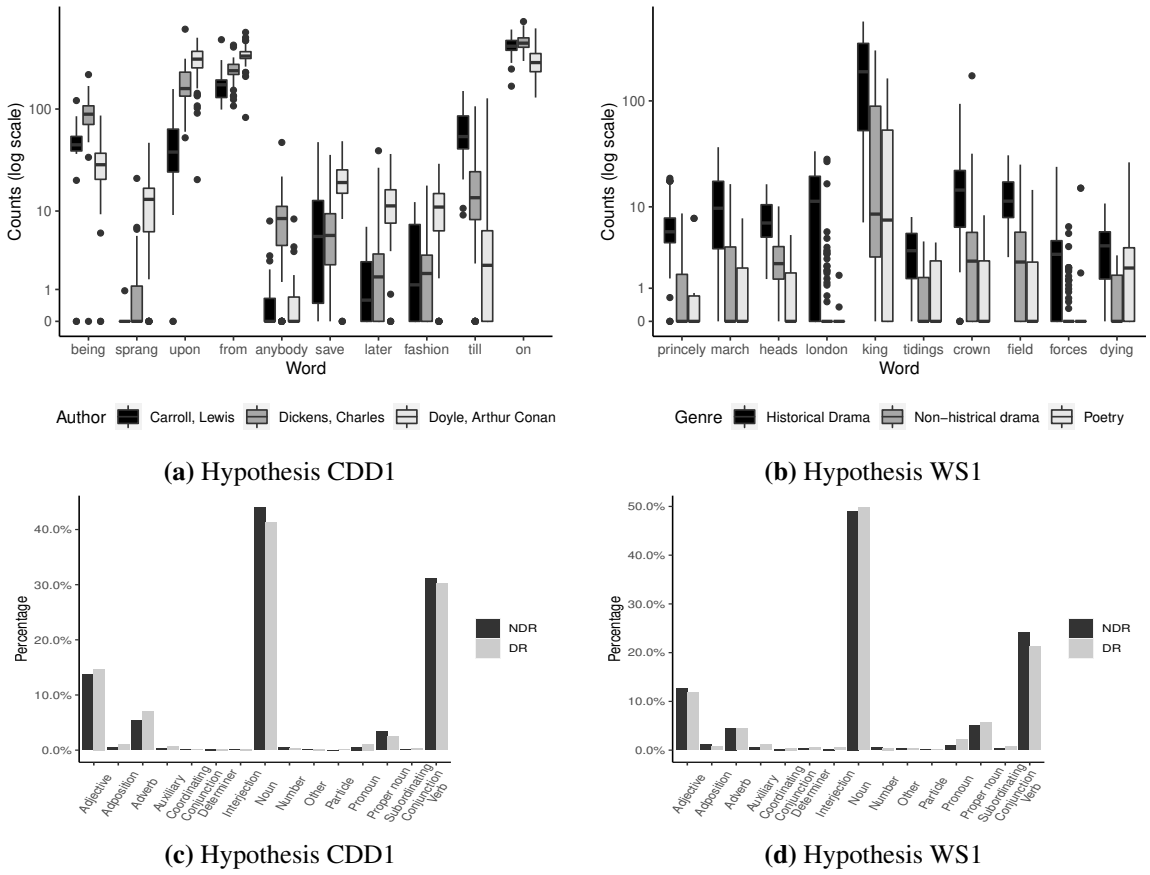
$$\text{(Hypothesis CDD1) } H_{0j}: \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \boldsymbol{\beta}_j = 0 \text{ versus } H_{aj}: \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \boldsymbol{\beta}_j \neq 0;$$

(Hypothesis CDD2) $H_{0j}: \alpha_{1j} = 0$ versus $H_{aj}: \alpha_{1j} \neq 0$; (Hypothesis CDD3) $H_{0j}: \alpha_{2j} = 0$ versus $H_{aj}: \alpha_{2j} \neq 0$; and (Hypothesis CDD4) $H_{0j}: (0,1,1)^{\mathrm{T}}\boldsymbol{\beta}_j = 0$ versus $H_{aj}: (0,1,1)^{\mathrm{T}}\boldsymbol{\beta}_j \neq 0$. Hypothesis CDD1 compares the three authors altogether, while the other hypotheses compare one author to the remaining two. With a nominal level 0.5%, our method detects 2595, 419, 1388, and 1445 differentially represented words for each hypothesis. The top 10 differentially represented words for the three authors, such as “being" and “sprang", are displayed in Figure 2.4(a), while the overall comparison is reported in the Venn diagram in Figure A.15 in the supplementary file. It is interesting to notice that Arthur Conan Doyle favored “sprang" while Lewis Carroll and Charles Dickens barely used it. In Figure 2.4(c), we further report the percentages of differentially represented words (DR) and non-differentially represented words (NDR) within each speech category [85, 86]. Differentially represented words among these three authors have higher percentages in adjectives, adverbs, and pronouns than non-differentially represented words. In contrast, differentially represented words have lower percentages in nouns, proper nouns, and verbs than non-differentially expressed words.

Following the above, we next investigate the genre difference among works of William Shakespeare based on three subject groups: poetry, non-historical drama, and historical drama. We model the normalized word counts by (2.2.1) with

$$\mathbf{X}_i = \begin{cases} (1,0,0)^{\mathrm{T}} & \text{the } i\text{th book is a poetry} \\ (1,1,0)^{\mathrm{T}} & \text{the } i\text{th book is a non-historical drama} \\ (1,1,1)^{\mathrm{T}} & \text{the } i\text{th book is a historical drama} \end{cases}$$

for $i = 1, \ldots, 176$ books and $\boldsymbol{\beta}_j = (\mu_j, \alpha_j, \gamma_j)^{\mathrm{T}}$ for $j = 1, \ldots, 4122$ words. We consider (Hypothesis WS1) $H_{0j}: (0,0,1)^{\mathrm{T}}\boldsymbol{\beta}_j = 0$ versus $H_{aj}: (0,0,1)^{\mathrm{T}}\boldsymbol{\beta}_j \neq 0$, which compares the non-historical



(a) Hypothesis CDD1 **(b)** Hypothesis WS1

(c) Hypothesis CDD1 **(d)** Hypothesis WS1

Figure 2.4: Panels (a) and (b): the top 10 differentially represented words placed in ascending order by their p-values (from left to right) for hypotheses CDD1 and WS1, respectively, and the vertical axis is counts under log-scale. Panels (c) and (d): percentages of differentially represented words (DR) and non-differentially represented words (NDR) within each speech category (https://universaldependencies.org/u/pos/all.html) for hypotheses CDD1 and WS1, respectively. The nominal level is 0.5%.

and historical dramas, and (Hypothesis WS2) $H_{0j} : (0, 2, 1)^{\mathrm{T}}\boldsymbol{\beta}_j = 0$ versus $H_{aj} : (0, 2, 1)^{\mathrm{T}}\boldsymbol{\beta}_j \neq 0$, which distinguishes poetry and all dramas. With a nominal level 0.5%, our method identifies 724 and 225 differentially represented words for each hypothesis. As a vast amount of historical dramas of Shakespeare are about kings of the Kingdom of England, the words “princely", “London", “king", and “crown" appear more in the historical dramas; see Figure 2.4(b). In addition, Shakespeare used vocabularies such as “march", “forces", “army", “battle", and “war" more frequently in the historical dramas than in the non-historical dramas. Interestingly, the love story related lexicons, such as “love" and “marry", appear more in his non-historical dramas. From Figure 2.4(d), the differentially represented words between historical dramas and non-historical

dramas of Shakespeare have higher percentages in nouns, pronouns, and proper nouns, whereas their percentages are low in adjectives, adverbs, and verbs.

In summary, our method provides a reliable addition to the existing toolkit in corpus linguistics and text/literature analysis. It can be employed to study the differences for specific words, which extends the current state-of-art that focuses on the overall distribution of word counts. An interesting follow-up analysis is to investigate how do the stopping words, such as “being" or “upon", affect the results and whether their removal will lead to different discovery. We leave this to future studies.

2.6 Discussions

We conclude this article by discussing several open issues. First, our inference method is based on normal approximation, which works well for a moderate sample size. Given a relatively small sample, the multiplier bootstrap may have a better finite sample performance. The pioneering work of [87] on the Gaussian approximation to the functional of high dimensional empirical processes sheds light on the multiplier bootstrap applied to the adaptive Huber regression. The validity of multiplier bootstrap for adaptive Huber regression will require a finite fourth moment condition on the errors, which is similar to Condition 1. Secondly, it is possible to extend our method to deal with a mixed-effects model $\mathbf{Y}_i = \mathbf{\Theta}\mathbf{Z}_i + \mathbf{A}\mathbf{f}_i + \boldsymbol{\epsilon}_i$, where $\mathbf{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p)^{\mathrm{T}} \in \mathbb{R}^{p\times(d+1)}$, $\mathbf{A} \in \mathbb{R}^{p\times K}$ is the loading matrix, $\mathbf{f}_i \in \mathbb{R}^K$ are zero-mean latent factors that are unobserved, and $\boldsymbol{\epsilon}_i \in \mathbb{R}^p$ are uncorrelated with $\mathbf{f}_i$ and $\mathbf{Z}_i$. For multiple mean effects testing, this model has been studied by [7]. For testing general linear hypotheses, a similar yet more involved procedure can be developed.

In addition, our framework can be generalized for design matrix with potentially heavy-tailed entries. In practice, take the mediation analysis involving the RNA-sequencing data for example, both the responses and entries in the design matrix are potentially heavy-tailed. To tackle this challenge, we may replace the entries in design matrix by its trimmed version $X_i^{\bar{\omega}} = (\varphi_{\bar{\omega}}(x_{i1}) \ldots, \varphi_{\bar{\omega}}(x_{id}))^{\mathrm{T}}$, where $\varphi_{\bar{\omega}}(u) = \min\{\max(-\bar{\omega}, u), \bar{\omega}\}$ with tuning parameter $\bar{\omega} > 0$. Here, the data driven selec-

tion on $\bar{\omega}$ is largely unknown and cross-validation is therefore mandatory for implementations. At the cost of extra tuning parameter $\bar{\omega}$ and additional $\log(np)$ term in the orders of both τ and $\bar{\omega}$, results similar to Theorem 2.3.1 can be established while the theoretical guarantee on $\widehat{\mathbf{\Sigma}}_j$ is more involved. Lastly, in this paper, we focus on tail-robustness against stochastic outliers caused by heavy-tailed and/or skewed error distributions. This is different from the classical robustness characterized by the breakdown point [88] under Huber's ϵ-contamination model, which emphasizes the tolerance of a statistical method to a fraction of arbitrary outliers. It is of great importance to develop proper inference methods for large-scale multivariate regression that are robust against arbitrary contamination. We leave these for future work.

Chapter 3

Model Specification Tests for Dependence Structures of Gaussian Markov Random Fields on Temporally Dependent Data

3.1 Introduction

Last decades have witnessed an ever-increasing capacity of data acquisition technologies in many areas, including biological domains such as epidemiology and genetics, engineering and industry such as distributional management, and social sciences such as social networks and human behavior. As a result, new statistical techniques are demanded to provide better understanding of the data of both unprecedented size and complex structures. Among many statistical tools, network models have been commonly employed for abstracting noisy data and providing an insight into regularities and dependencies among observations. For example, nodes of the network in an epidemiology study can represent regions or individuals under exposure, meanwhile edges can model the associations among regions or individuals [89, 90]. Similarly, in a genetic study, nodes and edges in a network can model the genes from a particular organism and intra-gene dependencies, respectively [91–93]. In a social domain, nodes and edges of a network can represent the individuals and human-human interactions [94]. In ecology, social network analysis is a flexible toolbox to analyze animal social system [95, 96]. From both statistical and computational perspectives, the probabilistic graphical models, specifically the Markov random fields (MRFs), pave a natural path to model and explore networks. Different from traditional approaches based on covariances, these models capture the conditional independence between random variables and therefore recover the intrinsic associations among nodes. With the structure of MRFs estimated, statistical predictions can be obtained via Kriging or the network can be visioned through connecting nodes that are conditionally dependent [97–99].

Recent popular techniques for modeling network is a probabilistic graphical model such as MRF and Bayesian network [98, 100, 101]. Bayesian network have been widely employed in social science, genomics studies, ecology, marketing researches, and public health. On the other hand, MRF is the most popular model of undirected graphs. Let $G = (V, E)$ be a undirected graph where V denotes the set of vertices, and E denotes the set of edges over vertices. A node $u \in V$ can represent a gene, an organism, an individual, or a region, and an edge $(u, v) \in E$ can represent a relationship between nodes. MRFs satisfy Markov properties in undirected graph: global Markov, local Markov, and pairwise Markov property, and it represents conditional independence between nodes. Let $\mathbf{Y} = (Y_1, ..., Y_p)'$ be a random vector of nodal states. Any MRF can be written as an exponential family in a canonical form by Hammersley-Clifford theorem, $P(\mathbf{y}|\boldsymbol{\theta}) = Z^{-1} \prod_{c \in C} \psi_c(\mathbf{y}_c|\boldsymbol{\theta}_c)$, where C is the set of all the maximal cliques of G, $\mathbf{y}_c$ is a realization of $\mathbf{Y}_c$ for a maximal clique c, and Z is the partition function, which is a positive real-valued function. There are various examples of MRFs such as Gaussian MRF (GMRF), Ising model, Hopfield networks, and Potts model.

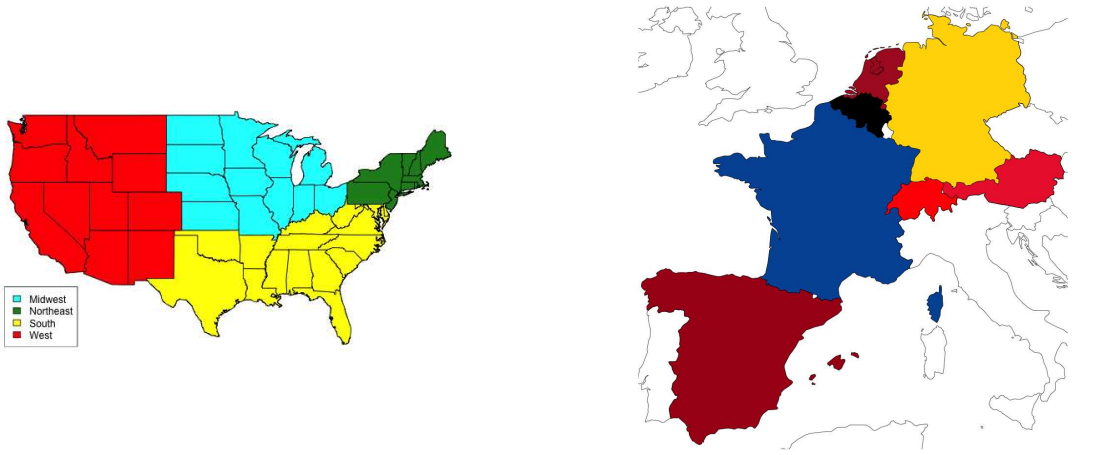
GMRFs play a key role in graphical models [55, 101]. An undirected graph defined by GMRF is equivalent to a precision matrix $\boldsymbol{\Omega} = (\omega_{l_1,l_2})_{1 \leq l_1,l_2, \leq p}$, and $\omega_{l_1,l_2} = 0$ implies conditional independence between nodes. Therefore, testing the dependency structure of a GMRF is equivalent to testing the structure of a precision matrix. Besides, GMRFs enjoy computational advantages from numerical methods for sparse matrices because of the sparse structure of $\boldsymbol{\Omega}$. In addition, applications of GMRF include not only graphical models, but also structural time-series analysis, longitudinal and survival data analysis, image analysis, and spatial statistics.

Understanding the neighborhood structures, $N(v_j)$ for $j = 1, \ldots, p$, and dependency structure of vertices is a prerequisite for constructing GMRF. However, such neighborhood structures are usually unknown and non-trivial in real problems, and they may change from case to case. MRF models with inappropriate neighborhoods could result in misleading results. Therefore, identifying the neighborhood structures based on data is vital in network studies. [56] considered the hypothesis (3.2.4) of four nearest neighborhood against eight nearest neighborhood by using empirical

likelihood. However, their method depends on partition the spatial locations into cliques, which is not trivial to extend to general null hypothesis. Same as the neighborhood structure, GMRF with inappropriate dependency structure could lead to misleading results. In applications, we often face a single time-evolving data with unknown dependency structure. However, the existing methods focused on the two-sample test, specific structures, or non-evolving data. Therefore, a unified approach for evolving data could serve an important role for scientific researches in different disciplines.

Many existing methods focus on estimation in both non-evolving and evolving data, but not on inference in their intrinsic structure. Several estimation methods for high-dimensional sparse precision matrix or Gaussian graphs have been proposed, which are related to GMRF. For example, [35] proposed a neighborhood selection for high-dimensional graphs with Lasso. [37] and [39] used graphical Lasso for a penalized likelihood estimation. [45] considered the connection between multivariate linear regression and sparsity of the precision matrix by linear programming. [40] proposed the CLIME estimator. [102] proposed an asymptotically tuning-free approach via square-root Lasso based nodewise regression. [103] studied the problem of estimating conditional precision matrix given an indexing variable. [104] studied sparsity and clustering structure of graphs using a regularized maximum likelihood method. Moreover, much efforts devoted to MRFs. For instance, [105] proposed ℓ_1-regularized logistic regression of each variable on other variable any MRF. [98] studied estimation of time-evolving Ising model. [106] studied time-varying Gaussian network with abrupt changes. [107] investigated a change point estimation in high dimensional MRFs.

On the other hand, a few testing procedures have been proposed for graphical models. For example, [56] proposed a testing procedure by blockwise empirical likelihood under MRF, whereas [108] proposed a procedure for testing diagonal spatial precision matrix under GMRFs. [109] proposed an inference procedure for evolving nonparanormal graphical model. However, all these models have structural limitation since the null hypothesis is not a general one. In addition, there



(a) US states and geographical regions by US Census Bureau

(b) European countries: Austria, Belgium, France, Germany, Netherlands, Spain, and Switzerland

Figure 3.1: States or provinces used in Google Flu Trends analysis in Section 3.4.

have been a few papers about two-sample test. For example, [110] proposed a test for differential networks under GMRF in genomics applications.

In this paper, we consider the hypotheses testing for the dependence structures of GMRFs. The procedure consists of three parts as follows: We first obtain the residuals from the debiased Lasso estimates of nodewise regressions [52,111], so we obtain the testing statistic. This ℓ_1-type approach can model the sparsity of precision matrix, which is equivalent to the sparsity of GMRF. Next, we estimate a long-run covariance matrix to obtain empirical distribution of the test statistic [111], which allows us to study the asymptotic distribution of the average of time dependent data. In the last step, we use the wild bootstrap [112] to obtain the empirical distribution of an L_∞-type test statistic, which allows us to test the hypothesis.

We applied our method to Google Flu Trends described in Section 3.4. Figure 3.1 displays regions in EU and US that we used in our analysis. Our findings via GMRF are as follows: We found that the number of neighborhood regions in GFT network was larger than one of its geographical neighbors. Moreover, we observed EU has narrower neighborhood than US. For dependency structure, it was more likely the distance structure than isotropic structure. Some sub-

graph had different dependency structure to the overall dependency structure. For example, Spain had more likely isotropic structure. These findings suggest further investigation on their possible impact on flu transmission dynamics. See Section 3.4 for more discussions.

The rest of the paper is organized as follows. In Section 3.2, we review the GMRF and explore a few important local dependence structures based on the GMRF formulation and parameters. In Section 3.3, we provide the testing procedure for general structure GMRF. In Section 3.4, we apply our methods on Google Flu Trends to investigate regional network structure. Section 3.5 is devoted to simulation studies.

3.2 Dependence Structures in GMRF

In this section, we discuss the connection between the model specification tests for conditional dependence structure and the unified inference procedure on structures of the precision matrix. Model-specification test for conditional dependence can be formulated through the conditional dependency parameters $\{\eta_{j_1 j_2}\}$ in GMRF in (3.2.2), and hypotheses in functions of $\{\eta_{j_1 j_2}\}$ are equivalent to those in the functions of precision coefficients. Taking advantage of such a connection, assessing the GMRF's structures can be formulated as a hypothesis testing problem. A few examples are presented in details to elaborate this.

Let the response be $\mathbf{y}_t = (y_{1,t}, \ldots, y_{p,t})'$ for $t = 1, \ldots, n$, where $y_{j,t}$ denotes the response variable at location s_j and time t. Let $V = \{s_j\}_{j=1}^p$ is the set of locations. For each time point t, suppose $\{\mathbf{y}_t\}$ follows a GMRF model, where the full conditional distribution of $y_{j,t}$ given all the other responses at time t is equal to the condition distribution given the responses in $N(s_j)$, where $N(s_j)$ is a subset of V denoting the neighborhood of s_j. This means equivalence between the global and the local Markov properties that

$$\begin{aligned} f_{j_1}(y_{j_1,t}|\{y_{j_2,t} : s_{j_2} \in V\}, \theta_{j_1}) &= f_{j_1}(y_{j_1,t}|\{y_{j_2,t} : s_{j_2} \in N(s_{j_1})\}, \theta_{j_1}) \\ &= (2\pi\sigma^2)^{-1/2} \exp\{-(y_{j_1,t} - A_{j_1})^2/(2\sigma^2)\} \qquad (3.2.1) \end{aligned}$$

for $j_1 = 1, \ldots, p$, where $f_{j_1}(\cdot|\cdot, \theta_{j_1})$ is the conditional normal density of $y_{j_1,t}$ with parameter θ_{j_1},

$$A_{j_1} = A_{j_1}(\{y_{j_1,t} : s_{j_1} \in N(s_{j_1})\}, \theta_{j_1}) = \mu_{j_1} + \sum_{s_{j_2} \in N(s_{j_1})} \eta_{j_1 j_2}(y_{j_2,t} - \mu_{j_2}) \tag{3.2.2}$$

is the conditional mean of $y_{j_1,t}$ given its neighborhood, μ_j is the marginal mean of $y_{j,t}$, and $\theta_j = \{\sigma^2, \mu_j, \kappa_j, \eta_{jj} : v_j \in N(v_j)\}$. Under such a model, $y_{j_1,t}$ is conditional independent with $\{y_{j_3,t} : s_{j_3} \notin N(s_{j_1})\}$ given $\{y_{j_2,t} : s_{j_2} \in N(s_{j_1})\}$, and the parameters $\{\eta_{j_1 j_2}\}$ reflect the conditional dependence among $\mathbf{y}_t$.

Let $\mathcal{E}_h \subset \{1, \cdots, p\}^2$ for $h = 1, \cdots, H$ be the sets of interest, where $\mathcal{E}_{h_i} \cap \mathcal{E}_{h_j} = \emptyset$. We are interested in testing the conditional dependence structure of GMRF through $\{\eta_{jj}\}$:

$$H_0 : \tilde{g}_{h,j_1 j_2}(\eta_{j_1,j_2}) = c_h \text{ for all } (j_1, j_2) \in \mathcal{E}_h \text{ and all } h \text{ vs. } H_a : \text{Not } H_0 \tag{3.2.3}$$

for $h = 1, \cdots, H$, where $\tilde{g}_{h,j_1 j_2}(\cdot)$ is known functions which may depend on the indices (j_1, j_2), and $\{c_h\}$ are unknown constants. To simplify the notation, we drop the subscript j_1, j_2 in $\tilde{g}_{h,j_1 j_2}(\cdot)$ when there is no confusion. In the following, we provide several common dependency structures in spatial statistics which are included in the framework of (3.2.3). Suppose we observe a data on vertices $V = \{v_1, \ldots, v_p\} \subset \mathbf{R}^2$ on a regular grid over time, where $v_i \in \{(m_1, m_2) : 1 \le m_1 \le k_1, 1 \le m_2 \le k_2\}$ for $p = k_1 k_2$.

- Detection of neighborhood sizes is of great importance in GMRF models. One example is the four nearest and eight nearest neighborhood structures which are commonly used in practice, where for $j = (m_1 - 1)k_1 + m_2$,

$$\begin{aligned} N_4(s_j) &= \{(m_1, m_2 - 1), (m_1, m_2 + 1), (m_1 - 1, m_2), (m_1 + 1, m_2)\} \text{ and} \\ N_8(s_j) &= N_4(s_j) \cup \{(m_1 - 1, m_2 - 1), (m_1 - 1, m_2 + 1), (m_1 + 1, m_2 - 1), \\ &\quad (m_1 + 1, m_2 + 1)\} \end{aligned}$$

  respectively. To assess the validity of the neighborhood, we consider to test

$$
\begin{aligned}
H_0 &: \ N_w(s_j) \text{ is the neighborhood for any } j = 1, \cdots, p, \\
H_a &: \ N_w(s_j) \text{ is not the neighborhood for all } j = 1, \cdots, p,
\end{aligned}
\tag{3.2.4}
$$

where $w = 4$ or 8. For four nearest and eight nearest neighborhoods, the supports of Ω are

$$\mathcal{E}_{N_4} = \{(j,j), (j-1,j), (j+1,j), (j+k_1,j), (j-k_1,j) : 1 \le j \le p\} \text{ and}$$

$$\mathcal{E}_{N_8} = \mathcal{E}_{N_4} \cup \{(j-k_1+1,j), (j+k_1-1,j), (j+k_1+1,j), (j-k_1-1,j) : 1 \le j \le p\},$$

respectively. Therefore, (3.2.4) is a special case of (3.2.3) with $\tilde{g}(\eta_{j_1,j_2}) = \eta_{j_1,j_2}$ and $\mathcal{E}_1 = \mathcal{E}_{N_w}^c$:

$$H_0 : \eta_{j_1,j_2} = 0 \text{ for } (j_1,j_2) \in \mathcal{E}_{N_w}^c \ \text{ vs. } \ H_a : \text{Not } H_0. \tag{3.2.5}$$

- Given the neighborhood sizes, we are interested in the conditional dependence structures within each neighborhood. For isotropic dependence structures, the conditional dependence of $y_{j,t}$ between each variable in its neighborhood $\{y_{j,t} : s_j \in N(s_j)\}$ are the same. Under this case, $\eta_{j_1 j_2}$ in (3.2.2) are constant for all vertices $s_j \in N(s_j)$.

  Testing for the isotropic structure under four nearest neighborhood is equivalent to hypothesis (3.2.3) with $\tilde{g}_1(\eta_{j_1,j_2}) = \tilde{g}_2(\eta_{j_1,j_2}) = \eta_{j_1,j_2}$, $\mathcal{E}_1 = \mathcal{E}_{N_4}$, $\mathcal{E}_2 = \mathcal{E}_{N_4}^c$, $c_1 = c$ and $c_2 = 0$:

$$H_0 : \eta_{j_1,j_2} = \begin{cases} c & (j_1,j_2) \in \mathcal{E}_{N_w} \\ 0 & (j_1,j_2) \in \mathcal{E}_{N_w}^c \end{cases} \ \text{ vs. } \ H_a : \text{Not } H_0. \tag{3.2.6}$$

- In the case that the dependence structures are directional, the dependence parameters η_{jj} may be different between the horizontal and vertical neighborhoods of y_{ij}. We have

$$\text{Directional: } \ A_{j_1} = \mu_{j_1} + \eta_u \sum_{s_{j_2} \in N_u(s_{j_1})} (y_{j_2,t} - \mu_{j_2}) + \eta_v \sum_{s_{j_2} \in N_v(s_{j_1})} (y_{j_2,t} - \mu_{j_2}), \tag{3.2.7}$$

where $N_u(s_{j_1}) = \{s_{j_2} : |j_1 - j_2| = k_1\}$ and $N_v(s_{j_1}) = \{s_{j_2} : |j_1 - j_2| = 1\}$ are the neighborhoods of s_j in horizontal and vertical directions, respectively. Therefore, setting $\tilde{g}_1(\eta_{l_1,j_2}) = \tilde{g}_2(\eta_{l_1,j_2}) = \tilde{g}_3(\eta_{l_1,j_2}) = \eta_{l_1,j_2}$, $\mathcal{E}_1 = \{(j_1, j_2) : |j_1 - j_2| = 1\}$, $\mathcal{E}_2 = \{(j_1, j_2) : |j_1 - j_2| = k_1\}$ and $\mathcal{E}_3 = \mathcal{E}_{N_4}^c$ with $c_3 = 0$ in the hypothesis (3.2.3) is for testing the directional dependence structure under four nearest neighborhood:

$$H_0 : \tilde{g}(\eta_{j_1,j_2}) = \begin{cases} \eta_u = c_1 & (j_1, j_2) \in \mathcal{E}_1 \\ \eta_v = c_2 & (j_1, j_2) \in \mathcal{E}_2 \\ \eta_{\{N_u \cup N_v\}^c} = 0 & (j_1, j_2) \in \mathcal{E}_3 \end{cases} \text{ vs. } H_a : \text{Not } H_0 \tag{3.2.8}$$

- Another structure is based on the distance between two vertices, where $\eta_{j_1 j_2}$ is reciprocal to the distance between s_{j_1} and s_{j_2}. The corresponding expression of A_{j_1} in (3.2.2) is

$$\text{Distance: } A_{j_1} = \mu_{j_1} + \frac{\eta}{d_{j_1 j_2}} \sum_{s_{j_2} \in N(s_{j_1})} (y_{j_2,t} - \mu_{j_2}), \tag{3.2.9}$$

where $d_{j_1 j_2} = \|s_{j_1} - s_{j_2}\|$ is the Euclidean distance between s_{j_1} and s_{j_2}. It is clear that testing for the distance based dependence structure coincides with the hypothesis (3.2.3) with $\tilde{g}_1(\eta_{j_1,j_2}) = \eta_{j_1,j_2}/d_{j_1 j_2}$, $\tilde{g}_2(\eta_{j_1,j_2}) = \eta_{j_1,j_2}$, $\mathcal{E}_1 = \mathcal{E}_{N_w}$ and $\mathcal{E}_2 = \mathcal{E}_{N_w}^c$:

$$H_0 : \tilde{g}(\eta_{j_1,j_2}) = \begin{cases} \eta_{j_1,j_2}/d_{j_1,j_2} = c & (j_1, j_2) \in \mathcal{E}_1 \\ \eta_{j_1,j_2} = 0 & (j_1, j_2) \in \mathcal{E}_2 \end{cases} \text{ vs. } H_a : \text{Not } H_0 \tag{3.2.10}$$

It can be shown that the joint distribution of $\mathbf{y}_t$ under (3.2.1) and (3.2.2) is $N(\mu, (I_p - C)^{-1}M)$, where $\mu = (\mu_1, \cdots, \mu_p)'$, $C = (\eta_{j_1 j_2})$, $M = \text{diag}(\sigma^2)$ and I_p is the $p \times p$ identity matrix. Let $\mathbf{\Sigma} = (\sigma_{j_1,j_2})$ be the covariance of $\mathbf{y}_t$, and $\mathbf{\Omega} = (\omega_{j_1,j_2}) = \mathbf{\Sigma}^{-1} = M^{-1}(I_p - C)$ be the precision matrix of $\mathbf{y}_t$. Note that C is a symmetric matrix with zero diagonal entries and $\omega_{j_1,j_2} = -\sigma^{-2}\eta_{j_1,j_2}$ for $j_1 \neq j_2$. If all $\eta_{j_1,j_2} = 0$, then the joint distribution of $\mathbf{y}_t$ is $N(\mu, M)$ which corresponds to an independence model. Testing neighborhood dependence structures as in (3.2.5), (3.2.6), (3.2.8)

and (3.2.10) are equivalent to testing the structure of the precision matrix $\mathbf{\Omega}$ as

$$H_0 : g_{h,j_1j_2}(\omega_{j_1,j_2}) = c_h \text{ for all } (j_1, j_2) \in \mathcal{E}_h \text{ and all } h \text{ vs. } H_a : \text{Not } H_0 \tag{3.2.11}$$

for some known function $g_{h,j_1j_2}(\cdot)$ and unknown constants $\{c_h\}$.

The above hypotheses (3.2.11) on the precision matrix can be also applied to other covariance classes. One of the most popular variogram model in geostatistics is the Matérn covariance class.The Matérn covariance function between vertices s_{j_1} and s_{j_2} is defined as

$$\sigma_{j_1j_2} = \mathrm{Cov}(y_{j_1,t}, y_{j_2,t}) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)}(\rho d_{j_1j_2})^{\nu} K_{\nu}(\rho d_{j_1j_2}). \tag{3.2.12}$$

Here, $K_\nu(\cdot)$ is the modified Bessel function of the second kind and order $\nu > 0$, ρ is a scaling parameter and σ^2 is the marginal variance. [99] showed that the inverse of the Matérn covariance is approximately sparse, and the Gaussian fields with Matérn covariance (3.2.12) can be well approximated by GMRF models. Due to the sparsity of the precision matrix brought by the Markov property, they suggested to use GMRF representation to compute the Gaussian fields with Matérn covariance. Meanwhile, we could test the structure of the precision matrix that serves a way to check the Matérn covariance structure. When $\nu = 1$, [99] showed that GMRF representation for the Matérn fields have the precision matrix in the form that $\omega_{jj} = 4 + c^2$, $\omega_{j_1j_2} = -2c$ for $|j_1 - j_2| = 1$ or k where $c = 4 + 8/r^2$ for the range parameter $r > 0$, $\omega_{j_1j_2} = 2$ for $|j_1 - j_2| = k \pm 1$ and $\omega_{j_1j_2} = 1$ for $|j_1 - j_2| = 2$ or $2k$. To test for the Matérn covariance with $\nu = 1$, we set the hypothesis (3.2.3) in such a way that

$$\begin{aligned}
g_1(\omega_{j_1,j_2}) &= \sqrt{\omega_{j_1,j_2} - 4} \quad \text{for } \mathcal{E}_1 = \{(j_1, j_2) : j_1 = j_2\} \text{ with } c_1 = c; \\
g_2(\omega_{j_1,j_2}) &= -\omega_{j_1,j_2}/2, \quad \text{for } \mathcal{E}_2 = \{(j_1, j_2) : |j_1 - j_2| = 1 \text{ or } k\} \text{ with } c_2 = c; \\
g_3(\omega_{j_1,j_2}) &= \omega_{j_1,j_2}, \quad \text{for } \mathcal{E}_3 = \{(j_1, j_2) : |j_1 - j_2| = k \pm 1\} \text{ with } c_3 = 2; \\
g_4(\omega_{j_1,j_2}) &= \omega_{j_1,j_2}, \quad \text{for } \mathcal{E}_4 = \{(j_1, j_2) : |j_1 - j_2| = 2 \text{ or } 2k\} \text{ with } c_4 = 1.
\end{aligned}$$

The hypothesis of interest is a special case of (3.2.11) that

$$H_0 : g(\omega_{j_1,j_2}) = \begin{cases} \sqrt{\omega_{j_1,j_2} - 4} = c & (j_1, j_2) \in \mathcal{E}_1 \\ -\omega_{j_1,j_2}/2 = c & (j_1, j_2) \in \mathcal{E}_2 \\ \omega_{j_1,j_2} = 2 & (j_1, j_2) \in \mathcal{E}_3 \\ \omega_{j_1,j_2} = 1 & (j_1, j_2) \in \mathcal{E}_4 \\ \omega_{j_1,j_2} = 0 & \text{otherwise} \end{cases} \quad \text{vs. } H_a : \text{Not } H_0. \tag{3.2.13}$$

3.3 Methodology

In this section, we propose a testing procedure for hypothesis (3.2.11) that is adaptive to the time dependence among observations. Let $\mathcal{Y}_n = \{\mathbf{y}_t\}_{t=1}^n$ be a stationary p-dimensional time series, where each observation $\mathbf{y}_t$ follows the GMRF model (3.2.1) and (3.2.2). Let $\bar{\mathbf{y}} = (\bar{y}_1, \ldots, \bar{y}_p)'$ be the sample mean of $\{\mathbf{y}_t\}_{t=1}^n$. Notice that the sample version of (3.2.2) is the p node-wise linear regressions

$$y_{j_1,t} = \bar{y}_{j_1} + \sum_{j_2 \neq j_1} \alpha_{j_1,j_2}(y_{j_2,t} - \bar{y}_{j_2}) + \epsilon_{j_1,t} \tag{3.3.1}$$

for $t = 1, \ldots, n$ and $j_1 = 1, \ldots, p$. Let $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_p)^{\mathrm{T}}$ and $\mathbf{V} = \mathrm{Cov}(\boldsymbol{\epsilon}) = (v_{j_1,j_2})_{p\times p}$. Denote $\boldsymbol{\alpha}_j = (\alpha_{j,1}, \ldots, \alpha_{j,j-1}, -1, \alpha_{j,j+1}, \ldots, \alpha_{j,p})^{\mathrm{T}}$, and let $\widehat{\boldsymbol{\alpha}}_j$ be the Lasso estimator [36] for $\boldsymbol{\alpha}_j$ as

$$\widehat{\boldsymbol{\alpha}}_j = \underset{\boldsymbol{\alpha}\in\Theta_j}{\arg\min} \left[\frac{1}{n} \sum_{t=1}^{n} \{\boldsymbol{\alpha}^{\mathrm{T}}(\mathbf{y}_t - \bar{\mathbf{y}})\}^2 + 2\lambda_j |\boldsymbol{\alpha}|_1 \right], \tag{3.3.2}$$

where $\Theta_j = \{\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)^{\mathrm{T}} \in \mathbb{R}^p : \alpha_j = -1\}$ and λ_j is the tuning parameter at the order $\sqrt{\log(p)/n}$. Let

$$\widehat{\boldsymbol{\epsilon}}_t = (\widehat{\epsilon}_{1,t}, \ldots, \widehat{\epsilon}_{p,t})^{\mathrm{T}} \text{ with } \widehat{\epsilon}_{j,t} = -\widehat{\boldsymbol{\alpha}}_j^{\mathrm{T}}(\mathbf{y}_t - \bar{\mathbf{y}}) \tag{3.3.3}$$

be the residuals from fitting (3.3.1) by Lasso. Let $\widetilde{\mathbf{V}} = (\widetilde{v}_{j_1,j_2})_{p\times p}$ be the sample covariance of $\{\widehat{\boldsymbol{\epsilon}}_t\}_{t=1}^n$, where $\widetilde{v}_{j_1,j_2} = n^{-1}\sum_{t=1}^n \widehat{\epsilon}_{j_1,t}\widehat{\epsilon}_{j_2,t}$. Following [111], we employ the de-biased estimator

$$\widehat{v}_{j_1,j_2} = \begin{cases} -\dfrac{1}{n}\sum_{t=1}^{n}(\widehat{\epsilon}_{j_1,t}\widehat{\epsilon}_{j_2,t} + \widehat{\alpha}_{j_1,j_2}\widehat{\epsilon}^2_{j_2,t} + \widehat{\alpha}_{j_2,j_1}\widehat{\epsilon}^2_{j_1,t}), & j_1 \neq j_2; \\ \dfrac{1}{n}\sum_{t=1}^{n}\widehat{\epsilon}_{j_1,t}\widehat{\epsilon}_{j_2,t}, & j_1 = j_2. \end{cases} \tag{3.3.4}$$

for the error variance $\mathbf{V}$. We then estimate $g(\omega_{j_1,j_2})$ as

$$\widehat{g}(\omega_{j_1,j_2}) = g(\widehat{\omega}_{j_1,j_2}) \text{ for } \widehat{\omega}_{j_1,j_2} = \frac{\widehat{v}_{j_1,j_2}}{\widehat{v}_{j_1,j_1}\widehat{v}_{j_2,j_2}} \tag{3.3.5}$$

for any j_1 and j_2.

To construct the test statistic, we also need to estimate the unknown constants $\{c_h\}$ under H_0 of (3.2.11). For any $h = 1, \cdots, H$, let

$$\widehat{c}_h = \sum_{(j_1,j_2)\in\mathcal{E}_h} g(\widehat{\omega}_{j_1,j_2})/q_h,$$

where $q_h = |\mathcal{E}_h|$ is the cardinality of $\mathcal{E}_h$. The test statistic for the hypothesis (3.2.11) is constructed as

$$T_n = \max_h \max_{(j_1,j_2)\in\mathcal{E}_h} \sqrt{n}|g(\widehat{\omega}_{j_1,j_2}) - \widehat{c}_h|. \tag{3.3.6}$$

Under some regularity conditions, by using the similar argument of [111], it gives the asymptotic expansion of $g(\widehat{\omega}_{j_1,j_2})$ for any first order continuous function $g(\cdot)$.

$$g(\widehat{\omega}_{j_1,j_2}) - g(\omega_{j_1,j_2}) = -g'(\omega_{j_1,j_2})\frac{\delta_{j_1,j_2}}{v_{j_1,j_1}v_{j_2,j_2}} + o_p\{(n\log p)^{-1/2}\},$$

and $\widehat{c}_h - c_h = o_p\{s^{1/2}(nq_h)^{-1/2}\} = o_p(n^{-1/2})$ for any $h = 1, \cdots, H$ if $q_h \gg s$, where $\delta_{j_1,j_2} = n^{-1}\sum_{t=1}^{n}(\epsilon_{j_1,t}\epsilon_{j_2,t} - v_{j_1,j_2})$, and $o_p\{(n\log p)^{-1/2}\}$ is a uniform higher order term for all j_1 and j_2. Those results imply that $T_n = \sqrt{n}|\mathbf{\Psi}|_\infty + o_p(1)$, where $\mathbf{\Psi} = \mathbf{G} \circ \{\text{diag}(\mathbf{V})\}^{-1}\mathbf{\Delta}\{\text{diag}(\mathbf{V})\}^{-1}$, $\mathbf{\Delta} = -n^{-1}\sum_{t=1}^{n}\mathbf{\Delta}_t$, $\mathbf{G} = (g'(\omega_{j_1,j_2}))_{p\times p}$ and $\mathbf{\Delta}_t = \boldsymbol{\epsilon}_t\boldsymbol{\epsilon}_t^{\mathrm{T}} - \mathbf{V}$. Let $\boldsymbol{\varsigma}_t$ be the vectorization of $\mathbf{G} \circ \{\text{diag}(\mathbf{V})\}^{-1}\mathbf{\Delta}_t\{\text{diag}(\mathbf{V})\}^{-1}$, where each element of $\boldsymbol{\varsigma}_t$ takes the form $\frac{g'(\omega_{j_1,j_2})}{v_{j_1,j_1}v_{j_2,j_2}}(\epsilon_{j_1,t}\epsilon_{j_2,t} -$

v_{j_1,j_2}). Then, T_n is asymptotically equally distributed as $\sqrt{n}|\sum_{t=1}^n \boldsymbol{\varsigma}_t/n|_\infty$. Due to the form of partial sum, we can approximate the distribution of T_n via Gaussian approximation.

Let $\mathbf{W}$ be the long run covariance of $\{\boldsymbol{\varsigma}_t\}_{t=1}^n$, which takes the form

$$\mathbf{W} = \mathbb{E}\Big\{\Big(\frac{1}{n^{1/2}}\sum_{t=1}^n \boldsymbol{\varsigma}_t\Big)\Big(\frac{1}{n^{1/2}}\sum_{t=1}^n \boldsymbol{\varsigma}_t\Big)^{\mathrm{T}}\Big\}. \tag{3.3.7}$$

It is clear the distribution of T_n is related to the long run covariance $\mathbf{W}$. In practice, we estimate $\boldsymbol{\varsigma}_t$ by plugging in $\widehat{\omega}_{j_1,j_2}$, $\widehat{v}_{j_1,j_2}$ and residuals $\widehat{\epsilon}_{j,t}$. Denote this estimator by

$$\widehat{\boldsymbol{\varsigma}}_t = \mathrm{Vec}\{(\frac{g'(\widehat{\omega}_{j_1,j_2})}{\widehat{v}_{j_1,j_1}\widehat{v}_{j_2,j_2}}(\widehat{\epsilon}_{j_1,t}\widehat{\epsilon}_{j_2,t} - \widehat{v}_{j_1,j_2}))_{p\times p}\}.$$

We propose a kernel-type estimator suggested by [113] for $\mathbf{W}$ as

$$\widehat{\mathbf{W}} = \sum_{k=-n+1}^{n-1} \mathcal{K}\Big(\frac{k}{S_n}\Big)\widehat{\boldsymbol{\Gamma}}_k \ \text{ for } \ \widehat{\boldsymbol{\Gamma}}_k = \begin{cases} \dfrac{1}{n}\displaystyle\sum_{t=k+1}^{n} \widehat{\boldsymbol{\varsigma}}_t\widehat{\boldsymbol{\varsigma}}_{t-k}^{\mathrm{T}}, & k \geq 0; \\ \dfrac{1}{n}\displaystyle\sum_{t=-k+1}^{n} \widehat{\boldsymbol{\varsigma}}_{t+k}\widehat{\boldsymbol{\varsigma}}_t^{\mathrm{T}}, & k < 0. \end{cases} \tag{3.3.8}$$

where S_n is the bandwidth, $\mathcal{K}(\cdot)$ is a symmetric kernel function that is continuous at 0 and satisfying $\mathcal{K}(0) = 1$, $|\mathcal{K}(u)| \leq 1$ for any $u \in \mathbb{R}$, and $\int_{-\infty}^{\infty} \mathcal{K}^2(u)du < \infty$.

Let $\widehat{\boldsymbol{\xi}} \sim N(\mathbf{0}, \widehat{\mathbf{W}})$ for $\widehat{\mathbf{W}}$ specified in (3.3.8). Recall that $\mathcal{Y}_n = \{\mathbf{y}_t\}_{t=1}^n$ denotes the data of the sample. By Gaussian approximation results, it can be shown that

$$\sup_{x>0} \big|\mathbb{P}\big(T_n > x\big) - \mathbb{P}\big(|\widehat{\boldsymbol{\xi}}|_\infty > x|\mathcal{Y}_n\big)\big| \xrightarrow{p} 0 \ \text{ as } \ n \to \infty, \tag{3.3.9}$$

which indicates that the distribution of T_n can be approximated by that of $|\widehat{\boldsymbol{\xi}}|_\infty$.

Monte Carlo simulation can be used to obtain the distribution of $|\widehat{\boldsymbol{\xi}}|_\infty$ given the data $\mathcal{Y}_n$. Let $\widehat{\boldsymbol{\xi}}_1, \ldots, \widehat{\boldsymbol{\xi}}_M$ be i.i.d. drawn from $N(\mathbf{0}, \widehat{\mathbf{W}})$. Let

$$\widehat{F}_M(x) = \frac{1}{M}\sum_{m=1}^{M} \mathbb{I}\big\{|\widehat{\boldsymbol{\xi}}_m|_\infty \leq x\big\}$$

be the empirical distribution of $\{|\widehat{\boldsymbol{\xi}}_1|_\infty, \ldots, |\widehat{\boldsymbol{\xi}}_M|_\infty\}$. Based on the result (3.3.9),

$$\widehat{q}_\alpha = \inf\{x : \widehat{F}_M(x) \geq 1 - \alpha\}. \tag{3.3.10}$$

is an estimate of the upper α quantile of T_n. The α level test rejects the null hypothesis of (3.2.11) if $T_n > \widehat{q}_\alpha$.

The proposed inference procedure is summarized as the following algorithm.

Algorithm

Input: Observations $\mathbf{y}_t = (y_{1,t}, ..., y_{p,t})^{\mathrm{T}} \in \mathbb{R}^p$ for $t = 1, ..., n$ and a pre-specified level $\alpha \in (0, 1)$.

Procedure:

Step 1. Perform p node-wise regressions (3.3.1) $y_{j_1,t} = \bar{y}_{j_1} + \sum_{j_2 \neq j_1} \alpha_{j_1,j_2}(y_{j_2,t} - \bar{y}_{j_2}) + \epsilon_{j_1,t}$ by Lasso, where $t = 1, \ldots, n$. Then, get residuals $\widehat{\boldsymbol{\epsilon}}_t = (\widehat{\epsilon}_{1,t}, ..., \widehat{\epsilon}_{p,t})^{\mathrm{T}}$ where $\widehat{\epsilon}_{j,t} = -\widehat{\boldsymbol{\alpha}}_j^{\mathrm{T}}(\mathbf{y}_t - \bar{\mathbf{y}})$. Obtain the de-biased estimator $\widehat{\mathbf{V}}$ of $\mathrm{cov}(\boldsymbol{\epsilon}_t)$ as (3.3.4), and estimate the precision matrix $\boldsymbol{\Omega}$ of $\mathbf{y}_t$ by $\widehat{\boldsymbol{\Omega}} = \{\mathrm{diag}(\widehat{\mathbf{V}})^{-1}\}\widehat{\mathbf{V}}\{\mathrm{diag}(\widehat{\mathbf{V}})^{-1}\}$.

Step 2. Obtain the test statistic $T_n = \max_h \max_{(j_1,j_2)\in\mathcal{E}_h} \sqrt{n}|g(\widehat{\omega}_{j_1,j_2}) - \widehat{c}_h|$ in (3.3.6), where $\widehat{\omega}_{j_1,j_2} = \widehat{v}_{j_1,j_1}^{-1}\widehat{v}_{j_1,j_2}\widehat{v}_{j_2,j_2}^{-1}$ and $\widehat{c}_h = \sum_{(j_1,j_2)\in\mathcal{E}_h} g(\widehat{\omega}_{j_1,j_2})/q_h$.

Step 3. Let $\mathbf{A}$ be an $n \times n$ matrix whose (ℓ_1, ℓ_2) element is $\mathcal{K}(|\ell_1 - \ell_2|/S_n)$, and generate n-dimensional Gaussian random vector $(g_1, \ldots, g_n)^{\mathrm{T}}$ with mean zero and covariance $\mathbf{A}$.

Step 4. Calculate $\widehat{\boldsymbol{\varsigma}}_t = \mathrm{Vec}\{(\frac{g'(\widehat{\omega}_{j_1,j_2})}{\widehat{v}_{j_1,j_1}\widehat{v}_{j_2,j_2}}(\widehat{\epsilon}_{j_1,t}\widehat{\epsilon}_{j_2,t} - \widehat{v}_{j_1,j_2}))_{p\times p}\}$ and $\widehat{\boldsymbol{\xi}} = n^{-1/2}\sum_{i=1}^n g_t\widehat{\boldsymbol{\varsigma}}_t$.

Step 5. Repeat Steps 3 and 4 M times to obtain i.i.d. samples $\widehat{\boldsymbol{\xi}}_1, ..., \hat{\boldsymbol{\xi}}_M$ from $N(\mathbf{0}, \widehat{\mathbf{W}})$ for $\widehat{\mathbf{W}}$ given in (3.3.8). Calculate $\widehat{q}_\alpha = \inf\{x : \widehat{F}_M(x) \geq 1 - \alpha\}$, where $\widehat{F}_M(x)$ is the empirical distribution of $(|\widehat{\boldsymbol{\xi}}_1|_\infty, ..., |\widehat{\boldsymbol{\xi}}_M|_\infty)^{\mathrm{T}}$. Reject H_0 in (3.2.11) if $T_n > \widehat{q}_\alpha$.

3.4 Application to the Google Flu Trend

We applied our inference procedure in Algorithm 1 to study data from the Google Flu Trend to explore the intrinsic geographical dependency among regions and to provide some interesting epidemiology insights. Specifically, for both United States and European data, we considered different neighborhood sizes and also investigated different types of dependence structures such as isotropic and distance-based in Section 3.2.

3.4.1 Backgrounds

Understanding of disease transmission is the key question in epidemiology to recognize pattern of health events and develop prevention systems. A good example of studies is about flu pandemic. The 1918 H1N1 flu pandemic was uniquely fatal, which infected one third of world populations and caused 50 millions deaths [114]. Back in 2009, H1N1 flu virus involved another epic pandemic, which caused hundreds of thousands deaths [115]. Besides, annual influenza outbreak caused multiple deaths [116]. For these reasons, public health organizations offer influenza surveillance information. For example, the Centers for Disease Control Prevention (CDC) and the European Influenza Surveillance Scheme (EISS) have released weekly influenza surveillance reports for a few decades. However, these reports have 1-2 week reporting lag, so there have been attempts to create faster flu detection system. A particular example is GFT.

GFT was a web service operated by Google to offer flu activities estimation available at https://www.google.org/flutrends/about/. GFT uses individual Google search queries to provide faster detection, while the traditional weekly based benchmarks by CDC and EISS used on both virologic and clinical data. GFT is able to report influenza-like illness (ILI) 1-2 weeks ahead of the official ILI data from CDC [117]. To identify locations of inquires, users' IP address were used.

The original model in GFT was the following:

$$\operatorname{logit} P(t) = \beta_0 + \beta_1 \operatorname{logit} Q(t) + \epsilon$$

where $P(t)$ is the percentage of ILI Physician visits, and $Q(t)$ is ILI-related Query Fraction. $Q(t)$ is calculated by 45 top scoring queries, determined by cross validation, among 50 millions queries. Those queries do not have to be ILI-related [117].

GFT has attracted many researchers' attention. [117] claimed GFT prediction were 97% accurate comparing with GFT ILI. [118] examined that GFT tracked information about the 2009 flu pandemic in the United States. In February 2010, the CDC identified influenza cases spiking in the mid-Atlantic region of the United States. Surprisingly, Google search queries about flu symptoms was able to show that same spike two weeks prior to the CDC report being released. Moreover, GFT has been used in applied statistical papers. For example, [119] used GFT to understand effects of spatial parameters on influenza's transmission by the susceptible-infected-recovered-susceptible (SIRS) model, and [120] used GFT to develop influenza forecasting model using generalized ARMA model. [121] applied a humidity driven SIRS model jointly with either the ensemble adjustment Kalman filter or a particle filter to estimate key epidemiology parameters.

Our testing procedure can be applied in GFT. The key difference from the papers above is whether the network model framework is employed or not. Specifically, both [119] and [121] focused on estimating parameters in SIRS model under Bayesian framework, and [120] focused on time-series model selection and forecasting. The models in [119] focused on separating intrastate dynamics from interstate dynamics using time-varying covariates. [121] focused on estimating parameters in SIRS model. On the other hand, by employing our procedure, we can do inference on any type of interstate dependence structure in feasible time.

3.4.2 Data Pre-processing

Similar to [119], we focused on data from the 48 mainland states plus Washington D.C. for the United States. Data from these 49 areas have complete weekly records from Dec 2nd, 2007 to Aug 9th, 2015. We also studied data from the 73 areas from 7 European countries: Austria, Belgium, France, Germany, Netherlands, Spain, and Switzerland. These 73 areas possess complete weekly records from Dec 2nd, 2007 to Aug 9th, 2015. The geographical regions under considerations were

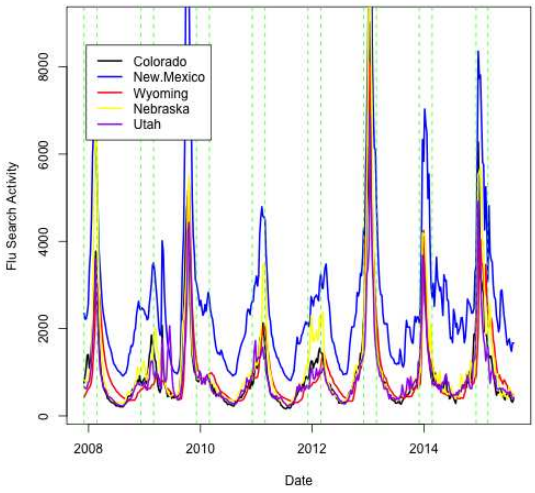
displayed in Figure 3.1 in Section 3.1. In summary, data for the United States can be abstracted as a network with 49 vertices evolving for 402 time points, while the data for Europe can be modeled as a network with 73 vertices evolving for 360 time points.

For data pre-processing, we regressed the log-transformed data against time to detrend and employed the first order difference procedure to remove the seasonality [122]. Specifically, for data at the area r and time t, $x_{t,r}$, we regressed $\log(x_{t,r}+1)$ against t to obtain detrended data $y_{t,r}$ and took $\epsilon_{t,r} = y_{t,r} - y_{t-1,r}$. Inspecting the autocorrelation and partial autocorrelation plots of $\epsilon_{t,r}$ for each area, the stationary assumption was reasonably satisfied. See Figure 3.2 as an example of state of Colorado. In addition, we standardized $\epsilon_{t,r}$ within each area and applied our proposed Algorithm 1 to the processed data.

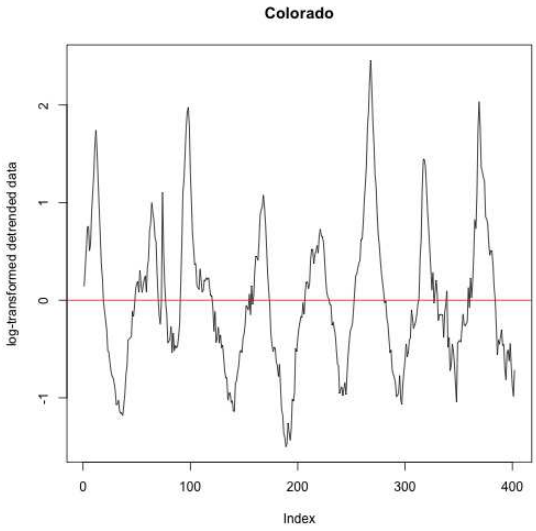
3.4.3 Data Analysis

Our goal is to explore and learn the structures of dependency among the 49 areas in the United States and the 73 areas in the Europe based on the temporal data from Google Flu Trend. In addition, we also studied the dependency structures within some pre-defined regions to gain more insights. Specially, we considered the four US Census regions: West, Mid-West, South, and North-east; and for Europe, we focused on four regions: Belgium-Netherlands, France, Germany, and Spain. Three types of dependency structures were studied. Those are the neighborhood, isotropic and distance as defined in Section 3.2.

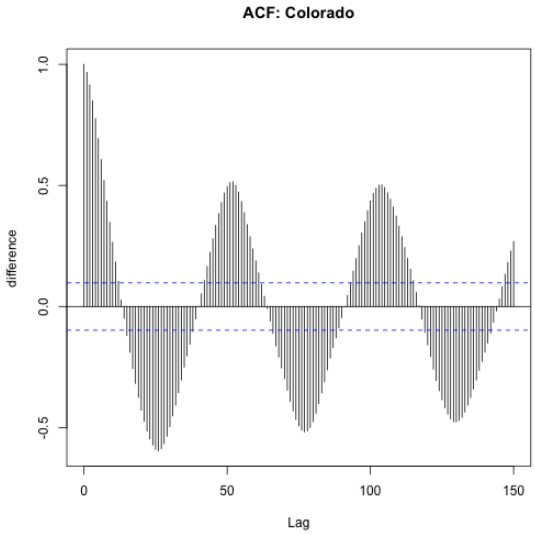
Knowing the neighborhood size is a key to investigate interstate transmission in epidemiology. Furthermore, it is not uncommon to investigate whether interstate effect is identical by regions and the effects depends on the distance between two locations. For these reasons, we first performed neighborhood test for five different size of neighborhoods, then we performed isotropic and distance tests if we failed to reject the neighborhood test. The null hypotheses are:



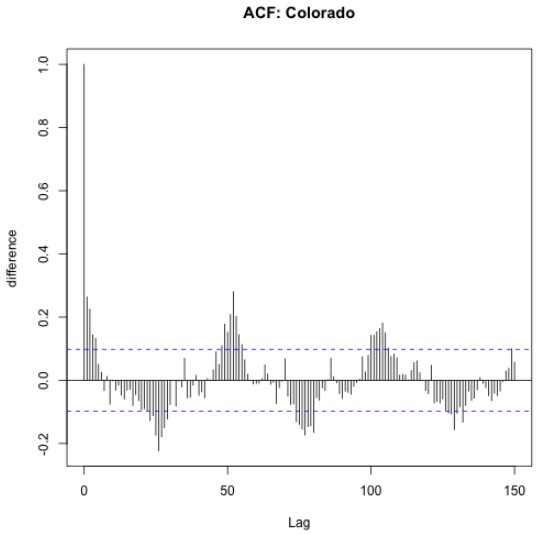
(a) GFT for Colorado and its neighboring states (From 2007-2015). Green dashed lines indicate the first week of December and the last week of February.



(b) Detrended data of log-transformed GFT (From 2007-2015)



(c) Auto correlation function for detrended data (From 2007-2015)



(d) Auto correlation function for first order differenced detrended data (From 2007-2015)

Figure 3.2: Plots displaying the pre-processing of Google Flu Trend data for the state of Colorado, USA.

$$\text{(Neighborhood) } H_0 : \omega_{i,j} = 0 \; \forall (i,j) \notin N$$

$$\text{(Isotropic) } H_0 : \omega_{i,j} = \begin{cases} c & \forall (i,j) \in N \\ 0 & \forall (i,j) \notin N \end{cases}$$

$$\text{(Distance) } H_0 : \omega_{i,j} = \begin{cases} c/d_{i,j} & \forall (i,j) \in N \\ 0 & \forall (i,j) \notin N \end{cases}$$

and their alternatives are H_a : Not H_0 where N indicates a set of neighborhoods. Similarly, we performed neighborhood and isotropic tests with geographical neighborhoods on sub-regions. We used the significance level $\alpha = 0.01$.

We used the following settings for this analysis. We used the tuning parameter λ for de-biased Lasso from our empirical sizes simulation in Section 3.5, specifically, $\lambda = \sqrt{2cn^{-1} \log p}$ where c was determined by the simulation for (D2) structure in Section 3.5 on a 10×10 regular grid, which represented the dependency structure of 100×100 GMRF. For example, Figures 3.3 and 3.4 display the square root of absolute value of precision matrix estimate entries, $\sqrt{|\omega_{i,j}|}$. The tests were performed by the wild bootstrap with 5000 replications.

Neighborhoods of Different Size

We considered five different sizes of neighborhood: the geographical neighborhood and four "distance-based" neighborhoods. These are commonly considered in epidemiology. For example, [119] estimated an interstate susceptibility from neighboring states. Similarly, distance could be an important factor in disease transmission because of indirect transmission by air particles, vehicles, and vectors [123].

First of all, the geographical neighborhood, denoted by N_{geo}, includes its neighboring regions on maps. For example, Figure 3.5 displays neighboring states of Colorado and neighboring regions of Îll-de-France: The neighboring states of Colorado are Arizona, Kansas, Oklahoma, Nebraska, New Mexico, Utah, and Wyoming, while the neighboring regions of Îll-de-France are upper-Normandy, Picardy, Champagne-Ardenne, Burgundy, and Centre-Val de Loire.

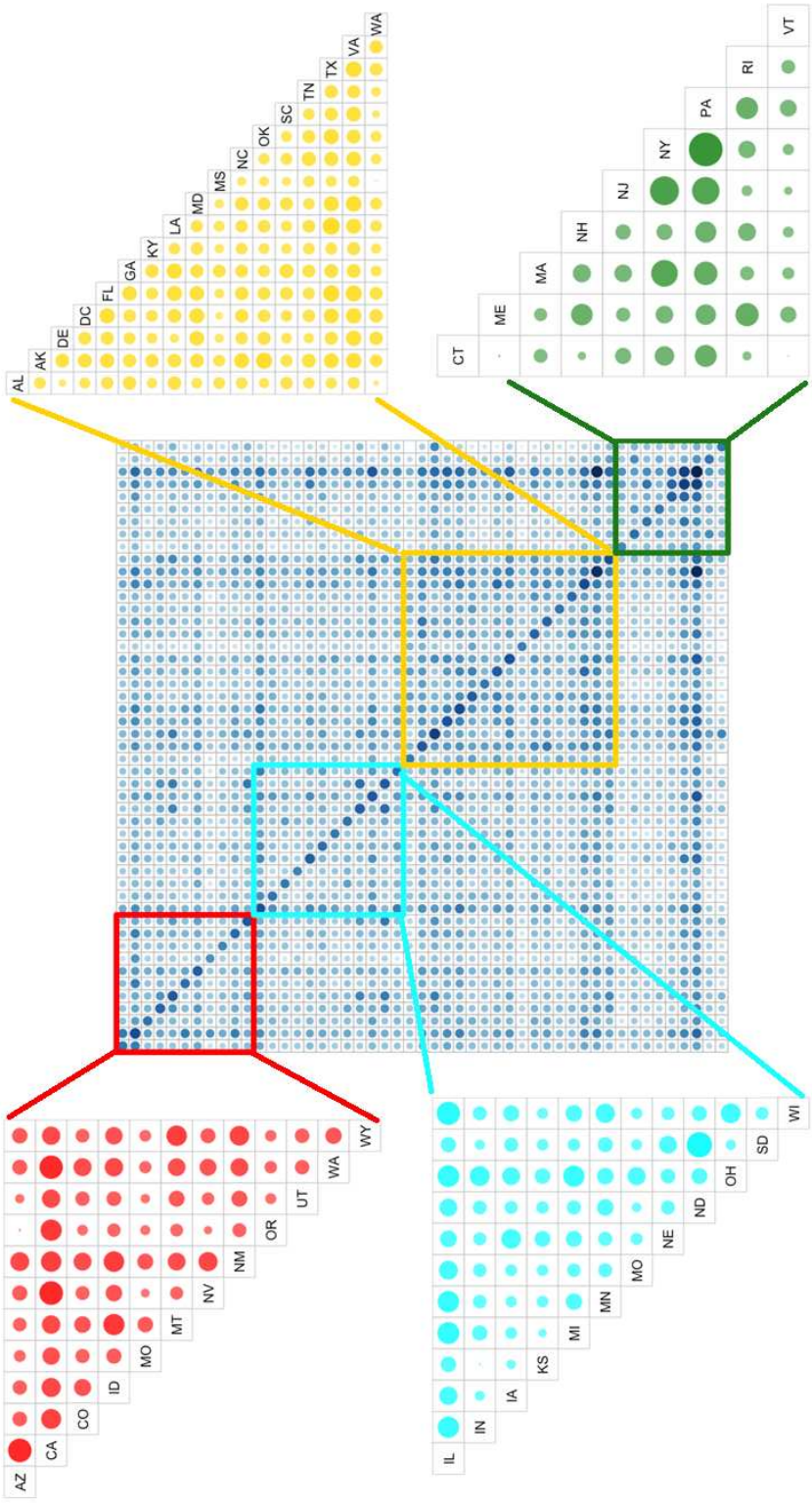


Figure 3.3: Estimated precision matrix of USA states (From 2007-2015): square root of absolute values of each entry, $\sqrt{|\omega_{i,j}|}$.

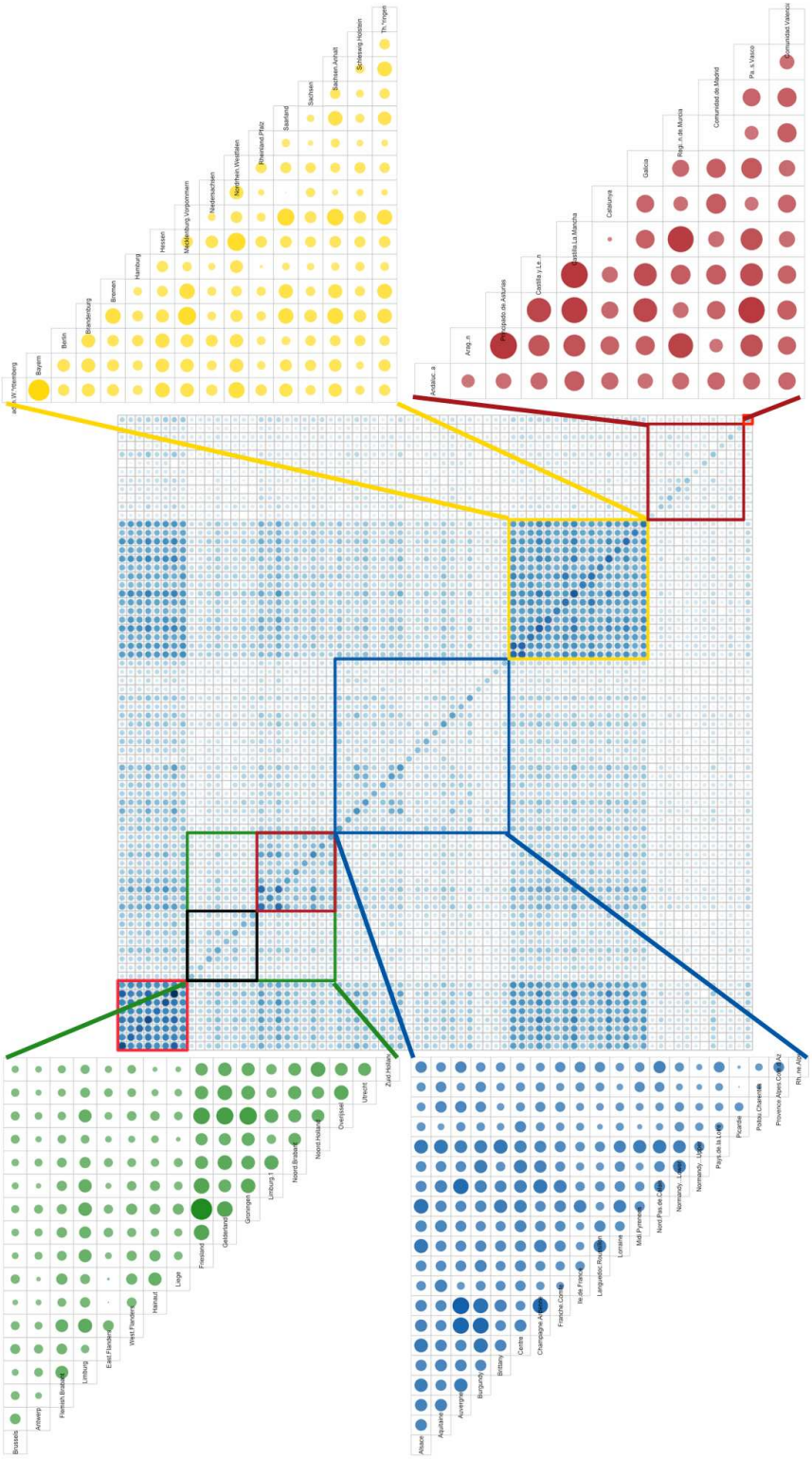


Figure 3.4: Estimated precision matrix of European provinces (From 2007-2015): square root of absolute values of each entry, $\sqrt{|\omega_{i,j}|}$.

(a) Colorado **(b)** Îll de France

Figure 3.5: Neighboring Regions (red) of selected state and region (blue). French map displays départments of France, which is a subdivision of regions of France.

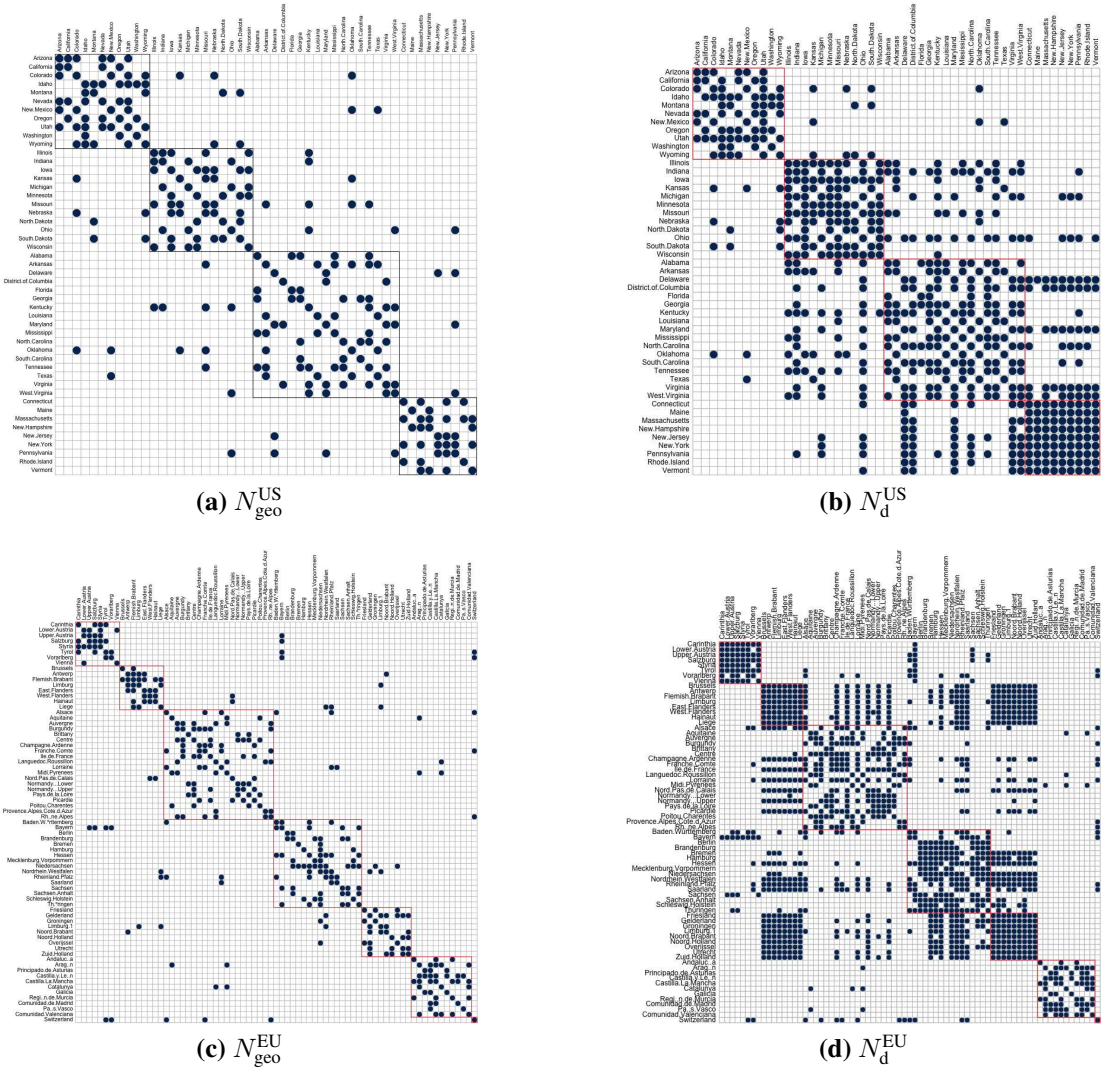
On the other hand, the "distance based" neighborhood, denoted by $N_{\mathrm{d},\alpha}$, defined by the distance between the geographic center of two regions:

$$N_{\mathrm{d},\alpha} = \{(i,j) : d_{i,j} < 2.5 d_\alpha \text{ where } d_\alpha \text{ is } \alpha\text{th percentile of } \{d_{i,j}\}_{i \neq j}\}. \tag{3.4.1}$$

We considered $N_{\mathrm{d},5}, N_{\mathrm{d},7.5}, N_{\mathrm{d},10}$, and $N_{\mathrm{d},15}$. We calculated regional distances using their geographical centers by `geosphere::distm` function in `R`. The coordinates of the US states' centers were acquired from `state.center` data set, whereas the coordinates for European regions were obtained from Dutch, English, French, German, and Spanish Wikipedia. Figure 3.6 displays adjacency matrices of N_{geo} and $N_{\mathrm{d},5}$ for both US and Europe. Table 3.1 displays the distance limit of $N_{\mathrm{d},\alpha}$.

3.4.4 Main Findings

In this section, we summarize the testing results from GFT. We observed a few interesting findings for overall US and Europe. Table 3.2 displays p-values of the tests for both US and Europe. For US, it was likely to have $N_{\mathrm{d},10}$ or $N_{\mathrm{d},15}$. We found the GFT of US states may connect



(a) $N_{\text{geo}}^{\text{US}}$ **(b)** N_{d}^{US}

(c) $N_{\text{geo}}^{\text{EU}}$ **(d)** N_{d}^{EU}

Figure 3.6: Adjacency Matrices by the geographical neighboring regions and distance based definition (3.4.1).

Table 3.1: Distance limit for $N_{\text{d},\alpha}$ (km)

Neighborhood	US	Europe
$N_{\text{d},5}$	939.64	390.68
$N_{\text{d},7.5}$	1168.92	474.67
$N_{\text{d},10}$	1303.49	559.41
$N_{\text{d},15}$	1648.05	712.39

Table 3.2: p-values of tests for N_{geo} and $N_{\text{d},\alpha}$ (Overall dependency)

	The United States of America			Europe		
	NBD	Isotropic	Distance	NBD	Isotropic	Distance
N_{geo}	0.0002	N/A	N/A	0.0046	N/A	N/A
$N_{\text{d},5}$	0.0002	N/A	N/A	0.0032	N/A	N/A
$N_{\text{d},7.5}$	0.0000	N/A	N/A	0.0302	0.0008	0.0032
$N_{\text{d},10}$	0.0688	0.0002	0.0000	0.1086	0.0004	0.0106
$N_{\text{d},15}$	0.0642	0.0002	0.0002	0.8354	0.0000	0.0958

Table 3.3: p-values of tests for N_{geo} (Substructure)

Regions	Neighborhood	Isotropic
West	0.0050	N/A
Mid-West	0.0532	0.0024
South	0.0148	0.0320
Northeast	0.0010	N/A
BEL-NLD	0.0002	N/A
France	0.0020	N/A
Germany	0.0016	N/A
Spain	0.0698	0.0858

even though they are far from each other since the distance limit of $N^{\text{US}}_{\text{d},10}$ is 1303.49km, which is close to the distance between Nebraska and Tennessee. Moreover, we rejected both isotropic and distance tests. On the other hand, we found European countries had narrower neighborhood than US. The distance limit of $N^{\text{EU}}_{\text{d},7.5}$ is only 474.67km, which is close to the distance between Schleswig-Holstein and Sachsen in Germany. The distance test also failed to reject for $N^{\text{EU}}_{\text{d},10}$ and $N^{\text{EU}}_{\text{d},15}$.

Furthermore, we observed that some sub-region had different neighborhood structure to overall US and Europe. Table 3.3 displays p-values of the tests for sub-domains of both US and Europe. US mid-west, US south, and Spain might have the geographical neighborhood structure. These regions have lower population density, so their network structures have different features to other regions. Moreover, US south and Spain failed to reject the isotropic test. Lower population densities in these regions may be a reason of these dependency structure, and further scientific investigations require.

There are possible explanations about the results above. First, high population density in European countries may be a reason of these results. As we observed above, lower population density might be a reason of N_{geo}. Second, difference on the mode of transportation in US and European countries may be a reason. It is well-known that long-distance trip and air travel are correlated with flu epidemic events [124]. US has higher percentage of air travel and lower percentage of railway travel than European countries. To verify these hypothesis, further epidemiological studies are required.

We compared our findings to [119]. They studied the susceptible parameter in SIRS model for both interstate transmission from neighboring states and intrastate transmission. They performed the model selection procedure based on the deviance information criterion, then they used the model omitting the air travel component of transmission although air travel may play a significant role in the spread of ILI. We found US south had N_{geo} as its neighborhood. From Figure 4 and 5 of [119], US south may be susceptible to interstate transmission, which is consistent with our finding. They also found that many of the mid-west states are less susceptible to interstate transmission than the states in the northwest, which is different to ours.

In summary, our inference procedure can give a new perspective to analyze network. Our analysis could be good for scientific studies as an analysis with epidemiological covariates even though the data is time-evolving. Further scientific studies require to verify whether our findings have epidemiological meaning.

3.5 Simulation studies

Using the numerical experiments, this section is devoted to investigate the finite sample performance of our proposed method under different settings.

3.5.1 Settings

We considered the following settings for our simulation studies. We set the regular grids with varying sizes among 10×10, 30×20, and 50×40, so that the number of vertices $p = 100, 600$, and

2000, respectively. For the number of time points, we let $n = 100, 300, and 500$. We report the empirical sizes and powers of our method for different scenarios based on 500 numerical replications, and 1000 bootstrap were drawn for each replication. The significance level is set as $\alpha = 0.05$. Let $\boldsymbol{\epsilon}_1, \cdots, \boldsymbol{\epsilon}_n$ be *i.i.d.* p-dimensional samples from $N(\mathbf{0}, \boldsymbol{\Sigma})$. For a p-dimensional time-evolving data $\{\mathbf{y}_t\}_{t=1}^n$, we considered the following three different data generation mechanisms.

- (D1) **IID:** Let $\mathbf{y}_t = \boldsymbol{\epsilon}_t \sim N(\mathbf{0}, \boldsymbol{\Sigma})$.
- (D2) **AR(1):** Let the components of $\{\mathbf{y}_t\}$ be the part of p independent AR(1) process. More precisely, $\mathbf{y}_t = (y_{1,t}, \cdots y_{p,t})^T$, $y_{i,t} = \rho y_{i,t-1} + \sqrt{1-\rho^2}\epsilon_{i,t}$ for $1 \leq i \leq p$, $\boldsymbol{\epsilon}_t = (\epsilon_{1,t}, \cdots, \epsilon_{p,t})^T$, and $\boldsymbol{\epsilon}_t \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ where $\rho = 0.3$.
- (D3) **Latent ARCH(1):** Let the components of $\{\mathbf{y}_t\}$ be the part of p dimensional latent ARCH(1) process. More precisely, $\mathbf{y}_t = (y_{1,t}, \cdots y_{p,t})^T$, $\mathbf{y}_t|\mathbf{z}_t = \mathbf{z}_t + \rho\boldsymbol{\epsilon}_t$, $\mathbf{z}_t|\mathbf{y}_{t-1} \sim N(\mathbf{0}, \boldsymbol{\Gamma}_0 + \boldsymbol{\Gamma}_1 \circ \mathbf{y}_{t-1}\mathbf{y}_{t-1}^T)$, $\boldsymbol{\epsilon}_t = (\epsilon_{1,t}, \cdots, \epsilon_{p,t})^T$, and $\boldsymbol{\epsilon}_t \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Gamma}_0 = \boldsymbol{\Gamma}_1 = 128^{-1}\mathbf{I}_p$, $\rho = 1 - 128^{-1}$.

Motivated from the disucssion on learning the Gaussian graphical model using the scaled Lasso in [52], we set the tuning parameter $\lambda = \sqrt{(2c\log p)/n}$ in (3.3.2) for each j where $0 < c < 0.5$. It has been known that the recovered graph is not sensitive to the choice of c in practice and we let $c = 0.25$ in simulations.

3.5.2 Empirical Size

For each of the dependence structures above, we examined the sizes of the proposed tests. We generated data from the following GMRFs as the null structures of tests in Section 3.2.

- (T1) Neighborhood: Let $\boldsymbol{\Omega}$ be a GMRF with four nearest neighborhood where $\eta = 0.15$ in (3.2.6). That is, $\mathcal{E}_{N_4}$ is the support of $\boldsymbol{\Omega}$, and $\eta_{j,l}$ are constant for all the pair in N_4.
- (T2) Directional (Anisotropic): Let $\boldsymbol{\Omega}$ be the precision matrix of directional dependence structure with four nearest neighborhood where $\eta_u = 0.15$ and $\eta_v = 0.1$ in (3.2.7).

  - (T2a) Isotropic: Let $\mathbf{\Omega}$ be the precision matrix of directional dependence structure with four nearest neighborhood where $\eta = 0.15$ in (3.2.6).

- (T3) Distance: Let $\mathbf{\Omega}$ be the precision matrix of the distance dependence structure where $\eta = -0.2$ in (3.2.9).

- (T4) Approximated Matérn: Let $\mathbf{\Omega} = (\omega_{k,l})_{1 \le k,l \le p}$ be the dependence structure from [99] with the range parameter, $r = 2$. More precisely, $\omega_{l,l} = 40$, $\omega_{l,l\pm1} = \omega_{l\pm1,l} = -12$, $\omega_{l\pm1,l\pm1} = 2$, and $\omega_{l\pm2,l} = \omega_{l,l\pm2} = 1$.

The empirical sizes of the proposed tests are controlled with respect to the nominal level in all of the scenarios as displayed in Table 3.4. With our choice of the tuning parameter λ from the procedure in Section 3.5.1, we observe that most of the empirical size in Table 3.4 were between 2.5% and 7.5%. The size are consistently controlled for all settings in (D1)–(D3), which provides numerical evidences that the proposed method performs well with temporally dependent data.

3.5.3 Empirical Power

We next report the empirical powers of our procedure against different alternatives, including the graph whose corresponding precision matrix has exponentially decaying off-diagonals and therefore admits a bandable structure, a sparse GMRFs with small perturbations, and three nested GMRF structures as follows.

- (A1) **Exponential decay structure:** $\mathbf{\Omega}_a = (\omega^*_{k,l})_{1 \le k,l \le p}$ is the exponential decay precision matrix: $\omega^*_{k,l} = C \cdot 0.4^{|k-l|^{1/3}}$ where $C = 1$ for (T1)-(T3) and $C = 40$ for (T4).

- (A2) **Sparse GMRF with perturbation:** $\mathbf{\Omega}_a = (\omega_{k,l})_{1 \le k,l \le p} = \mathbf{\Omega}_0 + \delta \mathbf{\Omega}^*$ where $\mathbf{\Omega}_0$ is the null structure, and $\delta = 0.075$ for (T1)-(T3), and $\delta = 2$ for (T4). $\mathbf{\Omega}^* = (\omega^*_{k,l})_{1 \le k,l \le p}$ is a sparse GMRF as follows: Let $\tau = \sqrt{4 \log p}$, and $\mathbf{\Omega}^{(1)} = (\omega^{(1)}_{k,l})_{1 \le k,l \le p}$. Diagonal elements $\omega^1_{k,k} = 3\tau/2$. We randomly picked twenty non-diagonal elements, B_{20}. For $(k,l) \in B_{20}$, $\omega^{(1)}_{k,l} = \omega^{(1)}_{l,k} \sim \text{Unif}(\tau/2, 3\tau/2)$. Let $\lambda_{\min}$ be the smallest eigenvalue of $\mathbf{\Omega}^{(1)}$. $\mathbf{\Omega}^* = (\omega^*_{k,l})_{1 \le k,l,\le p}$, $\omega^*_{k,k} = \omega^1_{k,k} - \min(0, \lambda_{\min})$, and $\omega^*_{k,l} = \omega^{(1)}_{k,l}$ for $k \neq l$.

Table 3.4: The empirical size $(\%)$ of the tests (T1)-(T4) for assessing different GMRFs at the 5% nominal level. T1, T2, T2a, T3, and T4 are neighborhood, anisotropic directional, isotropic, distance, approximated Matérn structure defined in Section 3.2. D1, D2, and D3 are IID, AR(1), and latent ARCH(1).

Model	n	Grid Size	T1	T2	T2a	T3	T4
D1	200	10×10	6.6	4.8	7.0	5.2	4.0
		30×20	4.4	5.0	5.4	5.0	3.6
		50×40	2.2	5.6	2.4	3.6	5.2
	300	10×10	5.0	4.8	5.2	5.6	3.4
		30×20	6.4	7.8	7.8	5.0	6.8
		50×40	3.2	3.2	5.6	3.6	4.2
	500	10×10	5.0	5.0	5.4	3.6	3.6
		30×20	4.4	6.6	3.8	5.0	3.6
		50×40	4.6	6.4	4.4	6.2	7.4
D2	200	10×10	4.8	5.0	3.8	3.8	2.2
		30×20	4.6	3.0	3.0	3.0	3.2
		50×40	3.8	3.8	3.8	3.4	4.2
	300	10×10	4.4	4.0	5.0	5.4	3.4
		30×20	4.4	3.6	5.2	5.4	4.4
		50×40	7.0	5.6	6.0	3.0	7.6
	500	10×10	4.2	5.8	5.2	5.4	5.6
		30×20	4.8	3.8	5.2	2.6	3.2
		50×40	3.6	4.0	3.6	3.8	5.2
D3	200	10×10	4.8	6.2	5.4	6.8	3.6
		30×20	6.4	8.2	6.6	5.6	7.8
		50×40	6.2	3.2	4.8	5.2	3.8
	300	10×10	4.8	6.8	5.6	8.2	6.0
		30×20	6.2	5.0	4.4	4.8	3.8
		50×40	3.0	4.2	4.2	3.6	3.8
	500	10×10	5.0	6.0	5.4	6.4	7.2
		30×20	5.6	5.2	5.4	5.6	3.4
		50×40	6.8	6.2	5.0	6.6	5.0

- (A3) **Four nearest vs Eight nearest neighborhood structures:** We considered the four nearest neighborhood structure N_4 as the null. The data was generated by the eight nearest neighborhood isotropic structure with $\eta = 0.125$.

- (A4) **Eight nearest vs Twelve nearest neighborhood structures:** We considered eight nearest neighborhood as the null. The data was generated by twelve nearest neighborhood isotropic structure with $\eta = -0.125$.

- (A5) **Isotropic vs Anisotropic Directional:** We set isotropic structure with four nearest neighborhood as the null. The data was generated by an anisotropic directional GMRF with $\eta_u = -0.12$ and $\eta_v = 0.12$.

Then, we applied the same tests in empirical size simulation for (A1) and (A2), the neighborhood test for (A3) and (A4), and the isotropic test for (A5).

Notice that the dimensionality of the problem grows exponentially fast. For example, to examine whether $\omega_{i,j} = 0$ for $\forall (i,j) \notin \mathcal{S}_{N_4}$, $|\mathcal{S}_{N_4}|$ is 9540, 357100, and 3990180 for $p = 100$, 600, and 2000, respectively. However, as observed in Tables 3.5, 3.6, and 3.7, the empirical powers quickly approach to 1 as the number of time points increase. Our procedure is therefore consistent against sparse signals, which reflects the expected advantage of the L_∞-type testing statistic.

The empirical powers are affected by the precision matrix structure. We observed this phenomenon that one test was easier to detect specific structure than other tests. For instance, (A1) structure had higher power for the distance test because of entries on the edges of regular grids; $\omega_{1,2} = \omega_{10,11} = 0.4$ in 10×10 A1, but $\omega_{1,2} \neq \omega_{10,11}$ in a distance structured 10×10 GMRF. However, the test has no significant power for (A2) structure than other tests.

We conclude the section with a discussion on for extra simulation studies for comparison of two tests that one is nested to the other such as a four-nearest neighborhood test versus an eight-nearest neighrobhod test on a twelve-nearest neighborhood dataset. From Table 3.7, we observed (A4) has lower powers than (A3). It is worth noting that (A3) uses more entries $\widehat{\omega}_{j_1.j_2}$ than (A4) to construct the test statistic (3.3.6). However, since we used different structures of dataset in the

Table 3.5: The empirical power $(\%)$ of the tests (T1)-(T4) for data generated by (A1), exponential precision matrix, at the 5% nominal level. T1, T2, T2a, T3, and T4 are neighborhood, anisotropic directional, isotropic, distance, approximated Matérn structure defined in Section 3.2. D1, D2, and D3 are IID, AR(1), and latent ARCH(1).

Model	n	Grid Size	T1	T2	T2a	T3	T4
D1	200	10×10	94.2	95.2	97.0	47.0	100.0
		30×20	38.0	36.6	64.8	34.4	100.0
		50×40	13.6	21.6	28.2	13.2	100.0
	300	10×10	100.0	100.0	100.0	70.8	100.0
		30×20	99.8	100.0	100.0	70.0	100.0
		50×40	89.8	83.4	98.4	66.4	92.4
	500	10×10	100.0	100.0	100.0	100.0	100.0
		30×20	100.0	100.0	100.0	100.0	100.0
		50×40	100.0	100.0	100.0	100.0	100.0
D2	200	10×10	91.8	97.8	95.6	55.4	100.0
		30×20	85.2	61.2	64.2	38.2	43.6
		50×40	54.2	56.0	75.0	43.8	100.0
	300	10×10	100.0	100.0	100.0	84.4	100.0
		30×20	99.8	100.0	100.0	95.8	100.0
		50×40	100.0	99.8	100.0	75.6	82.2
	500	10×10	100.0	100.0	100.0	100.0	100.0
		30×20	100.0	100.0	100.0	100.0	100.0
		50×40	100.0	100.0	100.0	100.0	100.0
D3	200	10×10	94.0	88.2	92.4	56.2	100.0
		30×20	44.8	55.8	74.4	29.8	86.8
		50×40	29.0	22.2	33.8	22.6	69.0
	300	10×10	100.0	100.0	100.0	79.0	100.0
		30×20	100.0	99.2	99.2	78.8	100.0
		50×40	95.6	96.2	98.6	66.2	100.0
	500	10×10	100.0	100.0	100.0	100.0	100.0
		30×20	100.0	100.0	100.0	100.0	100.0
		50×40	100.0	100.0	100.0	100.0	100.0

Table 3.6: The empirical power $(\%)$ of the tests (T1)-(T4) for data generated by (A2), sparse GMRF with perturbation, at the 5% nominal level. T1, T2, T2a, T3, and T4 are neighborhood, anisotropic directional, isotropic, distance, approximated Matérn structure defined in Section 3.2. D1, D2, and D3 are IID, AR(1), and latent ARCH(1).

Model	n	Grid Size	T1	T2	T2a	T3	T4
D1	200	10×10	84.8	81.8	83.2	28.8	51.8
		30×20	73.2	67.2	76.8	11.2	34.2
		50×40	52.8	55.4	58.8	4.8	3.6
	300	10×10	99.8	99.6	99.8	55.4	92.4
		30×20	99.4	98.8	99.6	15.2	86.6
		50×40	98.6	98.0	99.2	7.8	71.0
	500	10×10	100.0	100.0	100.0	82.4	100.0
		30×20	100.0	100.0	100.0	40.6	100.0
		50×40	100.0	100.0	100.0	24.8	100.0
D2	200	10×10	83.0	83.2	85.4	32.6	44.2
		30×20	78.6	72.0	73.6	10.4	18.8
		50×40	75.0	67.2	77.4	8.8	3.6
	300	10×10	99.6	99.6	98.4	54.6	88.8
		30×20	99.0	99.8	99.4	19.4	76.6
		50×40	100.0	99.0	99.2	9.0	47.6
	500	10×10	100.0	100.0	100.0	79.6	100.0
		30×20	100.0	100.0	100.0	41.2	100.0
		50×40	100.0	100.0	100.0	25.2	99.8
D3	200	10×10	86.0	83.4	82.0	32.6	49.8
		30×20	73.8	70.2	72.2	9.8	100.0
		50×40	64.8	57.8	65.0	6.2	41.2
	300	10×10	99.6	99.0	99.8	53.8	55.2
		30×20	98.8	99.4	99.8	17.4	100.0
		50×40	98.0	98.0	99.0	7.8	77.0
	500	10×10	100.0	100.0	100.0	79.8	98.8
		30×20	100.0	100.0	100.0	38.8	100.0
		50×40	100.0	100.0	100.0	22.6	99.2

Table 3.7: The empirical power $(\%)$ of the corresponding tests for data generated by (A3)-(A5) at the 5% nominal level. (A3) is comparing four nearest neighborhood with eight nearest neighborhood, whereas (A4) is comparing eight nearest neighborhood with twelve nearest neighborhood. (A5) is comparing anisotropic directional with isotropic. D1, D2, and D3 are IID, AR(1), and latent ARCH(1).

Model	n	Grid Size	A3	A4	A5
D1	200	10×10	68.2	38.4	46.2
		30×20	76.6	18.0	30.2
		50×40	73.0	7.8	13.0
	300	10×10	93.2	68.8	80.4
		30×20	94.6	60.6	78.2
		50×40	96.2	38.8	62.6
	500	10×10	100.0	99.6	100.0
		30×20	100.0	99.8	100.0
		50×40	100.0	99.4	100.0
D2	200	10×10	68.6	32.4	48.2
		30×20	68.6	47.8	28.8
		50×40	72.0	35.8	54.6
	300	10×10	91.8	68.6	83.8
		30×20	95.0	64.2	84.4
		50×40	98.8	82.6	82.4
	500	10×10	100.0	99.8	100.0
		30×20	100.0	99.8	100.0
		50×40	100.0	99.8	100.0
D3	200	10×10	67.2	37.0	40.6
		30×20	66.0	21.0	33.0
		50×40	54.2	16.8	17.8
	300	10×10	92.4	71.2	79.8
		30×20	94.4	62.2	65.4
		50×40	94.2	51.8	61.2
	500	10×10	100.0	99.4	100.0
		30×20	100.0	99.6	100.0
		50×40	100.0	99.6	100.0

settings of (A3) and (A4), the direct comparison of the two results is not meaningful. Nevertheless, for the same dataset, testing twelve-nearest neighborhood would be easier to capture signals than testing eight-nearest neighborhood since the latter is a nested structure of the former. To verify this intuition, it would be interesting to study the power of the four-nearest neighborhood test on twelve-nearest isotropic structured data by comparing the results from (A3) and (A4).

Chapter 4

Learning Gaussian Graphical Model with Uniform Performance via Distributional Robust Optimization

4.1 Introduction

The Gaussian graphical model is a convenient and powerful tool for studying network structure among large number of variables. It has been used in various scientific applications including gene network analysis, image analysis, and functional brain network analysis. Let $G = (V, E)$ be an undirected graph with vertices V and edges E, and each vertex corresponds to the each component of a random vector. Two vertices v_i and v_j are connected by an edge (i, j) if and only if there is a conditional dependence between the two random variables X_i and X_j given all other variables of $\mathbf{X}$. Meanwhile, the inverse of its covariance matrix, *the precision matrix*, is of great interest in statistics. Let $\mathbf{\Omega}$ be the precision matrix of a Gaussian random vector $\mathbf{X} = (X_1, \ldots, X_p) \sim N(\mathbf{0}, \mathbf{\Omega}^{-1})$, then $\mathbf{\Omega}$ provides the graph structure of $\mathbf{X}$. To be specific, for a Gaussian random vector $\mathbf{X}$, each entry of the precision matrix satisfies that $\omega_{ij} \neq 0$ if and only if there is an edge between v_i and v_j. In high-dimensional setting ($n \ll p$), various methods using ℓ_1 penalty have been proposed under sparsity conditions from pioneering works including the nodewise regression method [35], Glasso [37–39], CLIME [40].

However, Gaussian graphical models rely on normality or sub-Gaussianity assumption in theory. If a dataset violates the distributional assumptions, we encounter performance loss of statistical procedures in some sense. For example, micro RNA or RNA-seq data do not follow normal distributions, so gene network estimates from the data might have more false negatives or false positives. A practical remedy is log-transformation, but it does not guarantee that the transformed dataset follows a multivariate normal distribution. Besides, the data itself can be contaminated in data processing. Non-Gaussian graphical approaches also rely on another distributional as-

sumption, and it is not a remedy of the contamination. In practice, we do not know the type of contamination and how much the data is deviated from the distributional assumption.

If a procedure can endure distributional assumption violation than other methods, we can call the procedure is robust. This idea is from [1, 3], which is called *distributional robustness*. It studies the case where the true underlying distribution is slightly different from the distributional assumption. It would be quantified by a discrepancy measure or distance between two distributions. From the concept of robustness, statisticians have tackled various contamination models to consider small distributional perturbation. The most popular choice is Huber's ϵ-contamination model [1]. Statisticians also consider the deviation of estimators to study another type of robustness, called tail-robustness, even the underlying distribution has finite small order moments [8, 9]. Recently, statistical learning and machine learning communities consider various model setups such as adversarial contamination settings [21, 125–127] and covariate shifts [15, 128, 129].

In the past years, several precision matrix estimation methods have been proposed to weaken normality assumption or to achieve robustness under various settings. [59–61] proposed robust precision matrix estimators under cellwise contamination by the combination of robust covariance matrix estimators and standard high dimensional precision matrix estimation procedures. [46, 58] proposed estimators for Gaussian copula using *nonparanormal transformation* so that the transformed random vector follows a multivariate normal distribution. [130] proposed robust precision matrix estimators under the finite $(2 + \epsilon)$th moments assumption for $\epsilon \in (0, 2)$ rather than sub-Gaussianity assumption. [131] extended CLIME to transfer learning by combining information from auxiliary studies—these samples are independently drawn from slightly different distributions of the target study—to improve the performance of estimation and prediction.

Precision matrix estimations with different penalty term have been studied. Lasso provides sparsity of estimates from the ℓ_1 penalty geometry, but it is difficult to capture highly correlated variables. On the other hand, ridge penalty provides a closed form like ridge regressions and reduces variance of estimator, whereas it does not provide sparsity. There are a few studies on the graphical ridge estimator for precision matrix when sparsity is not required [132, 133]. Elastic

net type penalties motivated by elastic net regression [134] have also considered to obtain sparsity of the estimator and capture the highly correlated variables together. To the best of our knowledge, [135] was the first attempt to use the elastic net penalty for exponential family Markov Random Field (MRF) estimation. For Gaussian MRF, they used the original elastic net regression for nodewise regression. Recently, [136, 137] considered methods called graphical elastic net, but their motivations are not from distributionally robustness.

Distributionally robust optimization aims to find the minimizer of the worst-case expected loss with an ambiguity set or uncertainty set of probability distributions or parameters, denoted by $\mathcal{U}$:

$$\widehat{\theta} = \underset{\theta}{\operatorname{argmin}} \sup_{\mathbb{P}\in\mathcal{U}} \mathbb{E}_{\mathbb{P}}\ell(\theta; \mathbf{X}) \tag{4.1.1}$$

for some function $\ell(\cdot)$. (4.1.1) is often computationally intractable, but there exists an equivalent or asymptotic equivalent tractable dual formulation under mild conditions. The dual form is often a well-known problems in statistics with extra regularization term. For example, distributionally robust linear regression with Wasserstein ball is equivalent to penalized linear regressions [17, 20]. Distributionally robust square-root Lasso problem with an ambiguity set defined by ℓ_2 ball to constrain the sketch of data matrix, a low-rank approximation of data matrix, is equivalent to square-root elastic net [138]. In the past decade, there are several papers combining methods in classical statistics or statistical learning with distributionally robust optimization. For regressions, [15] considered distributionally robust linear regression with least square loss under covariate shift scenario. [16] studied distributionally logistic regressions with Wasserstein ball. [17, 20] provided properties of distributionally robust regressions. It is worth noting that the dual formulation of (4.1.1) depends on the ambiguity set $\mathcal{U}$, which is to consider all distributions that are slightly different to a specific distribution. The popular choices of probability metrics for ambiguity sets include Wasserstein distance [16, 17, 20], f-divergence [19, 28, 139], and L_p norm [138]. In this paper, we consider Wasserstein-2 distance defined by Euclidean norm, which is used for both statistical procedures and distributionally robust optimization [32, 140].

We propose a graph estimation method named DRAGON (Distributionally Robust grAph estimation via nodewise reGressiON) that aims to be robust against certain amount of distributional perturbation of the data. To this end, we employ distributionally robust regressions on Gaussian graph estimation using nodewise regression. Nodewise regression idea allows us to detour computation burden of $p \times p$ matrix optimization problem by solving p parallel sparse regression problems. Our distributionally robust regression formulation is equivalent to a square-root regression with a strictly convex penalty, which is related to square-root elastic net.

The rest of the paper proceeds as follows. In Section 4.2, we revisit graph and precision matrix estimation methods by nodewise regression, and introduce our method based on the distributionally robust regression. Especially, we revisit a computationally tractable dual form of distributinally robust regression. Section 4.3 is devoted to simulation studies. We provide the proofs and additional numerical analysis in the supporting information.

4.2 Method

4.2.1 Background

We first define some notation and introduce Gaussian graph and precision matrix estimation via nodewise regression. For a vector $\mathbf{v} = (v_1, \ldots, v_d) \in \mathbb{R}^d$, we define a vector norm $\|\mathbf{v}\|_q = (\sum_{i=1}^d v_i^q)^{1/q}$ for $q \in [1, \infty)$ and $\|\mathbf{v}\|_\infty = \max_{1 \le i \le d} |v_i|$. For a matrix $\mathbf{A}$, we define $\|\mathbf{A}\|_q = \sup_{\|\mathbf{u}\|_q = 1} \|\mathbf{A}\mathbf{u}\|_q$ and $\|\mathbf{A}\|_F = (\sum_{i,j} a_{ij}^2)^{1/2}$. We denote the trace of $\mathbf{A}$ by $\mathrm{Tr}(\mathbf{A})$. We denote a data matrix by $\mathbf{X} \in \mathbb{R}^{n \times p}$ and its column vectors by $\mathbf{X}_j = (X_{1j}, \ldots, X_{nj}) \in \mathbb{R}^n$ for $j = 1, \ldots, p$. We denote a submatrix of $\mathbf{X}$ without the jth column by $\mathbf{X}_{-j}$. We denote the ith row vector of $\mathbf{X}_{-j}$ by $\mathbf{X}_{i,-j} = (X_{i,1}, \ldots, X_{i,j-1}, X_{i,j+1}, \ldots, X_{i,p}) \in \mathbb{R}^{p-1}$. For an index set A, we denote a submatrix of $\mathbf{X}$ without the jth columns for $j \in A$ by $\mathbf{X}_{-A}$. We denote the subvector of $\boldsymbol{\beta}_j$ except $\beta_{j,k}$ by $\boldsymbol{\beta}_{j,-k}$. We use $\mathbf{X} \overset{\mathrm{d}}{=} \mathbf{Y}$ when $\mathbf{X}$ and $\mathbf{Y}$ have the same distribution.

Nodewise regression method proposed by [35] is to estimate the neighborhood set of a vertice v_i in high-dimensional setting, $\{v_k \in \mathbf{V} : \widehat{\beta}_{jk} \neq 0, k \neq j\}$, where $\widehat{\beta}_{jk}$ is estimated by ℓ_1-penalized linear regression

$$
\begin{aligned}
\widehat{\boldsymbol{\beta}}_j &= (\widehat{\beta}_{j,1}, \ldots, \widehat{\beta}_{j,j-1}, \widehat{\beta}_{j,j+1}, \ldots, \widehat{\beta}_{j,p})^{\mathrm{T}} \\
&= \underset{\boldsymbol{\beta} \in \mathbb{R}^{p-1}}{\operatorname{argmin}} \left(\frac{1}{n} \sum_{i=1}^{n} (X_{ij} - \mathbf{X}_{i,-j}^{\mathrm{T}} \boldsymbol{\beta})^2 + \lambda \|\boldsymbol{\beta}\|_1 \right) \\
&= \underset{\boldsymbol{\beta} \in \mathbb{R}^{p-1}, \|\boldsymbol{\beta}\|_1 \leq c}{\operatorname{argmin}} \left(\frac{1}{n} \sum_{i=1}^{n} (X_{ij} - \mathbf{X}_{i,-j}^{\mathrm{T}} \boldsymbol{\beta})^2 \right),
\end{aligned}
\tag{4.2.1}
$$

for some $c \geq 0$ and $\lambda \geq 0$ where $\mathbf{X}_i \in \mathbb{R}^p$ is a random sample from a normally distributed random vector $\mathbf{X} \sim N(0, \boldsymbol{\Omega}^{-1})$. An intuitive explanation on the nodewise regression is that the linear regression estimator provides information on conditional dependence between predictor variables and the response variable, and the graphical models aim to capture the conditional dependence among edges. Thus, if $\widehat{\beta}_{jk} \neq 0$, there is a conditional dependence between X_j and X_k given $\mathbf{X}_{-\{j,k\}}$.

We can apply another regression method for estimating Gaussian graphs or precision matrices. [50] and [102] considered square-root Lasso [141] in the nodewise regression approach for Gaussian graph estimation:

$$
\widehat{\boldsymbol{\beta}}_j = \underset{\boldsymbol{\beta} \in \mathbb{R}^{p-1}}{\operatorname{argmin}} \left\{ \left(\frac{1}{n} \sum_{i=1}^{n} (X_{ij} - \mathbf{X}_{i,-j}^{\mathrm{T}} \boldsymbol{\beta})^2 \right)^{1/2} + \lambda \|\boldsymbol{\beta}\|_1 \right\}, \tag{4.2.2}
$$

for some $\lambda \geq 0$. One advantage of the nodewise regression with square-root Lasso is its asymptotically tuning-free property, which inherits the property of square-root Lasso. This reduces computation burdens from cross-validation.

To estimate the precision matrix from p nodewise regression results, the block matrix inversion formula is used [142]. Let $\widehat{\sigma}_j^2 = n^{-1} \sum_{i=1}^{n} (X_{ij} - \mathbf{X}_{i,-j} \widehat{\boldsymbol{\beta}}_j)^2$. Then, the preliminary estimate $\widetilde{\boldsymbol{\Omega}} = \{\widetilde{\omega}_{jk}\}_{1 \leq j,k \leq p}$ is defined by

$$
\widetilde{\omega}_{jj} = \widehat{\sigma}_j^{-2}, \ \widetilde{\boldsymbol{\Omega}}_{-j,j} = -\widehat{\sigma}_j^{-2} \widehat{\boldsymbol{\beta}}_j. \tag{4.2.3}
$$

for $j = 1, \ldots, p$. Following that, we symmetrize the preliminary estimate to obtain the final estimate $\widehat{\mathbf{\Omega}}$. In this paper, we use the symmetrized matrix $\widehat{\mathbf{\Omega}} = \{\widehat{\omega}_{jk}\}_{1 \leq j,k, \leq p}$ where $\widehat{\omega}_{jk} = \widetilde{\omega}_{jk}\mathbb{I}(|\widetilde{\omega}_{jk}| \leq |\widetilde{\omega}_{kj}|) + \widetilde{\omega}_{kj}\mathbb{I}(|\widetilde{\omega}_{jk}| > |\widetilde{\omega}_{kj}|)$ proposed by [40].

4.2.2 Wasserstein distance and distributional perturbation

Wasserstein distance is originated from optimal transport theory starting from late 18th century [25]. Wasserstein distance have received much attention in optimization, statistics, and deep learning communities. One of its famous application is Wasserstein generative adversarial network [143]. We refer to [140] for a review of Wasserstein distance in statistics.

In this paper, we consider Wasserstein-2 distance with Euclidean norm $d(x, y) = \|x - y\|_2$.

Definition 4.2.1 (Wasserstein-2 Distance with Euclidean norm)**.** For the Euclidean norm $\|\cdot\|_2$ on $\mathbb{R}^p$, Wasserstein-2 distance is defined by

$$W_2(\mathbb{P}, \mathbb{Q}) := \left[\inf_{\pi \in \Pi(\mathbb{P}, \mathbb{Q})} \mathbb{E}_{(X,Y)\sim\pi}\|\mathbf{X} - \mathbf{Y}\|_2^2\right]^{1/2} \tag{4.2.4}$$

where $\Pi(\mathbb{P}, \mathbb{Q})$ denote the set of all joint distributions $\pi(X, Y)$ whose marginal distributions are $\mathbb{P}$ and $\mathbb{Q}$, respectively.

Throughout this paper, we call (4.2.4) Wasserstein-2 distance for convenience. Quantifying the distributional perturbation by Wasserstein-2 distance is the next question. To this end, we consider examples that provide bounds of Wasserstein distance between the true distribution and the perturbed distributions to quantify the distributional perturbation under various scenarios. We first introduce the closed form of Wasserstein-2 distance with Euclidean norm between two normal distributions.

Proposition 4.2.1 (Proposition 7 in [144])**.** *The Wasserstein-2 distance with Euclidean norm in Definition 4.2.1 between two multivariate normal distributions* $\mathbb{P}_1 \sim N(\boldsymbol{\mu}_1, \mathbf{\Sigma}_1)$ *and* $\mathbb{P}_2 \sim N(\boldsymbol{\mu}_2, \mathbf{\Sigma}_2)$ *satisfies*

$$W_2^2(\mathbb{P}_1, \mathbb{P}_2) = \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_2^2 + \mathrm{Tr}\left\{\mathbf{\Sigma}_1 + \mathbf{\Sigma}_2 - 2\left(\mathbf{\Sigma}_2^{1/2}\mathbf{\Sigma}_1\mathbf{\Sigma}_2^{1/2}\right)^{1/2}\right\}, \tag{4.2.5}$$

where $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}$ are non-singular matrices.

It is worth noting that Wasserstein-2 distance between two elliptical distributions with the same density generator has the same closed form [145]. Also, this closed form serves as a lower bound of the Wasserstein-2 distance [146], which is often called Gelbrich bound.

Proposition 4.2.2 ([146])**.** *Consider Wasserstein-2 distance with Euclidean norm in Definition 4.2.1. Suppose $\mathbb{P}_1$ and $\mathbb{P}_2$ are distributions with mean vectors $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathbb{R}^p$ and invertible covariance matrices $\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2$, respectively. Then, it holds that*

$$W_2^2(\mathbb{P}_1, \mathbb{P}_2) \geq \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_2^2 + \mathrm{Tr}\left\{\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2 - 2\left(\boldsymbol{\Sigma}_2^{1/2}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2^{1/2}\right)^{1/2}\right\}. \tag{4.2.6}$$

The equality holds when $\mathbb{P}_1$ and $\mathbb{P}_2$ are elliptical distributions the same density generator.

Throughout this paper, we consider $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{0}$ for simplicity, so we can use simpler form of (4.2.5) and (4.2.6). It is natural to think that Wasserstein distance depends on the contamination level α under a specific contamination model. With Propositions 4.2.1 and 4.2.2, we check the bounds for Wasserstein distance from a distribution under selected settings. With these building blocks, we check the bound for the rowwise contamination setting, which is a generalized version of Huber's ϵ-contamination model [1].

Example 4.2.1 (Rowwise contamination with a normal contaminant)**.** Consider $\mathbf{X} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_X)$ and $\mathbf{Y} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_Y)$ in $\mathbb{R}^p$ where $\mathbf{X}$ and $\mathbf{Y}$ are independent. Let $\mathbf{W} = (1-\epsilon)\mathbf{X} + \epsilon\mathbf{Y}$ where $\epsilon \sim Bernoulli(\alpha)$, and ϵ is independent of $\mathbf{X}$ and $\mathbf{Y}$.

By Gelbrich bound, it holds that

$$\begin{aligned} W_2^2(\mathbf{X}, \mathbf{W}) &\geq \mathrm{Tr}\left\{\boldsymbol{\Sigma}_X + (1-\alpha)\boldsymbol{\Sigma}_X + \alpha\boldsymbol{\Sigma}_Y\right\} \\ &\quad - 2\mathrm{Tr}\left[\left[\boldsymbol{\Sigma}_X^{1/2}\{(1-\alpha)\boldsymbol{\Sigma}_X + \alpha\boldsymbol{\Sigma}_Y\}\boldsymbol{\Sigma}_X^{1/2}\right]^{1/2}\right] \\ &= \mathrm{Tr}\left\{(2-\alpha)\boldsymbol{\Sigma}_X + \alpha\boldsymbol{\Sigma}_Y\right\} \\ &\quad - 2\mathrm{Tr}\left[\left\{(1-\alpha)\boldsymbol{\Sigma}_X^2 + \alpha\boldsymbol{\Sigma}_X^{1/2}\boldsymbol{\Sigma}_Y\boldsymbol{\Sigma}_X^{1/2}\right\}^{1/2}\right]. \end{aligned} \tag{4.2.7}$$

By the definition of Wasserstein-2 distance, it holds that

$$W_2^2(\mathbf{X}, \mathbf{W}) := \inf_{\Pi(\mathbf{X},\mathbf{W})} \mathbb{E}\|\mathbf{X} - \mathbf{W}\|_2^2 \leq \mathbb{E}\|\mathbf{X} - \mathbf{W}\|_2^2 = \alpha \mathrm{Tr}(\mathbf{\Sigma}_X + \mathbf{\Sigma}_Y). \tag{4.2.8}$$

where $\Pi(\mathbf{X}, \mathbf{W})$ denotes the set of joint probability distributions whose marginal distributions are $\mathbf{X}$ and $\mathbf{W}$.

Combining (4.2.7) with (4.2.8), we have

$$\begin{aligned} &\mathrm{Tr}\left[(2-\alpha)\mathbf{\Sigma}_X + \alpha\mathbf{\Sigma}_Y - 2\left[(1-\alpha)\mathbf{\Sigma}_X^2 + \alpha\mathbf{\Sigma}_X^{1/2}\mathbf{\Sigma}_Y\mathbf{\Sigma}_X^{1/2}\right]^{1/2}\right] \\ &\leq W_2^2(\mathbf{X}, \mathbf{W}) \leq \alpha\mathrm{Tr}(\mathbf{\Sigma}_X + \mathbf{\Sigma}_Y), \end{aligned} \tag{4.2.9}$$

which provides a rough bound.

In Example 4.2.1, we considered a normal contaminant with zero mean. If the contaminant does not follow multivariate normal distribution, the bound changes since the result above relies on $\|\mathbf{X}\|_2^2 \overset{\mathrm{d}}{=} \sum_{j=1}^p \lambda_{Xj} Z_j^2$ where λ_{Xj} is the jth largest eigenvalue of $\mathbf{\Sigma}_X$ for $j = 1, \ldots, p$.

Using the same argument, we can obtain a bound for the cellwise contamination model [6].

Example 4.2.2 (Cellwise contamination with a normal contaminant)**.** Consider $\mathbf{X} \sim N(\mathbf{0}, \mathbf{\Sigma}_X)$ and $\mathbf{Y} \sim N(\mathbf{0}, \mathbf{\Sigma}_Y)$ in $\mathbb{R}^p$ where $\mathbf{X}$ and $\mathbf{Y}$ are independent. Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are *i.i.d.* sample from $\mathbf{X}$, and $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ are *i.i.d.* sample from $\mathbf{Y}$.

Let $\mathbf{W}_i = (\mathbf{I} - \mathbf{D}_i)\mathbf{X}_i + \mathbf{D}_i\mathbf{Y}_i$ where $\mathbf{D}_i = \mathrm{diag}(d_{i1}, \ldots, d_{in})$ is a diagonal matrix, and $d_{ik} \sim Bernoulli(\alpha_k)$ for $k = 1, \ldots, n$, and $\{d_{ik}\}_{k=1}^n$ are independent of $\mathbf{X}$ and $\mathbf{Y}$. Let $\mathbf{A} = \mathrm{diag}(\alpha_1, \ldots, \alpha_n)$ be a diagonal matrix. We compute Wassderstein-2 distance between $\mathbf{X}$ and $\mathbf{W}_i$.

Using the same argument as Example 4.2.1 with careful algebra, we have

$$\begin{aligned} &\mathrm{Tr}\left[(2\mathbf{I} - \mathbf{A})\mathbf{\Sigma}_X + \mathbf{A}\mathbf{\Sigma}_Y - 2\left[\mathbf{\Sigma}_X^{1/2}\{(\mathbf{I} - \mathbf{A})\mathbf{\Sigma}_X + \mathbf{A}\mathbf{\Sigma}_Y\}\mathbf{\Sigma}_X^{1/2}\right]^{1/2}\right] \\ &\leq W_2^2(\mathbf{X}, \mathbf{W}) \leq \mathrm{Tr}\{\mathbf{A}(\mathbf{\Sigma}_X + \mathbf{\Sigma}_Y)\}. \end{aligned} \tag{4.2.10}$$

It is worth noting that both the rowwise contamination and the cellwise contamination models are special case of the contamination model considered in [6] when $d_{i1} = \cdots = d_{in}$ and ,d_{ik} are fully independent, respectively. We can also derive (4.2.9) from (4.2.10) when $\alpha_1 = \cdots = \alpha_n = \alpha$.

From Examples 4.2.1 and 4.2.2, we observe the contamination levels affects the bound for Wasserstein-2 distance. For the rowwise and cellwise contamination models, the upper bound of Wasserstein-2 distance increases as the contamination level increases. The lower bound also depends on the contamination level.

4.2.3 Wasserstein distributionally robust regression

Before introducing our method, we discuss the regression formulation motivated from Wasserstein distributionally robust regression. For the data matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$, we consider a regression where jth column $\mathbf{X}_j$ as a response variable and the sub-matrix except jth column, $\mathbf{X}_{-j}$, as predictor variables. The goal of distributionally robust regression is to obtain an estimate that minimizes worst-case loss among distributions in an uncertainty set.

For a Wasserstein-2 ball centered at the empirical distribution $\widehat{\mathbb{P}}$ with radius ρ, it follows from the proof of Proposition 2 in [20] when $q = 2$ and $\delta = \rho^2$.

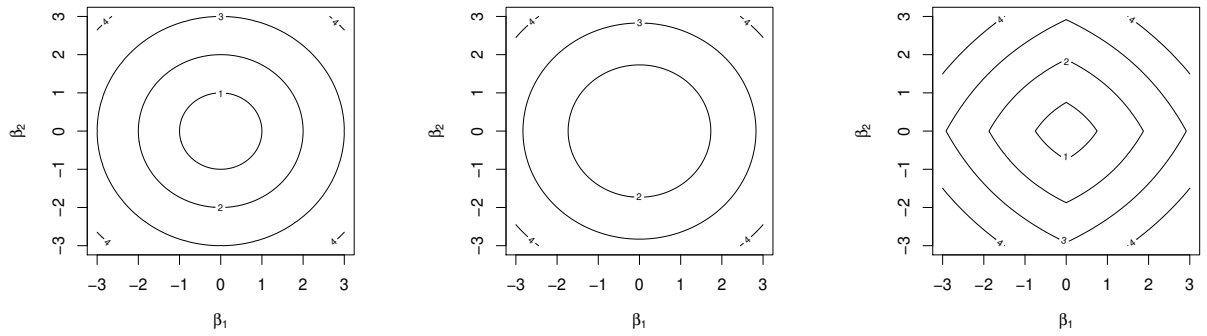
Proposition 4.2.3 (Corollary of Proposition 2 in [20])**.** *Consider the least square loss, $\ell(\mathbf{x}; \boldsymbol{\beta}_j) = (\mathbf{x}_j - \mathbf{x}_{-j}^{\mathrm{T}}\boldsymbol{\beta})^2$, and $c(\mathbf{x}, \mathbf{x}') = \|\mathbf{x} - \mathbf{x}'\|_2^2$. Then, it holds that*

$$\sup_{\mathbf{X}\sim\mathbb{Q},\mathbb{Q}\in B_\rho(\widehat{\mathbb{P}})} \mathbb{E}_{\mathbb{Q}} \left(X_j - \mathbf{X}_{-j}\boldsymbol{\beta}_j\right)^2 = \left(n^{-1/2}\|\mathbf{X}_j - \mathbf{X}_{-j}\boldsymbol{\beta}_j\|_2 + \rho\|(1, -\boldsymbol{\beta}_j)\|_2\right)^2. \tag{4.2.11}$$

That is, we can interpret the worst-case risk among the distributions in $B_\rho(\widehat{\mathbb{P}})$ by the empirical risk plus extra term depending on ρ.

Using (4.2.11), we can derive a computationally tractable dual form of distributionally robust linear regression. We consider the distributionally robust formulation of Lasso regression [36]:

$$\widehat{\boldsymbol{\beta}}_j = \underset{\boldsymbol{\beta}\in\mathbb{R}^{p-1}}{\operatorname{argmin}} \left\{ \lambda\|\boldsymbol{\beta}_j\|_1 + \sup_{\mathbf{X}\sim\mathbb{Q},\mathbb{Q}\in B_\rho(\widehat{\mathbb{P}})} \mathbb{E}_{\mathbb{Q}} \left(X_j - \mathbf{X}_{-j}\boldsymbol{\beta}_j\right)^2 \right\}. \tag{4.2.12}$$



(a) $\|\boldsymbol{\beta}\|_2$ **(b)** $\|(1, -\boldsymbol{\beta})\|_2$ **(c)** $0.5\|\boldsymbol{\beta}\|_1 + 0.5\|(1, -\boldsymbol{\beta})\|_2$

Figure 4.1: Geometries of penalty terms in $\mathbb{R}^2$.

Plugging (4.2.11) into (4.2.12), we obtain

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_j &= \underset{\boldsymbol{\beta}_j \in \mathbb{R}^{p-1}}{\operatorname{argmin}} \left\{ \left(n^{-1/2} \|\mathbf{X}_j - \mathbf{X}_{-j}\boldsymbol{\beta}_j\|_2 + \rho \|(1, -\boldsymbol{\beta}_j)\|_2 \right)^2 + \lambda \|\boldsymbol{\beta}_j\|_1 \right\} \\ &= \underset{\boldsymbol{\beta}_j \in \mathbb{R}^{p-1}}{\operatorname{argmin}} \{ n^{-1} \|\mathbf{X}_j - \mathbf{X}_{-j}\boldsymbol{\beta}_j\|_2^2 + \rho^2 \|(1, -\boldsymbol{\beta}_j)\|_2^2 + \lambda \|\boldsymbol{\beta}_j\|_1 \\ &\qquad + 2n^{-1/2} \rho \|\mathbf{X}_j - \mathbf{X}_{-j}\boldsymbol{\beta}_j\|_2 \|(1, -\boldsymbol{\beta}_j)\|_2 \}. \end{aligned} \tag{4.2.13}$$

It is worth noting that (4.2.13) is the same as LASSO (4.2.1) when $\rho \to 0$.

Although (4.2.12) is a distributionally robust version of standard Lasso, its dual form (4.2.13) is not straightforward to interpret due to $\|\mathbf{X}_j - \mathbf{X}_{-k}\boldsymbol{\beta}_j\|_2 \|(1, -\boldsymbol{\beta}_j)\|_2$ term. Without this term, it looks similar to the elastic net [134], but they are not exactly the same. For this reason, we consider another regression formulation for given ρ, which is related to (4.2.13) and easier to interpret.

$$\widehat{\boldsymbol{\beta}}_j = \underset{\boldsymbol{\beta}_j \in \mathbb{R}^{p-1}}{\operatorname{argmin}} \left\{ n^{-1/2} \|\mathbf{X}_j - \mathbf{X}_{-j}\boldsymbol{\beta}_j\|_2 + \lambda \|\boldsymbol{\beta}_j\|_1 + \rho \|(1, -\boldsymbol{\beta}_j)\|_2 \right\}, \tag{4.2.14}$$

which is similar to "square-root" elastic net since $\|(1, -\boldsymbol{\beta}_j)\|_2$ has similar geometry to $\|\boldsymbol{\beta}\|_2^2$ around $\mathbf{0}$ and $\|\boldsymbol{\beta}\|_2$ otherwise. Figure 4.1 displays geometry of the penalty term in (4.2.14). The term $\lambda\|\boldsymbol{\beta}\|_1 + \rho\|(1, -\boldsymbol{\beta}_j)\|_2$ looks similar to the elastic net penalty, so it would provide similar advantages of the elastic net. We present the relationships between (4.2.13) and (4.2.14).

Proposition 4.2.4. *Given $\rho \geq 0$, the two optimization problems* (4.2.13) *and* (4.2.14) *are related to each other. Specifically, let λ and λ' be the tuning parameters of* (4.2.13) *and* (4.2.14)*. Denote a common solution of the two optimization problems by $\widehat{\boldsymbol{\beta}}_j$, then it holds that $\lambda' = \{n^{-1/2}\|\mathbf{X}_j - \mathbf{X}_{-j}\widehat{\boldsymbol{\beta}}_j\|_2 + \rho\|(1, -\widehat{\boldsymbol{\beta}}_j)\|_2\}^{-1}\lambda$.*

Interestingly, we observe that (4.2.14) is equivalent to the following distributionally robust regression form by (4.2.11):

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^{p-1}}{\operatorname{argmin}} \left[\lambda\|\boldsymbol{\beta}\|_1 + \sup_{\mathbf{X}\sim\mathbb{Q}, \mathbb{Q}\in B_\rho(\widehat{\mathbb{P}})} \left\{\mathbb{E}_{\mathbb{Q}}\left(X_j - \mathbf{X}_{-j}\boldsymbol{\beta}\right)^2\right\}^{1/2} \right], \tag{4.2.15}$$

which is the distributionally robust version of square-root Lasso [141]:

$$\widehat{\boldsymbol{\beta}} := \underset{\boldsymbol{\beta}\in\mathbb{R}^{p-1}}{\operatorname{argmin}} \left[\left\{ \frac{1}{n}\sum_{i=1}^{n}(X_{ij} - \mathbf{X}_{i,-j}^{\mathrm{T}}\boldsymbol{\beta})^2 \right\}^{1/2} + \lambda\|\boldsymbol{\beta}\|_1 \right]. \tag{4.2.16}$$

The difference between (4.2.14) and square-root elastic net formulation [138, 147] is that we use $\|(1, -\boldsymbol{\beta})\|_2 = (1 + \|\boldsymbol{\beta}\|_2^2)^{1/2}$ rather than ℓ_2 norm $\|\boldsymbol{\beta}\|_2$. This allows us to utilize the radius of Wasserstein-2 ball ρ directly in the formulation, so it is straightforward to understand the role of ρ. In addition to this, [138] aims to speed-up square-root Lasso using the robust sketching of the matrix of predictor variables where the sketch is constrained by ℓ_2 norm of the difference between the original data and its sketch. On the other hand, we consider a Wasserstein-2 ball for the ambiguity set. [147] used the penalty term as a convex combination of $\|\boldsymbol{\beta}\|_1$ and $\|\boldsymbol{\beta}\|_2$, $\lambda\{(1-\alpha)\|\boldsymbol{\beta}\|_2 + \alpha\|\boldsymbol{\beta}\|_1\}$ for $\alpha \in [0, 1]$, and we use ℓ_1 norm as a penalty term to obtain sparsity of estimate and have $\rho\|(1, -\widehat{\boldsymbol{\beta}}_j)\|_2$ term as a part of worst-case risk.

Remark 4.2.1. The ℓ_2 penalty term $\|(1, -\boldsymbol{\beta}_j)\|_2$ can be reformulated by $\|\boldsymbol{\beta}\|_2^2$ since

$$\|(1, -\boldsymbol{\beta}_j)\|_2 = \|(1, -\boldsymbol{\beta}_j)\|_2^{-1}(1 + \|\boldsymbol{\beta}\|_2^2).$$

Therefore, (4.2.14) is related to square-root elastic net like problems:

$$\widehat{\boldsymbol{\beta}}_j = \underset{\boldsymbol{\beta}\in\mathbb{R}^{p-1}}{\operatorname{argmin}} \left\{ n^{-1/2} \|\mathbf{X}_j - \mathbf{X}_{-j}\boldsymbol{\beta}_j\|_2 + \lambda\|\boldsymbol{\beta}\|_1 + \rho'\|\boldsymbol{\beta}\|_2 \right\}, \tag{4.2.17}$$

$$\widehat{\boldsymbol{\beta}}_j = \underset{\boldsymbol{\beta}\in\mathbb{R}^{p-1}}{\operatorname{argmin}} \left\{ n^{-1/2} \|\mathbf{X}_j - \mathbf{X}_{-j}\boldsymbol{\beta}_j\|_2 + \lambda\|\boldsymbol{\beta}\|_1 + \frac{\rho''}{2}\|\boldsymbol{\beta}\|_2^2 \right\}, \tag{4.2.18}$$

for some $\rho' \geq 0$ and $\rho'' \geq 0$.

From Propositions 4.2.4, these optimization problems are related. The rest of the paper considers (4.2.14) to emphasize the radius of the ambiguity set in the optimization problem. It is also easier to understand.

To obtain intuition, we consider toy examples to study the rule of λ and ρ. First, we provide an example for bivariate normal vector to illustrate the properties of ρ in (4.2.14) with $\lambda = 0$.

Example 4.2.3. Considering a bivariate random vector $(X_1, X_2) \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ where $\mathrm{var}(X_1) = 1$, $\mathrm{var}(X_2) = 1$, and $\mathrm{cov}(X_1, X_2) = r$. For $\beta \in \mathbb{R}$, $\beta_* = r$ is the minimizer of $\mathbb{E}\{(X_1 - X_2\beta)^2\} = (1 - r^2) + (\beta - r)^2$. Consider the ridge regression with the ridge penalty term $\rho\|\beta\|_2^2 = \rho\beta^2$. $\beta_{ridge} = r/(1+\rho)$ is the minimizer of $\mathbb{E}\{(X_1 - X_2\beta)^2\} + \rho\beta^2$. This shows the shrinkage property of ridge regression. Similarly, the optimization problem

$$\underset{\beta}{\operatorname{argmin}} [\mathbb{E}\{(X_1 - X_2\beta)^2\}]^{1/2} + \rho(1 + \beta^2)^{1/2} \tag{4.2.19}$$

has no simple closed form as ridge regression, but we observe the shrinkage property due to $\rho\|(1, -\boldsymbol{\beta})\|_2$ in (4.2.14). Let $\beta_{*,\rho}$ be the minimizer of (4.2.19). For large ρ or $r \neq 0$ with small magnitude, the solution $\beta_{*,\rho}$ is close to $r/(1+\rho)$.

We investigate the role of ρ in this example. The next two Lemmas reveal that ρ provides shrinkage of the regression estimator, which is similar to the ridge regression. We first study a population version of optimization problem.

Lemma 4.2.3.1. *Under the setting in Example 4.2.3, the solution of* (4.2.19)*, denoted by* $\beta_{*,\rho} = \beta_*(\rho, r)$*, satisfies (i)* $\mathrm{sgn}(\beta_{*,\rho}) = \mathrm{sgn}(r)$*; (ii)* $0 \leq |\beta_{*,\rho}| \leq |r|$*,* $\beta_{*,\rho} = r$ *only if* $\rho = 0$*, and* $\beta_{*,\rho} = 0$

only if $r = 0$*; (iii)* $\beta_{*,1} = r^{-1}\{1 - (1 - r^2)^{1/2}\}$ *when* $r \neq 0$*; (iv)* $\operatorname{sgn}(\partial\beta_{*,\rho}/\partial\rho) = -\operatorname{sgn}(\beta_{*,\rho})$*, that is,* ρ *provides shrinkage on* $\beta_{*,\rho}$*; (v) As* $\rho \to \infty$*,* $\beta_{*,\rho} \to 0$*.*

Following that, we consider the empirical version of (4.2.19). We observe a similar behavior from ρ.

Lemma 4.2.3.2. *Assuming we draw* n *random samples from* $N(\mathbf{0}, \mathbf{\Sigma})$ *in the setting of Example 4.2.3, the solution of the optimization problem,*

$$\underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (X_{1i} - X_{2i}\beta)^2 \right\}^{1/2} + \rho(1 + \beta^2)^{1/2}. \tag{4.2.20}$$

denoted by $\widehat{\beta}_\rho$*, satisfies (i)* $\widehat{\beta}_0 = \widehat{\beta}_{OLS} := (n^{-1}\sum_{i=1}^{n} X_{2i}^2)^{-1}(n^{-1}\sum_{i=1}^{n} X_{1i}X_{2i})$*; (ii)* $0 \leq |\widehat{\beta}_\rho| \leq |\widehat{\beta}_0|$*; (iii)* $\operatorname{sgn}(\partial\widehat{\beta}_\rho/\partial\rho) = -\operatorname{sgn}(\widehat{\beta}_\rho)$*; (iv) As* $\rho \to \infty$*,* $\widehat{\beta}_\rho \to 0$*.*

From Example 4.2.3, we observe that ρ provides shrinkage on nodewise regression in the bivariate example. The next result provides a generalized result for higher dimensional.

Proposition 4.2.5. *Consider the optimization problem* (4.2.14)*. For given* λ*, the solution of this problem satisfies the following properties:*

i) If $\widehat{\beta}_{j,k} \neq 0$*, then* $\operatorname{sgn}(\partial\widehat{\beta}_{j,k}/\partial\rho) = -\operatorname{sgn}(\widehat{\beta}_{j,k})$*,*

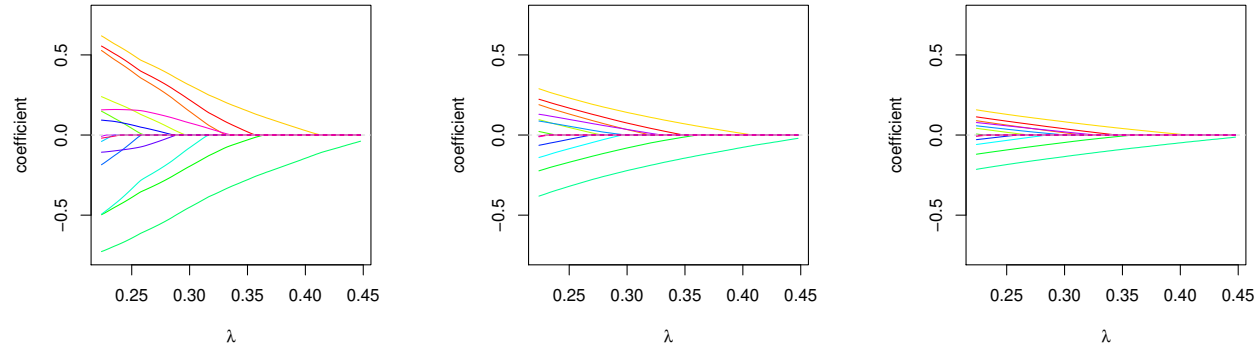
ii) As $\rho \to \infty$*,* $\widehat{\boldsymbol{\beta}}_j \to 0$*.*

In the following example, we investigate the behavior of $\widehat{\boldsymbol{\beta}}$ by different λ and ρ value from independently generated design matrix $\mathbf{X}$.

Example 4.2.4. We generate $X_{ij} \sim N(0, 1)$ for $i = 1, \ldots, 100$, $j = 1, \ldots, 150$ and $\epsilon_i \sim N(0, 1)$. We obtain $Y_i = \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta} + \epsilon_i$ where

$$\boldsymbol{\beta} = (1, 1, 1, 0.5, 0.5, -1, -1, -1, -0.5, -0.5, 0, 0, \ldots, 0) \in \mathbb{R}^{150}.$$

Figure 4.2 displays the solution paths from (4.2.14) when we replace $\mathbf{X}_j$ and $\mathbf{X}_{-j}$ by $\mathbf{Y}$ and $\mathbf{X}$, respectively. The solution paths of (4.2.14) show knowledge on ℓ_1 and ℓ_2 penalty parameters on



(a) $\rho = 0.1$ **(b)** $\rho = 0.5$ **(c)** $\rho = 1$

Figure 4.2: Solution paths of the optimization problem (4.2.14) by λ and ρ.

the elastic net are applicable. As λ increases, the estimate is getting sparse. On the other hand, as ρ increases, we observe shrinkage of estimates.

Another characteristic of elastic net is to capture highly correlated variables simultaneously. We observe the that (4.2.14) can handle grouping effect like elastic net.

Example 4.2.5. With the sample size $n = 100$, we consider two settings for $Z_1, Z_2, \epsilon, \xi_1, \xi_2, \ldots, \xi_6$:

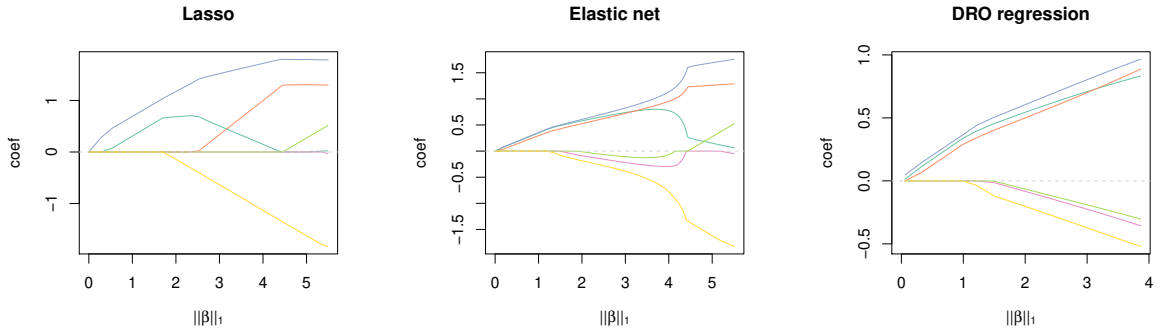
i) All of them are generated from $N(0, 1)$ [148];

ii) $\epsilon \sim t_3$, and the rest of them are generated from $N(0, 1)$;

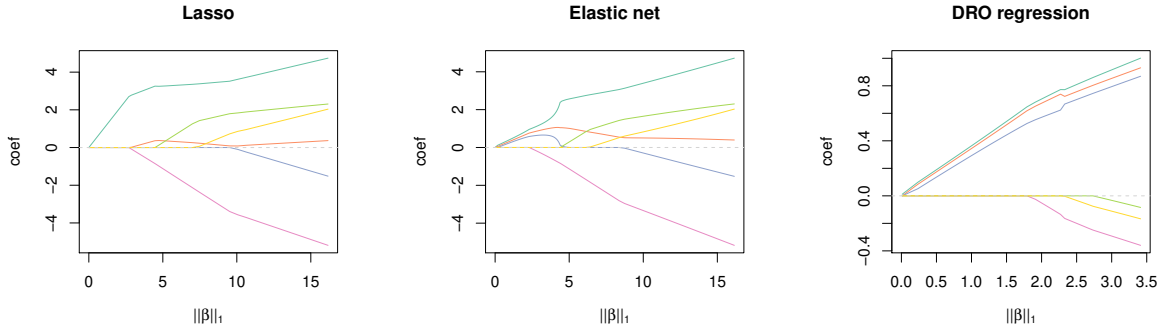
To generate highly correlated variables, we obtain $X_j = Z_1 + \xi_j/5$ for $j = 1, 2, 3$, $X_j = Z_2 + \xi_j/5$ for $j = 4, 5, 6$, and $Y = 3Z_1 - 1.5Z_2 + 2\epsilon$. In addition to this two settings, we consider one more setting:

iii) All of them are generated from $N(0, 1)$ as i), and obtain $X_1, \ldots, X_6, Y$ as the above. Then, we add random noises generated from $N(0, 0.2)$ to $X_{1i}, \ldots, X_{6i}, Y_i$ for $i = 1, \ldots, 20$.

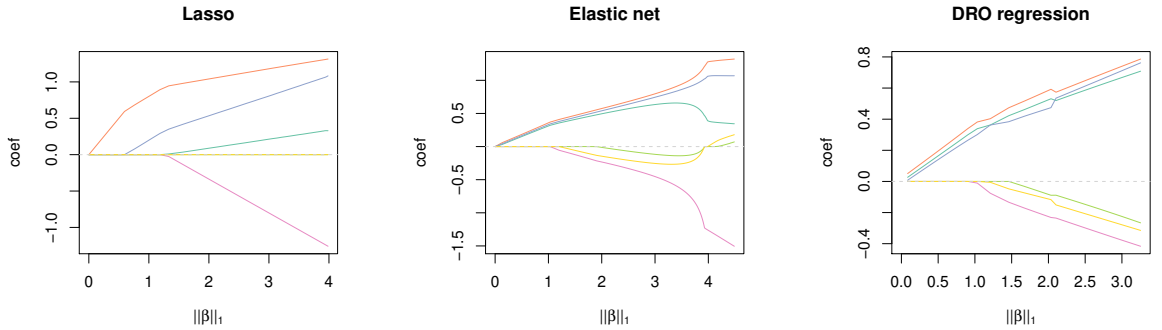
We fit the linear model Y against X_j for $j = 1, \ldots, 6$ with no intercept. Figure 4.3 displays the solution paths of Lasso, elastic net with $\alpha = 0.3$, and (4.2.14) with $\rho = 0.5$. We observe that the estimate from (4.2.14) captures the group structure of highly correlated predictor variables.



(a) All from $N(0,1)$



(b) $\epsilon \sim t_3$, others are from $N(0,1)$.



(c) Contaminated $(\mathbf{X}, Y)$

Figure 4.3: Solution paths of Lasso (left), Elastic net with $\alpha = 0.3$ (middle), (4.2.14) with $\rho = 0.5$ (right) for highly correlated predictor variables. We generate (a) X_k, ξ_k for $k = 1, \ldots, 6$, Z_1, Z_2, and ϵ from $N(0,1)$ as (4.1) from [148], (b) generate $\epsilon \sim t_3$ and others are from $N(0,1)$, (c) generate contaminated $(\mathbf{X}, Y)$ by adding $N(0, 0.2)$ noise on 20% of (X, Y) from (a).

4.2.4 DRAGON for precision matrix estimation

For Gaussian sparse graphs in high-dimensional setting, we assume $\mathbb{P} \sim N(\mathbf{0}, \mathbf{\Omega}^{-1})$ whose precision matrix is sparse. We want to construct a robust graph estimator against distributional perturbation in Wasserstein-2 ball around $\widehat{\mathbb{P}}$ with a few restrictions. To acquire the sparsity in high-dimensional Gaussian graphs theory and make the optimization problem be convex, we use an ℓ_1-regularization term $\|\boldsymbol{\beta}_j\|_1$. To this end, we consider (4.2.14) for each column, which is the Wasserstein distributionally robust regression version of (4.2.16):

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_j &= \underset{\boldsymbol{\beta}\in\mathbb{R}^{p-1}}{\operatorname{argmin}} \left[\lambda\|\boldsymbol{\beta}\|_1 + \sup_{\mathbf{X}\sim\mathbb{Q},\mathbb{Q}\in B_\rho(\widehat{\mathbb{P}})} \left\{\mathbb{E}_{\mathbb{Q}}\left(X_j - \mathbf{X}_{-j}\boldsymbol{\beta}\right)^2\right\}^{1/2}\right] \\ &= \underset{\boldsymbol{\beta}\in\mathbb{R}^{p-1}}{\operatorname{argmin}} \left\{n^{-1/2}\|\mathbf{X}_j - \mathbf{X}_{-j}\boldsymbol{\beta}\|_2 + \lambda\|\boldsymbol{\beta}\|_1 + \rho\|(1, -\boldsymbol{\beta})\|_2\right\},\end{aligned} \tag{4.2.21}$$

for $j = 1, \ldots, p$ where $B_\rho(\widehat{\mathbb{P}}) = \{\mathbb{Q} : W_2(\mathbb{Q}, \widehat{\mathbb{P}}) \le \rho\}$, a Wasserstein-2 ball with radius ρ centered at the empirical distribution $\widehat{\mathbb{P}}$.

To solve (4.2.14) numerically, we introduce iterating equations for DRAGON. Similar to the relationship between the scaled Lasso [53] and the square-root Lasso [141], (4.2.14) has the following equivalent form: For $j = 1, \ldots, p$,

$$\widehat{\boldsymbol{\beta}}_j = \underset{\boldsymbol{\beta}\in\mathbb{R}^{p-1}}{\operatorname{argmin}} \left\{\frac{\|\mathbf{X}_j - \mathbf{X}_{-j}\boldsymbol{\beta}\|_2^2}{2n\sigma_j} + \frac{\sigma_j}{2} + \lambda\|\boldsymbol{\beta}\|_1 + \rho\|(1, -\boldsymbol{\beta})\|_2\right\}, \tag{4.2.22}$$

where $\sigma_j^2 = n^{-1}\|\mathbf{X}_j - \mathbf{X}_{-j}\boldsymbol{\beta}\|_2^2$. With a careful algebra, we have the following algorithm based on iterating equations: We set the initial values $\boldsymbol{\beta}_j^{(0)} = \mathbf{0}$ and $(\sigma_j^2)^{(0)} = n^{-1}\|\mathbf{X}_j\|_2^2$ in the implementation, but we could use an estimate from other methods or algorithms as an initial value for warm start. For a selected $\lambda > 0$ and given $\rho \ge 0$,

$$\begin{aligned}\beta_{j,k}^{(t)} &= \frac{1}{d_{jk}^{(t-1/2)}} S_\lambda\left(\frac{1}{n\sigma_j^{(t-1)}}\langle\mathbf{X}_k, \mathbf{r}_{jk}^{(t-1/2)}\rangle\right),\ k \ne j \\ \sigma_j^{(t)} &= n^{-1/2}\|\mathbf{X}_j - \mathbf{X}_{-j}\boldsymbol{\beta}_j^{(t)}\|_2\end{aligned} \tag{4.2.23}$$

where

$$d_{jk}^{(t-1/2)} := \frac{\|\mathbf{X}_k\|_2^2}{n\sigma_j^{(t-1)}} + \frac{\rho}{\|(1, -\boldsymbol{\beta}_j^{(t-1)})\|_2},$$
$$\mathbf{r}_{jk}^{(t-1/2)} := \mathbf{X}_j - \sum_{l=1, l\neq j}^{k-1} \beta_{j,l}^{(t)} \mathbf{X}_l - \sum_{l=k+1, l\neq j}^{p} \beta_{j,l}^{(t-1)} \mathbf{X}_l,$$

and $S_\lambda(\cdot)$ is the soft-thresholding operator, and $\mathbf{r}_{jk}^{(t-1/2)}$ is a partial residual vector. It seems that $d_j^{(t-1/2)}$ term provides shrinkage when $\rho > 0$. We provide derivation details in Appendix B.1.

From (4.2.23), we deduce properties of $\widehat{\boldsymbol{\beta}}_j$ regarding λ and ρ. The solution of jth regression $\widehat{\boldsymbol{\beta}}_j$ satisfies

$$\widehat{\beta}_{j,k} = \left(\frac{\|\mathbf{X}_k\|_2^2}{n} + \frac{\rho \widehat{\sigma}_j}{\|(1, -\widehat{\boldsymbol{\beta}}_j)\|_2} \right)^{-1} S_{\lambda \widehat{\sigma}_j} \left(\frac{1}{n} \langle \mathbf{X}_k, \mathbf{r}_{jk} \rangle \right), \ k \neq j.$$

where $\widehat{\sigma}_j = n^{-1/2} \|\mathbf{X}_j - \mathbf{X}_{-j} \widehat{\boldsymbol{\beta}}_j\|_2$ and $\mathbf{r}_{jk} = \mathbf{X}_j - \sum_{l \in \{j,k\}}^{p} \widehat{\beta}_{j,l} \mathbf{X}_l$ is a partial residual. First, λ provides selection of non-zero coefficients in $\widehat{\boldsymbol{\beta}}_j$ that is apparent from the soft-thresholding operator. Second, ρ provides shrinkage of $\widehat{\boldsymbol{\beta}}$ as ρ increases.

In our implementation, we use `Rcpp` [149] and `RcppArmadillo` [150] for speeding-up and ease of implementation since the algorithm requires multiple levels of loop and matrix-vector operations. We also store constant values to reduce computation time further. For example, we store $\mathbf{G} = \mathbf{X}^{\mathrm{T}}\mathbf{X} = \{g_{j,k}\}_{1 \leq j,k \leq p}$ outside of the loop since $\|\mathbf{X}_j\|_2 = g_{j,j}$ and $\langle \mathbf{X}_k, \mathbf{X}_j - \mathbf{X}_{-j}\boldsymbol{\beta} \rangle = g_{k,j} - g_{k,-j}^{\mathrm{T}} \boldsymbol{\beta}$ are used in every iteration to update $\beta_{j,k}$ in level 3 loop. We check the relative error of $\boldsymbol{\beta}$, $\|\boldsymbol{\beta}_j^{(t)} - \boldsymbol{\beta}_j^{(t-1)}\|_2 / \|\boldsymbol{\beta}_j^{(t-1)}\|_2$, for the convergence criterion. We use 10^{-6} as the default value of the tolerance, and users can change it if needed.

There are a couple of algorithms related to the implementation of DRAGON. [102] provides the relationship between the scaled Lasso and square-root Lasso, which is a key to reduce computation time for square-root like optimization problem. They use standardized data for p regressions. [147] provide an algorithm for the scaled elastic net and square-root elastic net. They also used the cyclic coordinate descent with iterating equations to update each component of $\boldsymbol{\beta}$, but they only

considered the convex combination of $\|\boldsymbol{\beta}\|_1$ and $\|\boldsymbol{\beta}\|_2$ for the penalty term in their square-root elastic net formulation.

4.3 Simulation Studies

We conduct simulation studies to compare the computation speed and the estimation performance of the proposed procedure with competing methods. We compare the proposed methods with the following methods: i) Glasso [37–39]; ii) CLIME [40]; iii) TIGER [102]; iv) nonparanormal graph estimator (npn) proposed by [58]; and v) the robust precision matrix estimator (LT) proposed by [61]. For implementation, DRAGON implementation relies on `Rcpp` and `RcppArmadillo`. We also use the following R packages: `huge` [46] for TIGER and the nonparanormal transformation step of npn, `fastclime` [44] for CLIME, and `glassoFast` [42,43] for Glasso, npn, and LT.

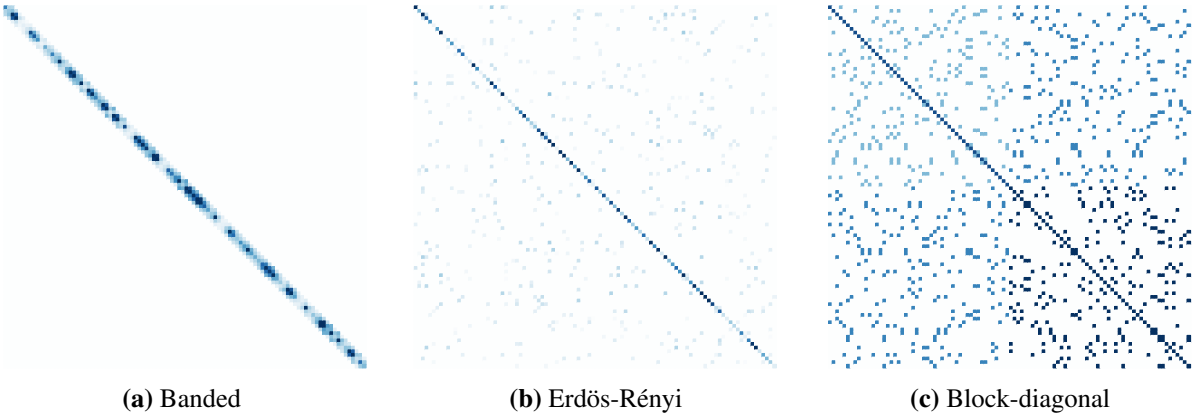
4.3.1 Simulation setting

We consider three precision matrix structures for the first stage of data generation procedure taken from [137]. Construct the precision matrix by $\boldsymbol{\Omega} = \mathbf{D}\widetilde{\boldsymbol{\Omega}}\mathbf{D}$ where $\widetilde{\boldsymbol{\Omega}} = \{\widetilde{\omega}_{j,k}\}_{1\leq j,k\leq p}$ and $\mathbf{D}$ is a diagonal matrix with elements $d_{j,j}$:

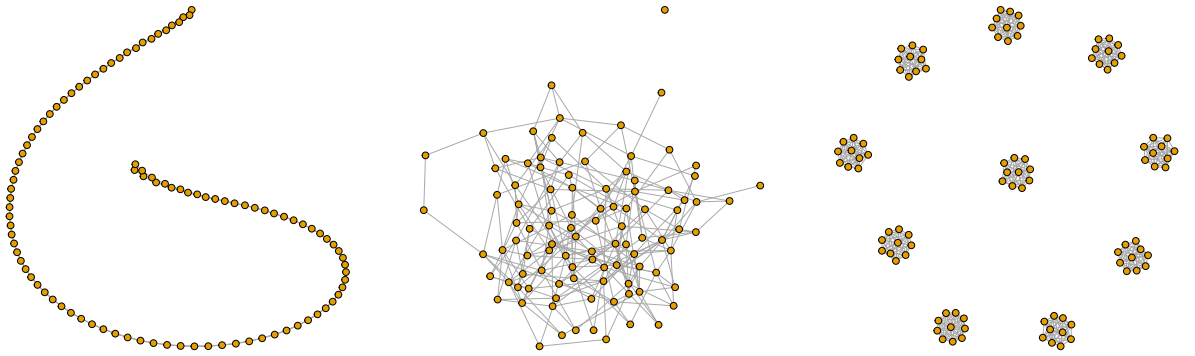
1. *Model 1* (Banded) Set $\widetilde{\omega}_{j,j} = 1$, $\widetilde{\omega}_{j,j+1} = \widetilde{\omega}_{j+1,j} = 0.6$, $\widetilde{\omega}_{j,j+2} = \widetilde{\omega}_{j+2,j} = 0.3$, $\widetilde{\omega}_{j,k} = 0$ for $|j-k| \geq 3$. Generate $d_{j,j} \sim uniform(1,5)$.

2. *Model 2* (Block diagonal) Set a block diagonal matrix with block size $p/10$ such that the diagonal entries are 1 and the off-diagonal entries are 0.5, then we permute the matrix by rows/columns to get $\widetilde{\boldsymbol{\Omega}}$. We use $d_{j,j} = 1$ for $j = 1,\ldots,p/2$ and $d_{j,j} = 1.5$ for $j = p/2+1,\ldots,p$ to obtain the final product $\boldsymbol{\Omega}$.

3. *Model 3* (Erdös-Rényi) Generate $\widetilde{\boldsymbol{\Omega}} = \{\widetilde{\omega}_{jk}\}_{1\leq j,k\leq p}$ $\widetilde{\omega}_{1,jj} = 1$, $\widetilde{\omega}_{1,jk} = \delta_{jk}u_{jk}$ for $j < k$ where $\delta_{jk} \sim Ber(0.05)$ and $u_{jk} \sim uniform(0.4, 0.8)$, and $\widetilde{\omega}_{1,kj} = \widetilde{\omega}_{1,jk}$. Generate $d_{j,j} \sim$



(a) Banded **(b)** Erdös-Rényi **(c)** Block-diagonal



(d) Banded **(e)** Erdös-Rényi **(f)** Block-diagonal

Figure 4.4: For $p = 100$, an illustration of the heatmaps of precision matrices and graph structures from the simulation settings. Darker blue color cells in the heatmap implies that the element of the precision matrix has larger magnitude.

$uniform(1, 5)$. Then set $\mathbf{\Omega} = \mathbf{D}\{\widetilde{\mathbf{\Omega}} + (|\lambda_{\min}(\widetilde{\mathbf{\Omega}})| + 0.05)\mathbf{I}\}\mathbf{D}$ where $\lambda_{\min}(\cdot)$ is the smallest eigenvalue of the matrix.

We consider three distributional perturbation scenarios on the data generation. We consider $(n, p) \in \{(100, 150), (200, 150), (200, 300)\}$. For each pair of (n, p), (i) Row-wise contamination- We Generate $n_1 = \lceil n \times (1 - \alpha) \rceil$ samples from $N(\mathbf{0}, \mathbf{\Omega}^{-1})$ where Ω is the precision matrix specified above and $\alpha \in \{0.1, 0.2, 0.3\}$. Then, we generate $n - n_1$ samples from the multivariate t_3-distribution with the true precision matrix $\mathbf{\Omega}$; (ii) Cell-wise contamination- Generate n samples from $N(\mathbf{0}, \mathbf{\Omega}^{-1})$. We randomly select αnp indices to add cell-wise contaminants drawn from $N(0, 1)$ where $1 \leq i \leq n$, $1 \leq j \leq p$, and $\alpha \in \{0.1, 0.2, 0.3\}$; (iii) Tail deviation; we draw n samples from the multivariate t_3-distribution to have its true precision matrix be $\mathbf{\Omega}$.

Table 4.1: Average timing performance (in milliseconds) with standard errors in parentheses for a single fit with $n = 100$ and $\lambda = 0.75(n^{-1} \log p)^{1/2}$ on the banded structured true precision matrix.

Method	Implementation	$p = 100$	$p = 200$
DRAGON	`RcppArmadillo`	16.61 (2.66)	94.30 (6.48)
CLIME	C++	678.98 (22.63)	2707.05 (149.38)
Glasso	FORTRAN	2.54 (0.38)	9.50 (1.38)
LT	FORTRAN + R	296.34 (8.86)	1170.33 (69.43)
npn	FORTRAN + R	95.14 (8.27)	100.93 (4.52)
TIGER	`RcppEigen`	371.43 (20.18)	390.42 (21.67)

4.3.2 Computation speed

We compare the timing performance of DRAGON for sparse precision matrix estimation with the competing methods. We set $n = 100$ and $p = \{100, 200\}$. We focus on banded graph structure, Model 1 in Section 4.3.1, using a fixed $\lambda = 0.75(n^{-1} \log p)^{1/2}$ and $\rho = 1$ without tuning parameter selection step. All comparisons are made on a computer with MacBook Pro 2019, 1.4 GHz Quad-core Intel Core i5 and 8GB RAM on R version 4.0.5. We use single thread to run the experiment.

Our implementation for DRAGON is built on `Rcpp` and `RcppArmadillo`. `glassoFast` R package used for Glasso, LT, and npn is based on Glasso original algorithm with technical modifications to reduce computation time and resolve non-termination issue in `glasso` R package, which is written in FORTRAN. It means Glasso should be the fastest due to the advantage of FORTRAN and its formulation. `fastclime` R package is written in C for the core functions. TIGER in `huge` R package is implemented on `Rcpp` and `RcppEigen`. Robust covariance estimation for LT is implemented in R.

Table 4.1 displays the summary of timing performance. We observe DRAGON achieves good timing performance for a single fit. Due to the difficult formulation, CLIME is the slowest one among the competing methods. Glasso shows the best timing performance from FORTRAN. Timing performances of LT and npn are due to robust covariance estimation and nonparanormal transformation, respectively.

4.3.3 Numerical performance

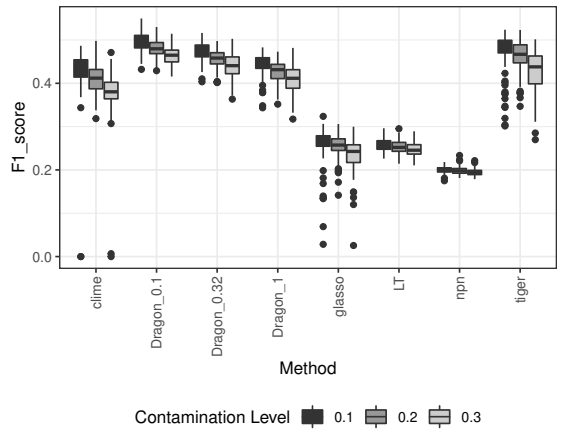
We demonstrate the numerical performance of the proposed method by comparing with existing methods. In this experiment, we use the following tuning parameter selections. We use $\lambda = c(n^{-1}\log p)^{1/2}$ for DRAGON where $c = 0.5$ for Erdös-Rényi graph and $c = 0.75$ for banded and block diagonal structures. We use $\lambda = (n^{-1}\log p)^{1/2}$ for TIGER as their claim. For all other competing methods, we use 5-fold cross-validation to select λ. For DRAGON, we consider the pre-specified radius of the ambiguity set, $\rho = \{10^{-1}, 10^{-0.5}, 1\}$, to check the effect of ρ in both selection and estimation performance. We use DRAGON_0.1, DRAGON_0.32, DRAGON_1 to distinguish the simulation results of DRAGON from different ρ values in summary figures.

We consider measures for selection and estimation performance evaluation. We define the following quantities for the true adjacency matrix $\mathbf{A} = \{a_{jk}\}_{1\le j,k,\le p}$ where $a_{jk} = \mathbb{I}(\omega_{jk} \neq 0)$ and the estimated adjacency matrix $\widehat{\mathbf{A}} = \{\widehat{a}_{jk}\}_{1\le j,k,\le p}$ where $\widehat{a}_{jk} = \mathbb{I}(\widehat{\omega}_{jk} \neq 0)$:

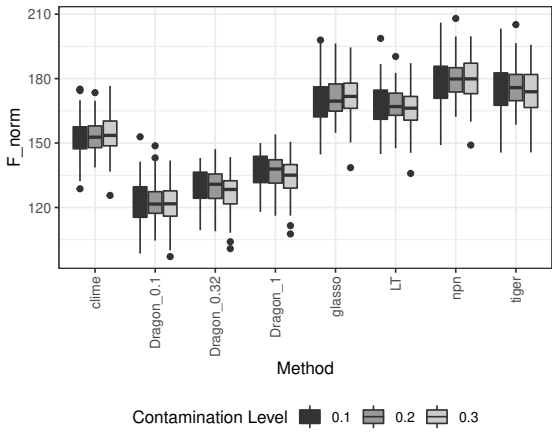
- True positives: $\mathrm{TP} = |\{(j,k) : a_{jk} = 1, \widehat{a}_{jk} = 1, j < k\}|$
- False positives: $\mathrm{FP} = |\{(j,k) : a_{jk} = 0, \widehat{a}_{jk} = 1, j < k\}|$
- True negatives: $\mathrm{TN} = |\{(j,k) : a_{jk} = 0, \widehat{a}_{jk} = 0, j < k\}|$
- False negatives: $\mathrm{FN} = |\{(j,k) : a_{jk} = 1, \widehat{a}_{jk} = 0, j < k\}|$

where $|S|$ is the number of elements of a set S. We consider F1 score, $2\mathrm{TP}/(2\mathrm{TP} + \mathrm{FP} + \mathrm{FN})$, for the graph recovery measure. We also consider Frobenius norm $\|\widehat{\mathbf{\Omega}} - \mathbf{\Omega}\|_F$ to compare the estimation performance.

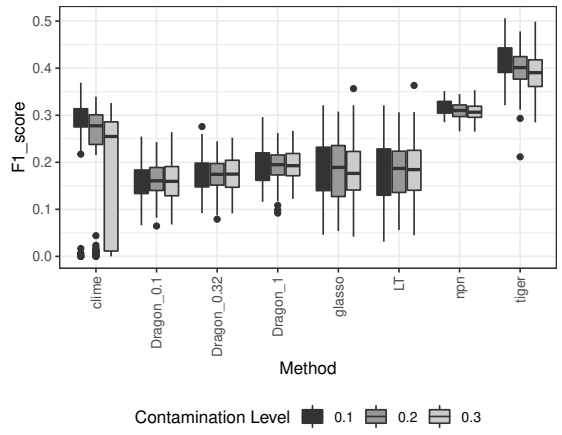
Figures 4.5–4.7 display boxplots of F_1-scores and Frobenius norms from 100 repetitions of experiments when $(n, p) = (100, 150)$. Each of the them corresponds to the contamination setting and consists of six sub-figures. Each row of Figures 4.5–4.7 corresponds to three precision matrix structures. The left and right side of figures present F1 scores and Frobenius norm results by estimation methods, respectively. Higher F1 score implies better selection performance, and smaller Frobenius norm indicates better estimation performance. We provide extra figures for $(n, p) = (200, 150)$ and $(200, 300)$ in Appendix B.5.



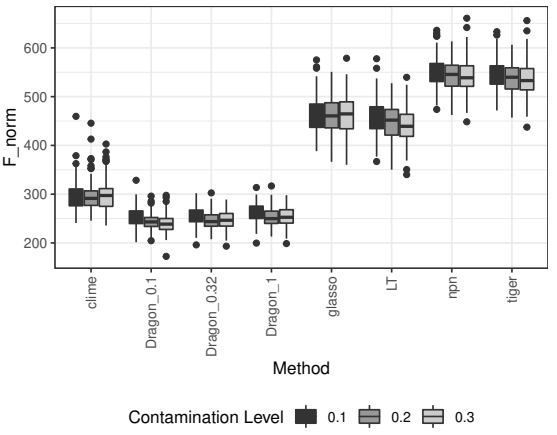
(a) Banded, F1 score



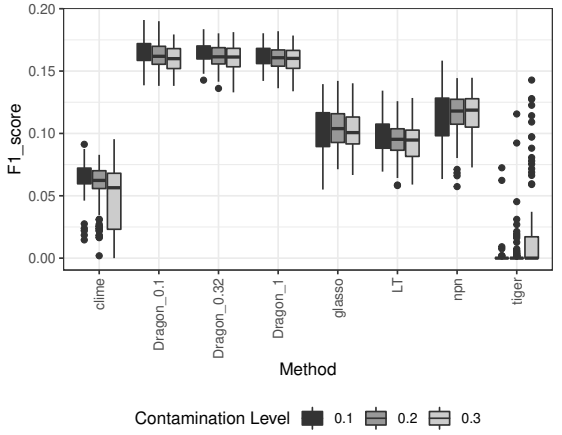
(b) Banded, Frobenius norm



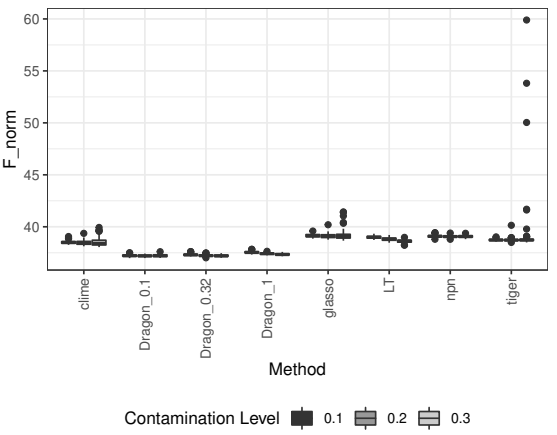
(c) Erdös-Rényi, F1 score



(d) Erdös-Rényi, Frobenius norm

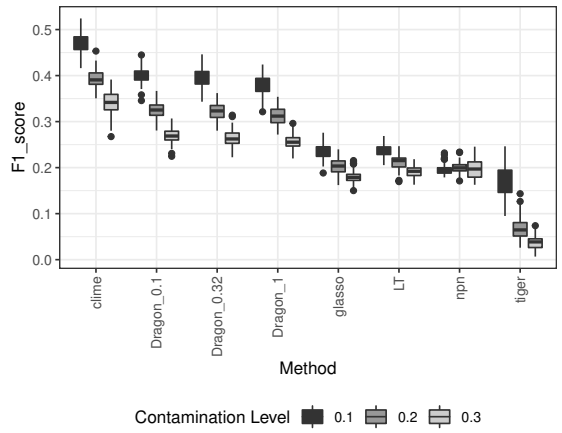


(e) Block diagonal, F1 score

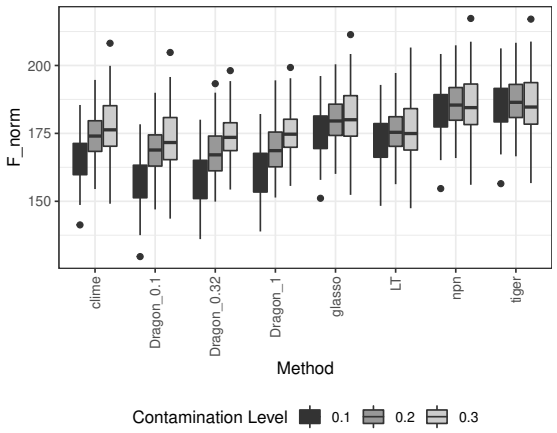


(f) Block diagonal, Frobenius norm

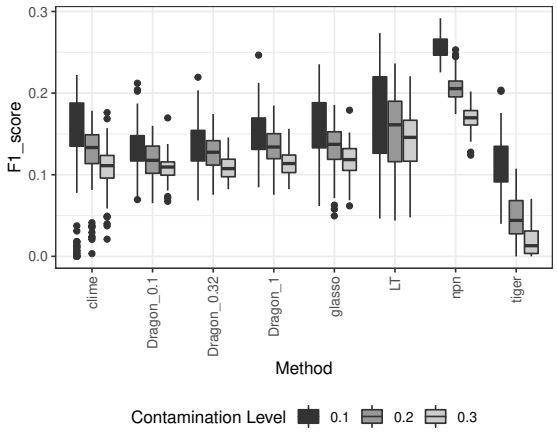
Figure 4.5: F1 score (left) and Frobenius norm (right) under the rowwise contamination setting when $(n, p) = (100, 150)$. Each boxplot summarizes the results from 100 repetitions of experiment.



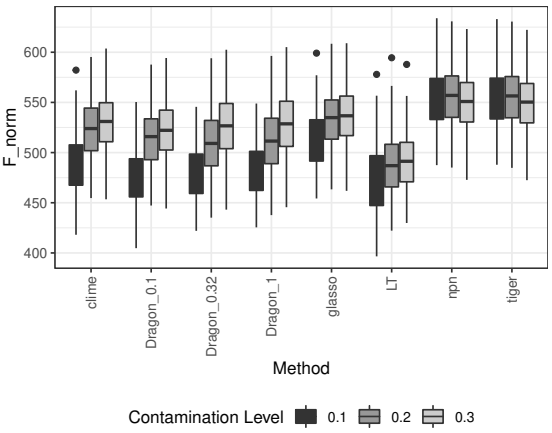
(a) Banded, F1 score



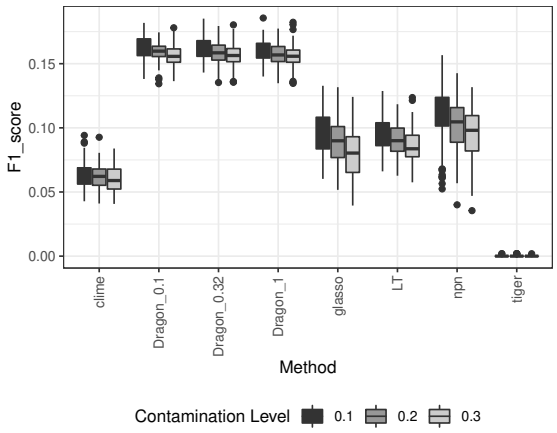
(b) Banded, Frobenius norm



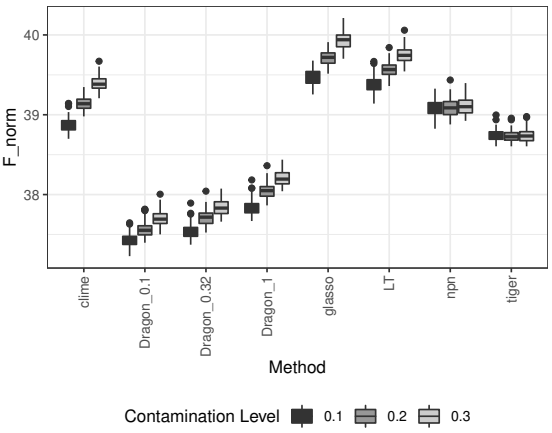
(c) Erdös-Rényi, F1 score



(d) Erdös-Rényi, Frobenius norm

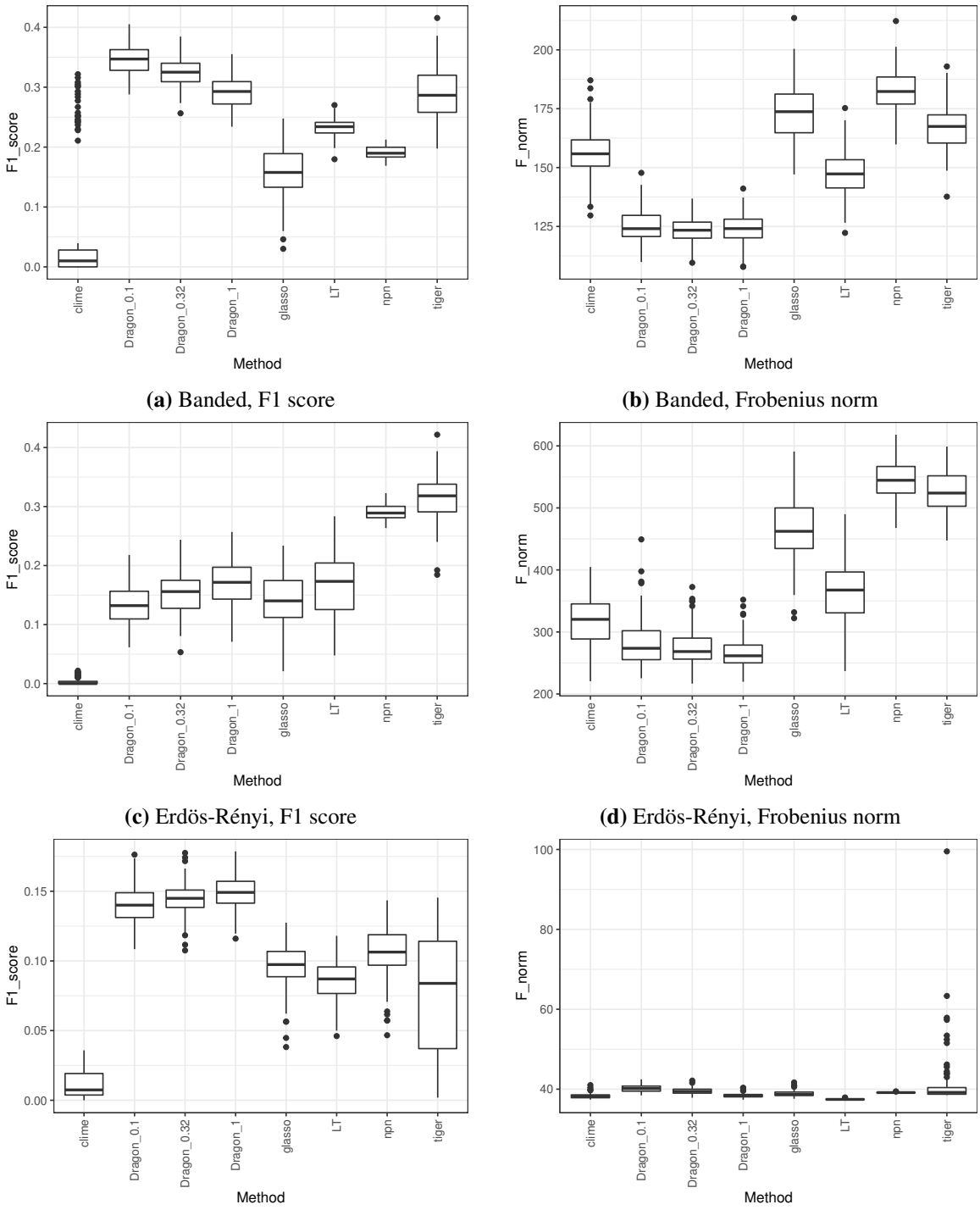


(e) Block diagonal, F1 score



(f) Block diagonal, Frobenius norm

Figure 4.6: F1 score (left) and Frobenius norm (right) under the cellwise contamination setting when $(n, p) = (100, 150)$. Each boxplot summarizes the results from 100 repetitions of experiment.



(a) Banded, F1 score

(b) Banded, Frobenius norm

(c) Erdös-Rényi, F1 score

(d) Erdös-Rényi, Frobenius norm

(e) Block diagonal, F1 score

(f) Block diagonal, Frobenius norm

Figure 4.7: F1 score (left) and Frobenius norm (right) under the tail deviation setting when $(n, p) = (100, 150)$. Each boxplot summarizes the results from 100 repetitions of experiment.

We observe DRAGON has smaller Frobenius norm than competing methods in almost all cases. On the other hand, DRAGON provides good selection performance, but it shows difficulties on Erdös-Rényi graph even we use smaller λ. An intuitive explanation for this phenomenon is that Erdös-Rényi graph does not have clear grouping structure, while the banded and block diagonal structures have correlated chains or blocks of variables. Our formulation (4.2.14) is related to square-root elastic net, and it would be able to handle grouping effect as Example 4.2.5. Hence, DRAGON has the best selection performance on block diagonal structure over the competing methods.

Besides, DRAGON is not always dominated by one of the competing methods over all settings. For example, CLIME shows comparable selection performance with DRAGON under rowwise contamination with the banded structure, but DRAGON shows better selection and estimation performance in the tail deviation setting. TIGER shows higher F1-score than DRAGON for Erdös-Rényi graphs, but DRAGON provides better selection and estimation performance in banded and block diagonal structures. Especially, DRAGON is more stable than TIGER for block diagonal case. LT shows the smallest Frobenius norm under cellwise contamination for Erdös-Rényi graph, but DRAGON shows better or comparable estimation performance than LT for all other settings. We observe similar behavior for $(n, p) = (200, 150)$ and $(200, 300)$; see Figures B.1–B.6 in the supplementary material.

We discuss the effect of ρ on the performance of DRAGON. We observe that the performance is getting worse as a trade-off from larger ρ in the most of the settings. There are exceptions in this observation that we observe better selection and estimation performance as ρ increases in tail deviation settings for Erdös-Rényi and block diagonal structure. We do not have good explanation for this at this moment. It would be of interest to check the distance between the multivariate normal distribution and the multivariate t distribution with our precision matrix settings.

Chapter 5

Discussion and Future Works

We conclude the dissertation with discussions on potential extensions of the proposed methods and future works. First, it is common that the text analysis will benefit from authorships' connections, such as citations, and the publication's information such as the resource of the text and the publication date. These naturally suggest to combine the multivariate regression framework in Section 2 with network information from authors and publications, and therefore leads to a framework on drawing robust inference of text-topic associations. This extension, however, will face a few challenges including the more sophisticated false discovery controlling procedure due to the existence of network structures, the large size of design matrix ($d > n$, potentially) that may require regularization-adaptive Huber estimates, and the extra sparsity in data due to the large number of topics. We will defer these to the future works.

We will investigate the theoretical properties of our model specification tests for GMRFs based on temporally dependent data proposed in Section 3. We will show the validity and consistency of the proposed tests. Also, motivated from the functional false discovery rate and the associated test on continuum, we will generalize the proposed global test to multiple testing problems, which will detect the local region on random field that violates the model assumptions. In addition, we plan to develop an R package for the proposed method.

Several additional studies will focus on the open questions about DRAGON to better understand its superior performance over competing methods. First, it is interesting and necessary to investigate the relationship between λ and ρ. For fixed ρ, one may expect that the optimal λ depends on ρ as in the elastic net [134], and a careful characterization of such dependence is important for tuning λ in practice. In terms of tuning λ, the standard approach such as cross-validation is computationally expensive for nodewise regressions. Other than the cross-validation, a variety of information criteria have been proposed [37, 151–153], and we would like to generalize them for tuning λ in DRAGON. Furthermore, the data-adaptive selection of the radius of Wasserstein-2 ball

ρ is a long standing challenge in distributional stochastic optimization. This depends on quantifying/estimating the deviation between the assumed data generation distribution and the empirical distribution from data, which may be resolved using the lately developed concentration results on the Wasserstein distance in high-dimensional regime. We plan to investigate the performance of DRAGON in analyzing the Genotype-Tissue Expression (GTEx) data, which is known to possess multiple potentially heterogeneous sub-populations. [131] analyzed this dataset to recover the gene network in a target brain tissue under the transfer learning setting, which assume the full knowledge of sub-populations. We will leave these to the immediate future works.

Theoretically, we will study the consistency of DRAGON under different contamination settings and investigate the gain of DRAGON in terms of robustness compared to the traditional methods, such as the vanilla nodewise regression estimator and Glasso. Inference based on the DRAGON estimator will be a separate and important question. Similar to the Lasso-based Gaussian graphical model methods, it is expected that the de-biasing step [48, 50, 51] will be needed to draw DRAGON-based inference on edges. However, the shrinkage effect of ρ will make the de-biasing nontrivial and requires the understanding of the relationship between ρ and λ, as mentioned above. These will be studied in separate works.

Lastly, extending the proposed methods in both Sections 3 and 4 to non-Gaussian graphical model is highly relevant to real scientific problems. Particularly, DRAGON can be easily generalized via combining the exponential graphical model with distributionally robust generalized linear regressions. Several nodewise regression-based exponential family graph estimations have flourished in the literature such as the high-dimensional Ising models [105, 154]. On the other hand, distributionally robust regressions for the generalized linear model have also been studied, such as the distributionally robust logistic regression [16] and the general maximum likelihood estimation under the distributionally robust framework [155]. These help shed light on generalizing DRAGON for non-Gaussian graphical models and will be studied in the future.

Bibliography

[1] Peter J. Huber. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964.

[2] Peter J. Huber and Elvezio M. Ronchetti. *Robust Statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, USA, 2009.

[3] John W. Tukey. A survey of sampling from contaminated distributions. In *Contributions to probability and statistics: Essays in Honor of Harold Hotelling*, pages 448–485. 1960.

[4] Frank R. Hampel. A General Qualitative Definition of Robustness. *The Annals of Mathematical Statistics*, 42(6):1887–1896, 1971.

[5] Frank R. Hampel. The Influence Curve and Its Role in Robust Estimation. *Journal of the American Statistical Association*, 69(346):383, 1974.

[6] Fatemah Alqallaf, Stefan Van Aelst, Victor J. Yohai, and Ruben H. Zamar. Propagation of outliers in multivariate data. *The Annals of Statistics*, 37(1):311–331, 2009.

[7] Jianqing Fan, Yuan Ke, Qiang Sun, and Wen-Xin Zhou. FarmTest: Factor-Adjusted Robust Multiple Testing With Approximate False Discovery Control. *Journal of the American Statistical Association*, 114(528):1880–1893, 2019.

[8] Olivier Catoni. Challenging the empirical mean and empirical variance: A deviation study. *Annales de l'institut Henri Poincare (B) Probability and Statistics*, 48(4):1148–1185, 2012.

[9] Jianqing Fan, Quefeng Li, and Yuyan Wang. Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(1):247–265, 2017.

[10] Stanislav Minsker. Sub-Gaussian estimators of the mean of a random matrix with heavy-tailed entries. *The Annals of Statistics*, 46(6A):2871–2903, 2018.

[11] Yuan Ke, Stanislav Minsker, Zhao Ren, Qiang Sun, and Wen-Xin Zhou. User-Friendly Covariance Estimation for Heavy-Tailed Distributions. *Statistical Science*, 34(3):454–471, 2019.

[12] Qiang Sun, Wen-Xin Zhou, and Jianqing Fan. Adaptive Huber Regression. *Journal of the American Statistical Association*, 115(529):254–265, 2020.

[13] Aurore Delaigle, Peter Hall, and Jiashun Jin. Robustness and accuracy of methods for high dimensional data analysis based on Student's t-statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3):283–301, 2011.

[14] Stanislav Minsker. Distributed statistical estimation and rates of convergence in normal approximation. *Electronic Journal of Statistics*, 13(2):5213–5252, 2019.

[15] Huan Xu, Constantine Caramanis, and Shie Mannor. Robust Regression and Lasso. *IEEE Transactions on Information Theory*, 56(7):3561–3574, 2010.

[16] Soroosh Shafieezadeh-Abadeh, Peyman Mohajerin Esfahani, and Daniel Kuhn. Distributionally robust logistic regression. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 1576–1584, 2015.

[17] Rui Gao, Xi Chen, and Anton J. Kleywegt. Wasserstein Distributional Robustness and Regularization in Statistical Learning. *arXiv preprint arXiv:1712.06050*, 2017.

[18] Aman Sinha, Hongseok Namkoong, Riccardo Volpi, and John Duchi. Certifying Some Distributional Robustness with Principled Adversarial Training. *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, pages 1–49, oct 2017.

[19] John Duchi and Hongseok Namkoong. Learning Models with Uniform Performance via Distributionally Robust Optimization. *The Annals of Statistics*, 2021+. to appear, arXiv preprint arXiv:1810.08750.

[20] Jose Blanchet, Yang Kang, and Karthyek Murthy. Robust Wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, 56(3):830–857, 2019.

[21] Zheng Liu and Po-Ling Loh. Robust W-GAN-Based Estimation Under Wasserstein Contamination. *arXiv preprint arXiv:2101.07969*, 2021.

[22] Imre Csiszár. Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten. *Publications of the Mathematical Institute of Hungarian Academy of Sciences*, 8:85–108, 1963.

[23] Tetsuzo Morimoto. Markov Processes and the H -Theorem. *Journal of the Physical Society of Japan*, 18(3):328–331, 1963.

[24] S. M. Ali and S. D. Silvey. A General Class of Coefficients of Divergence of One Distribution from Another. *Journal of the Royal Statistical Society: Series B (Methodological)*, 28(1):131–142, 1966.

[25] Cédric Villani. *Optimal Transport: Old and New*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.

[26] Erick Delage and Yinyu Ye. Distributionally Robust Optimization Under Moment Uncertainty with Application to Data-Driven Problems. *Operations Research*, 58(3):595–612, 2010.

[27] Ruiwei Jiang and Yongpei Guan. Data-driven chance constrained stochastic program. *Mathematical Programming*, 158:291–327, 2016.

[28] John C. Duchi, Peter W. Glynn, and Hongseok Namkoong. Statistics of Robust Optimization: A Generalized Empirical Likelihood Approach. *Mathematics of Operations Research*, 2021. to appear, http://pubsonline.informs.org/doi/10.1287/moor.2020.1085.

[29] John Duchi and Hongseok Namkoong. Variance-based regularization with convex objectives. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[30] Viet Anh Nguyen, Daniel Kuhn, and Peyman Mohajerin Esfahani. Distributionally Robust Inverse Covariance Estimation: The Wasserstein Shrinkage Estimator. *arXiv preprint arXiv:1805.07194*, 2018.

[31] Peyman Mohajerin Esfahani and Daniel Kuhn. *Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations*, volume 171. Springer Berlin Heidelberg, 2018.

[32] Daniel Kuhn, Peyman Mohajerin Esfahani, Viet Anh Nguyen, and Soroosh Shafieezadeh-Abadeh. Wasserstein Distributionally Robust Optimization: Theory and Applications in Machine Learning. In *Operations Research & Management Science in the Age of Analytics*, number 2, pages 130–166. INFORMS, 2019.

[33] Jose Blanchet and Karthyek Murthy. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600, 2019.

[34] Hamed Rahimian and Sanjay Mehrotra. Distributionally Robust Optimization: A Review. *arXiv preprint arXiv:1908.05659*, 2019.

[35] Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics*, 34(3):1436–1462, 2006.

[36] Robert Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[37] Ming Yuan and Yi Lin. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.

[38] Onureena Banerjee, Laurent El Ghaoui, and Alexandre d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine Learning Research*, 9(15):485–516, 2008.

[39] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[40] Tony Cai, Weidong Liu, and Xi Luo. A Constrained l1 Minimization Approach to Sparse Precision Matrix Estimation. *Journal of the American Statistical Association*, 106:594–607, 2011.

[41] Daniela M. Witten, Jerome H. Friedman, and Noah Simon. New Insights and Faster Computations for the Graphical Lasso. *Journal of Computational and Graphical Statistics*, 20(4):892–900, 2011.

[42] Mátyás A. Sustik and B. Calderhead. Glassofast : An efficient glasso implementation. Technical report, UTCS Technical Report TR-12-29, 2012.

[43] Matyas A. Sustik, Ben Calderhead, and Julien Clavel. *glassoFast: Fast Graphical LASSO*, 2018. R package version 1.0.

[44] Haotian Pang, Han Liu, and Robert Vanderbei. The fastclime package for linear programming and large-scale precision matrix estimation in R. *Journal of Machine Learning Research*, 15:489–493, 2014.

[45] Ming Yuan. High Dimensional Inverse Covariance Matrix Estimation via Linear Programming. *Journal of Machine Learning Research*, 11:2261–2286, 2010.

[46] Han Liu, Fang Han, Ming Yuan, John Lafferty, and Larry Wasserman. High-dimensional semiparametric Gaussian copula graphical models. *The Annals of Statistics*, 40(4), 2012.

[47] Tingni Sun and Cun Hui Zhang. Sparse matrix inversion with scaled lasso. *Journal of Machine Learning Research*, 14:3385–3418, 2013.

[48] Weidong Liu. Gaussian graphical model estimation with false discovery rate control. *The Annals of Statistics*, 41(6):2948–2978, 2013.

[49] Emmanuel Candes and Terence Tao. The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, 2007.

[50] Jana Janková and Sara van de Geer. Honest confidence regions and optimality in high-dimensional precision matrix estimation. *TEST*, 26(1):143–162, 2015.

[51] Jana Jankova and Sara van de Geer. Confidence intervals for high-dimensional inverse covariance estimation. *Electronic Journal of Statistics*, 9:1205–1229, 2015.

[52] Zhao Ren, Tingni Sun, Cun-Hui Zhang, and Harrison H. Zhou. Asymptotic normality and optimalities in estimation of large Gaussian graphical models. *The Annals of Statistics*, 43(3):991–1026, 2015.

[53] Tingni Sun and Cun Hui Zhang. Scaled sparse linear regression. *Biometrika*, 99(4):879–898, 2012.

[54] Rong Zhang, Zhao Ren, and Wei Chen. SILGGM: An extensive R package for efficient statistical inference in large-scale gene networks. *PLoS Computational Biology*, 14(8):1–14, 2018.

[55] Håvard Rue and Leonhard Held. *Gaussian Markov random fields : theory and applications*. Monographs on statistics and applied probability ; 104. Chapman & Hall/CRC, Boca Raton, 2005.

[56] Mark S. Kaiser and Daniel J. Nordman. Blockwise empirical likelihood for spatial Markov model assessment. *Statistics and Its Interface*, 5(3):303–318, 2012.

[57] Mark S. Kaiser, Soumendra N. Lahiri, and Daniel J. Nordman. Goodness of fit tests for a class of Markov random field models. *The Annals of Statistics*, 40(1):104–130, 2012.

[58] Han Liu, John Lafferty, and Larry Wasserman. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research*, 10:2295–2328, 2009.

[59] G. Tarr, S. Müller, and N. C. Weber. Robust estimation of precision matrices under cellwise contamination. *Computational Statistics and Data Analysis*, 93:404–420, 2016.

[60] Viktoria Öllerer and Christophe Croux. *Robust high-dimensional precision matrix estimation*, volume 1, pages 325–350. Springer International Publishing, 2015.

[61] Po-ling Loh and Xin Lu Tan. High-dimensional robust precision matrix estimation: Cellwise corruption under ϵ-contamination. *Electronic Journal of Statistics*, 12(1):1429–1467, 2018.

[62] Matthew E. Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47, 2015.

[63] Reza Khany and Khalil Tazik. Levels of Statistical Use in Applied Linguistics Research Articles: From 1986 to 2015. *J. Quant. Linguist.*, 26(1):48–65, 2019.

[64] T. Tony Cai and Wenguang Sun. Large-Scale Global and Simultaneous Inference: Estimation and Testing in Very High Dimensions. *Annual Review of Economics*, 9(1):411–439, 2017.

[65] Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995.

[66] John D. Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498, 2002.

[67] John D. Storey and Robert Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16):9440–9445, 2003.

[68] Elizabeth Purdom and Susan P. Holmes. Error Distribution for Gene Expression Data. *Statistical Applications in Genetics and Molecular Biology*, 4(1), 2005.

[69] Anders Eklund, Thomas E. Nichols, and Hans Knutsson. Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences*, 113(28):7900–7905, 2016.

[70] George Kingsley Zipf. *Human behavior and the principle of least effort: An introduction to human ecology*. Ravenio Books, 1949.

[71] Adarsh Prasad, Arun Sai Suggala, Sivaraman Balakrishnan, and Pradeep Ravikumar. Robust estimation via robust gradient estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(3):601–627, 2020.

[72] Kean Ming Tan, Qiang Sun, and Daniela Witten. Robust Sparse Reduced Rank Regression in High Dimensions. *arXiv preprint arXiv:1810.07913*, 2018.

[73] Lili Wang, Chao Zheng, Wen-Xin Zhou, and Wen-Xin Zhou. A New Principle for Tuning-Free Huber Regression. *Statistica Sinica*, 2021. to appear.

[74] Peter J. Huber. Robust Regression: Asymptotics, Conjectures and Monte Carlo. *The Annals of Statistics*, 1(5):799–821, 1973.

[75] Martin Gerlach and Francesc Font-Clos. A Standardized Project Gutenberg Corpus for Statistical Analysis of Natural Language and Quantitative Linguistics. *Entropy*, 22(1):126, 2020.

[76] Chloé Friguet, Maela Kloareg, and David Causeur. A Factor Model Approach to Multiple Testing Under Dependence. *Journal of the American Statistical Association*, 104(488):1406–1415, 2009.

[77] Keyur H. Desai and John D. Storey. Cross-dimensional inference of dependent high-dimensional data. *Journal of the American Statistical Association*, 107(497):135–151, 2012.

[78] Jianqing Fan, Xu Han, and Weijie Gu. Estimating false discovery proportion under arbitrary covariance dependence. *Journal of the American Statistical Association*, 107(499):1019–1035, 2012.

[79] Gilles Blanchard and Etienne Roquain. Adaptive False Discovery Rate Control under Independence and Dependence. *Journal of Machine Learning Research*, 10:2837–2871, 2009.

[80] T. Tony Cai, Cun Hui Zhang, and Harrison H. Zhou. Optimal rates of convergence for covariance matrix estimation. *The Annals of Statistics*, 38(4):2118–2144, 2010.

[81] Roman Vershynin. *High-Dimensional Probability An Introduction with Applications in Data Science*. Cambridge University Press, Cambridge, United Kingdom, 2018.

[82] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.

[83] Stefan Th. Gries. Corpus Linguistics: Quantitative Methods. *The Encyclopedia of Applied Linguistics*, 2012.

[84] M.-A. Dillies, A. Rau, J. Aubert, C. Hennequet-Antier, M. Jeanmougin, N. Servant, C. Keime, G. Marot, D. Castel, J. Estelle, G. Guernec, B. Jagla, L. Jouneau, D. Laloe, C. Le Gall, B. Schaeffer, S. Le Crom, M. Guedj, and F. Jaffrezic. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Briefings in Bioinformatics*, 14(6):671–683, 2013.

[85] Dat Quoc Nguyen, Dai Quoc Nguyen, Dang Duc Pham, and Son Bao Pham. A robust transformation-based learning approach using ripple down rules for part-of-speech tagging. *Ai communications*, 29(3):409–422, 2016.

[86] *RDRPOSTagger: Parts of Speech Tagging based on the Ripple Down Rules-based Part-Of-Speech Tagger*, 2017. R package version 1.1, https://github.com/bnosac/RDRPOSTagger.

[87] Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Gaussian approximation of suprema of empirical processes. *The Annals of Statistics*, 42(4):1564–1597, 2014.

[88] Frank R. Hampel. A General Qualitative Definition of Robustness. *The Annals of Mathematical Statistics*, 42(6):1887–1896, 1971.

[89] Shweta Bansal, Bryan T Grenfell, and Lauren Ancel Meyers. When individual behaviour matters: homogeneous and network models in epidemiology. *Journal of The Royal Society Interface*, 4(16):879–891, 2007.

[90] Mark R. T. Dale and Marie-Josée Fortin. *Spatial analysis : a guide for ecologists*. YBP Print DDA. Cambridge University Press, Cambridge, second edition. edition, 2014.

[91] Jia Xue, Susanne V. Schmidt, Jil Sander, Astrid Draffehn, Wolfgang Krebs, Inga Quester, Dominic DeNardo, Trupti D. Gohel, Martina Emde, Lisa Schmidleithner, Hariharasudan Ganesan, Andrea Nino-Castro, Michael R. Mallmann, Larisa Labzin, Heidi Theis, Michael Kraut, Marc Beyer, Eicke Latz, Tom C. Freeman, Thomas Ulas, and Joachim L. Schultze. Transcriptome-Based Network Analysis Reveals a Spectrum Model of Human Macrophage Activation. *Immunity*, 40(2):274–288, 2014.

[92] Mark D.M. Leiserson, Fabio Vandin, Hsin Ta Wu, Jason R. Dobson, Jonathan V. Eldridge, Jacob L. Thomas, Alexandra Papoutsaki, Younhun Kim, Beifang Niu, Michael McLellan, Michael S. Lawrence, Abel Gonzalez-Perez, David Tamborero, Yuwei Cheng, Gregory A. Ryslik, Nuria Lopez-Bigas, Gad Getz, Li Ding, and Benjamin J. Raphael. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nature Genetics*, 47(2):106–114, 2015.

[93] GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*, 348(6235):648–660, 2015.

[94] John Scott. *Social network analysis*. SAGE, Los Angeles, third edition, 2013.

[95] Catherine Hobaiter, Timothée Poisot, Klaus Zuberbühler, William Hoppitt, and Thibaud Gruber. Social Network Analysis Shows Direct Evidence for Social Transmission of Tool Use in Wild Chimpanzees. *PLoS Biology*, 12(9), 2014.

[96] Damien R. Farine and Hal Whitehead. Constructing, conducting and interpreting animal social network analysis. *Journal of Animal Ecology*, 84(5):1144–1163, 2015.

[97] Linda Hartman and Ola Hössjer. Fast kriging of large data sets with Gaussian Markov random fields. *Computational Statistics and Data Analysis*, 52(5):2331–2349, 2008.

[98] Mladen Kolar and Eric P. Xing. Sparsistent Estimation of Time-Varying Discrete Markov Random Fields. *arXiv preprint arXiv:0907.2337*, pages 1–34, 2009.

[99] Finn Lindgren, Håvard Rue, and Johan Lindström. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498, 2011.

[100] Martin J. Wainwright and Michael I. Jordan. Graphical Models, Exponential Families, and Variational Inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2007.

[101] Kevin P. Murphy. Undirected Graphical Models (Markov Random Fields). In *Machine Learning: A Probabilistic Perspective*, volume 2, chapter 19, pages 661–705. MIT Press, 2012.

[102] Han Liu and Lie Wang. TIGER: A tuning-insensitive approach for optimally estimating gaussian graphical models. *Electronic Journal of Statistics*, 11(1):241–294, 2017.

[103] Jialei Wang and Mladen Kolar. Inference for Sparse Conditional Precision Matrices. *arXiv preprint arXiv:1412.7638*, 2014.

[104] Yunzhang Zhu, Xiaotong Shen, and Wei Pan. Structural pursuit over multiple undirected graphs. *Journal of the American Statistical Association*, 109(508):1683–1696, 2014.

[105] Pradeep Ravikumar, Martin J. Wainwright, and John D. Lafferty. High-dimensional Ising model selection using ℓ_1 -regularized logistic regression. *The Annals of Statistics*, 38(3):1287–1319, 2010.

[106] Mladen Kolar and Eric P. Xing. Estimating networks with jumps. *Electronic Journal of Statistics*, 6:2069–2106, 2012.

[107] Sandipan Roy, Yves Atchadé, and George Michailidis. Change point estimation in high dimensional Markov random-field models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology*, 79(4):1187–1206, 2017.

[108] Yin Xia and Lexin Li. Hypothesis testing of matrix graph model with application to brain connectivity analysis. *Biometrics*, 73(3):780–791, 2017.

[109] Junwei Lu, Mladen Kolar, and Han Liu. Post-regularization inference for time-varying nonparanormal graphical models. *Journal of Machine Learning Research*, 18, 2018.

[110] Yin Xia, Tianxi Cai, and T Tony Cai. Testing differential networks with applications to the detection of gene-gene interactions. *Biometrika*, 102(2):247–266, 2015.

[111] Jinyuan Chang, Yumou Qiu, Qiwei Yao, and Tao Zou. Confidence regions for entries of a large precision matrix. *Journal of Econometrics*, 206(1):57–82, 2018.

[112] Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Gaussian approximation of suprema of empirical processes. *The Annals of Statistics*, 42(4):1564–1597, 2014.

[113] Donald W. K. Andrews. Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation. *Econometrica*, 59(3):817–858, 1991.

[114] Jeffery K. Taubenberger and David M. Morens. 1918 Influenza: the Mother of All Pandemics. *Emerging Infectious Diseases*, 12(1):15–22, 2006.

[115] Fatimah S. Dawood, A. Danielle Iuliano, Carrie Reed, Martin I. Meltzer, David K. Shay, Po Yung Cheng, Don Bandaranayake, Robert F. Breiman, W. Abdullah Brooks, Philippe

Buchy, Daniel R. Feikin, Karen B. Fowler, Aubree Gordon, Nguyen Tran Hien, Peter Horby, Q. Sue Huang, Mark A. Katz, Anand Krishnan, Renu Lal, Joel M. Montgomery, Kåre Mølbak, Richard Pebody, Anne M. Presanis, Hugo Razuri, Anneke Steens, Yeny O. Tinoco, Jacco Wallinga, Hongjie Yu, Sirenda Vong, Joseph Bresee, and Marc Alain Widdowson. Estimated global mortality associated with the first 12 months of 2009 pandemic influenza A H1N1 virus circulation: A modelling study. *The Lancet Infectious Diseases*, 12(9):687–695, 2012.

[116] World Health Organization. Influenza (Seasonal).

[117] Jeremy Ginsberg, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014, 2009.

[118] Samantha Cook, Corrie Conrad, Ashley L. Fowlkes, and Matthew H. Mohebbi. Assessing Google Flu trends performance in the United States during the 2009 influenza virus A (H1N1) pandemic. *PLoS One*, 6(8):1–8, 2011.

[119] Mevin B. Hooten, Jessica Anderson, and Lance A. Waller. Spatial and Spatio-temporal Epidemiology Assessing North American influenza dynamics with a statistical SIRS model. *Spatial and Spatio-temporal Epidemiology*, 1(2-3):177–185, 2010.

[120] Andrea Freyer Dugas, Mehdi Jalalpour, Yulia Gel, Scott Levin, Fred Torcaso, Takeru Igusa, and Richard E. Rothman. Influenza Forecasting with Google Flu Trends. *PLoS One*, 8(2), 2013.

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

[122] Peter J. Brockwell and Richard A. Davis. *Introduction to Time Series and Forecasting*. Springer Texts in Statistics. Springer International Publishing, Cham, 2016.

[123] Richard C. Dicker, Fatima Coronado, Denise Koo, and R. Gibson Parrish. *Principles of Epidemiology in Public Health Practice*. 2006.

[124] Cécile Viboud, Ottar N. Bjornstad, David L. Smith, Lone Simonsen, Mark A. Miller, and Bryan T. Grenfell. Synchrony, Waves, and Spatial Hierarchies in the Spread of Influenza. *Science*, 312(5772):447–451, 2006.

[125] Nam H. Nguyen and Trac D. Tran. Robust lasso with missing and grossly corrupted observations. *IEEE Transactions on Information Theory*, 59(4):2036–2058, 2013.

[126] Arnak Dalalyan and Philip Thompson. Outlier-robust estimation of a sparse linear model using ℓ_1-penalized huber's m-estimator. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[127] Takeyuki Sasai and Hironori Fujisawa. Robust estimation with Lasso when outputs are adversarially contaminated. *arXiv preprint arXiv:2004.05990*, 2020.

[128] John Duchi, Tatsunori Hashimoto, and Hongseok Namkoong. Distributionally Robust Losses for Latent Covariate Mixtures. *arXiv preprint arXiv:2007.13982*, pages 1–39, 2020.

[129] Weibin Mo, Zhengling Qi, and Yufeng Liu. Learning Optimal Distributionally Robust Individualized Treatment Rules. *Journal of the American Statistical Association*, 116(534):659–674, 2021.

[130] Marco Avella-Medina, Heather S. Battey, Jianqing Fan, and Quefeng Li. Robust estimation of high-dimensional covariance and precision matrices. *Biometrika*, 105(2):271–284, 2018.

[131] Sai Li, T. Tony Cai, and Hongzhe Li. Transfer Learning in Large-scale Gaussian Graphical Models with False Discovery Rate Control. *arXiv preprint arXiv:2010.11037*, 2020.

[132] Wessel N. Van Wieringen and Carel F.W. Peeters. Ridge estimation of inverse covariance matrices from high-dimensional data. *Computational Statistics and Data Analysis*, 103:284–303, 2016.

[133] M. O. Kuismin, J. T. Kemppainen, and M. J. Sillanpää. Precision matrix estimation with rope. *Journal of Computational and Graphical Statistics*, 26(3):682–694, 2017.

[134] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

[135] Mihai Cucuringu, Jesus Puente, and David Shue. Model Selection in Undirected Graphical Models with the Elastic Net. *arXiv preprint arXiv:1111.0559*, 2011.

[136] Davide Bernardini, Sandra Paterlini, and Emanuele Taufer. New estimation approaches for graphical models with elastic net penalty. *arXiv preprint arXiv:2102.01053*, 2021.

[137] Solt Kovács, Tobias Ruckstuhl, Helena Obrist, and Peter Bühlmann. Graphical Elastic Net and Target Matrices: Fast Algorithms and Software for Sparse Precision Matrix Estimation. *arXiv preprint arXiv:2101:02148*, 2021.

[138] Vu Pham and Laurent El Ghaoui. Robust sketching for multiple square-root LASSO problems. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pages 753–761. PMLR, 2015.

[139] Jose Blanchet, Fei He, and Karthyek Murthy. On distributionally robust extreme value analysis. *Extremes*, 23(2):317–347, 2020.

[140] Victor M. Panaretos and Yoav Zemel. Statistical Aspects of Wasserstein Distances. *Annual Review of Statistics and Its Application*, 6(1):405–431, 2019.

[141] A. Belloni, V. Chernozhukov, and L. Wang. Square-root lasso: Pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011.

[142] Jianqing Fan, Yuan Liao, and Han Liu. An overview of the estimation of large covariance and precision matrices. *Econometrics Journal*, 19(1):C1–C32, 2016.

[143] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN. *arXiv preprint arXiv:1701.07875*, 2017.

[144] Clark R. Givens and Rae Michael Shortt. A class of Wasserstein metrics for probability distributions. *The Michigan Mathematical Journal*, 31(2):231–240, 1984.

[145] D. C. Dowson and B. V. Landau. The Fréchet distance between multivariate normal distributions. *Journal of Multivariate Analysis*, 12(3):450–455, 1982.

[146] Matthias Gelbrich. On a Formula for the L^2 Wasserstein Metric between Measures on Euclidean and Hilbert Spaces. *Mathematische Nachrichten*, 147(1):185–203, 1990.

[147] Elias Raninen and Esa Ollila. Scaled and square-root elastic net. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, pages 4336–4340. IEEE, 2017.

[148] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical Learning with Sparsity*. Chapman and Hall/CRC, 2015.

[149] Dirk Eddelbuettel. *Seamless R and C++ Integration with Rcpp*. Springer New York, New York, NY, 2013.

[150] Dirk Eddelbuettel and Conrad Sanderson. RcppArmadillo: Accelerating R with high-performance C++ linear algebra. *Computational Statistics and Data Analysis*, 71:1054–1063, 2014.

[151] Rina Foygel and Mathias Drton. Extended bayesian information criteria for gaussian graphical models. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 1*, NIPS'10, page 604–612, 2010.

[152] Shaun Lysen. *Permuted Inclusion Criterion: A Variable Selection Technique*. Phd thesis, 2009.

[153] Jie Peng, Pei Wang, Nengfeng Zhou, and Ji Zhu. Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association*, 104(486):735–746, 2009.

[154] Ali Jalali, Pradeep Ravikumar, Vishvas Vasuki, and Sujay Sanghavi. On learning discrete graphical models using group-sparse regularization. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15, pages 378–387, 2011.

[155] Viet Anh Nguyen, Xuhui Zhang, Jose Blanchet, and Angelos Georghiou. Distributionally Robust Parametric Maximum Likelihood Estimation. *arXiv preprint arXiv:2010.05321*, 1, 2020.

[156] Walter Rudin. *Principles of mathematical analysis*. McGraw-hill New York, 1976.

[157] Wen-Xin Zhou, Koushiki Bose, Jianqing Fan, and Han Liu. A new perspective on robust M-estimation: Finite sample theory and applications to dependence-adjusted multiple testing. *The Annals of Statistics*, 46(5):1904–1931, 2018.

[158] Vladimir Spokoiny and Mayya Zhilova. Bootstrap confidence sets under model misspecification. *The Annals of Statistics*, 43(6):2653–2675, 2015.

[159] V. Bentkus. A Lyapunov-type Bound in $\mathbb{R}^d$. *Theory of Probability & Its Applications*, 49(2):311–323, 2005.

[160] Wenbo V. Li and Qi Man Shao. A normal comparison inequality and its applications. *Probability Theory and Related Fields*, 122(4):494–508, 2002.

[161] W. F. Kibble. A Two-Variate Gamma Type Distribution. *Sankhya: The Indian Journal of Statistics*, 5(2):137–150, 1941.

[162] A. S. Krishnamoorthy and M. Parthasarathy. A Multivariate Gamma-Type Distribution. *The Annals of Mathematical Statistics*, 22(4):549–557, 1951.

[163] D. R. Jensen. An Inequality for a Class of Bivariate Chi-Square Distributions. *Journal of the American Statistical Association*, 64(325):333–336, 1969.

[164] Ilia Krasikov. Inequalities for orthonormal Laguerre polynomials. *Journal of Approximation Theory*, 144(1):1–26, 2007.

[165] Luc Devroye, Abbas Mehrabian, and Tommy Reddad. The total variation distance between high-dimensional Gaussians. *arXiv preprint arXiv:1810.08693*, 2018.

[166] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

[167] Roger A. Horn and Charles R. Johnson. *Matrix analysis*. Cambridge university press, 1990.

[168] Xiaoying Tian, Joshua R. Loftus, and Jonathan E. Taylor. Selective inference with unknown variance via the square-root lasso. *Biometrika*, 105(4):755–768, 2018.

[169] Ryan J. Tibshirani. The lasso problem and uniqueness. *Electronic Journal of Statistics*, 7(none):1–25, 2013.

[170] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2010.

Appendix A

Supplementary materials for Chapter 2

A.1 Proofs of Main Theorems

A.1.1 Proof of Theorem 2.3.1

For ease of exposition, we use z^2 in place of z in this proof. To prove Theorem 2.3.1, we will prove a stronger result that

$$p_0^{-1}V^\circ(z^2) = \mathbb{P}(\chi_q^2 > z^2) + O_{\mathbb{P}}(q^{1/2}p^{-\kappa_1} + n^{-1/2}q^{7/4} + q[n^{-1}\{\log(np) + d\}]^{\delta/(2+\delta)}) \tag{A.1.1}$$

uniformly over $z \geq 0$ as $n, p \to \infty$. Denote $\mathbf{A}_j := \mathbf{C}\boldsymbol{\Sigma}_j\mathbf{C}^{\mathrm{T}} \in \mathbb{R}^{q\times q}$ the true covariance matrix of $n^{1/2}\mathbf{C}(\widehat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j)$, and $\mathbf{A}_j^{1/2}$ the square root of $\mathbf{A}_j$. The proof consists of two steps: first, we sandwich the number of false discoveries using the Bahadur representation of

$$\mathbf{T}_j^\circ = n^{1/2}\mathbf{A}_j^{-1/2}\mathbf{C}(\widehat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j), \tag{A.1.2}$$

where $V_j^\circ = \|\mathbf{T}_j^\circ\|^2$; then we will show that the bounds converge to $\mathbb{P}(\chi_q^2 > z^2)$ as $n, p \to \infty$.

First, we will show that $\mathbf{T}_j^\circ$ can be approximated by a q-dimensional multivariate normal distribution, so that V_j° can be approximated by the χ_q^2 distribution due to Lemma A.2.2.5. Let

$$\begin{aligned}
\mathbf{S}_j &= n^{-1/2}(\mathbf{C}\boldsymbol{\Sigma}_Z^{-1}\mathbf{C}^{\mathrm{T}})^{-1/2}\mathbf{C}\boldsymbol{\Sigma}_Z^{-1/2}\sum_{i=1}^n \boldsymbol{\Sigma}_Z^{-1/2}\left[\ell'_{\tau_j}(\epsilon_{ij})\mathbf{Z}_i - \mathbb{E}\{\ell'_{\tau_j}(\epsilon_{ij})\mathbf{Z}_i\}\right], \\
\mathbf{R}_j &= n^{-1/2}(\mathbf{C}\boldsymbol{\Sigma}_Z^{-1}\mathbf{C}^{\mathrm{T}})^{-1/2}\mathbf{C}\boldsymbol{\Sigma}_Z^{-1/2}\sum_{i=1}^n \boldsymbol{\Sigma}_Z^{-1/2}\mathbb{E}\{\ell'_{\tau_j}(\epsilon_{ij})\mathbf{Z}_i\}.
\end{aligned} \tag{A.1.3}$$

Note that $\mathbf{R}_j$ is negligible by Conditions 1 and Proposition A.2.1. By Corollary A.2.2,

$$\|\mathbf{T}_j^\circ - \sigma_{\epsilon,jj}^{-1/2}(\mathbf{S}_j + \mathbf{R}_j)\| \leq C_2\tau_{0j}\frac{d+t}{(n\sigma_{\epsilon,jj})^{1/2}} \tag{A.1.4}$$

with probability $1-3\exp(-t)$ as long as $n \geq C_3(d+t)$. For $j=1,\ldots,p$, let $E_{1j}(t)$ be the event on which (A.1.4) holds. Set $E_1(t)=\bigcap_{j=1}^{p} E_{1j}(t)$, on which

$$\begin{aligned}&\sum_{j\in H_0} \mathbb{I}\{\|\sigma_{\epsilon,jj}^{-1/2}\mathbf{S}_j\| \geq z + C_2\tau_{0j}(n\sigma_{\epsilon,jj})^{-1/2}(d+t)\} \leq V^{\circ}(z^2)\\ &\leq \sum_{j\in H_0} \mathbb{I}\{\|\sigma_{\epsilon,jj}^{-1/2}\mathbf{S}_j\| \geq z - C_2\tau_{0j}(n\sigma_{\epsilon,jj})^{-1/2}(d+t)\}.\end{aligned} \tag{A.1.5}$$

with probability $1-3p\exp(-t)$ as long as $n \geq C_4(d+t)$.

Define $D_j = \mathbb{I}(\|\sigma_{\epsilon,jj}^{-1/2}\mathbf{S}_j\| \geq z)$ and $\mathcal{P}_j = \mathbb{P}(\|\sigma_{\epsilon,jj}^{-1/2}\mathbf{S}_j\| \geq z)$ for $j=1,\ldots,p$ and $z \geq 0$. Under Condition 1, $D_1,\ldots,D_p$ are weakly correlated. Recall that $\mathcal{H}_0 = \{j : 1 \leq j \leq p, H_{0j} \text{ is true}\}$, it holds

$$\begin{aligned}\operatorname{var}\left(p_0{}^{-1}\sum_{j\in\mathcal{H}_0} D_j\right) &= \frac{1}{p_0^2}\sum_{j\in\mathcal{H}_0}\operatorname{var}(D_j) + \frac{1}{p_0^2}\sum_{j,k\in\mathcal{H}_0;j\neq k}\operatorname{cov}(D_j,D_k)\\ &\leq \frac{1}{4p_0} + \frac{1}{p_0^2}\sum_{j,k\in\mathcal{H}_0;j\neq k}\{\mathbb{E}(D_jD_k) - \mathcal{P}_j\mathcal{P}_k\}.\end{aligned} \tag{A.1.6}$$

We first study $\mathcal{P}_j$. Note that $\mathbf{S}_j$ is a sum of independent random vectors with $\mathbb{E}(\mathbf{S}_j)=\mathbf{0}$ and $\operatorname{cov}(\mathbf{S}_j) = s_j^2\mathbf{I}$ where $s_j^2 = \mathbb{E}[\{\ell'_{\tau_j}(\epsilon_{ij})\}^2]$. Let $\mathbf{G}\sim N(\mathbf{0},\mathbf{I})$ be a standard normal random vector. Lemmas A.2.2.1 and A.2.2.2 imply that

$$\max_{1\leq j\leq p}|\mathcal{P}_j - \mathbb{P}(\|\mathbf{G}\|\geq z)| \precsim n^{-1/2}q^{7/4} + q^{1/2}\frac{2}{\delta\sigma_{\epsilon,jj}}\frac{v_{j,\delta}^{2+\delta}}{\tau_{0j}^{\delta}}\left(\frac{d+t}{n}\right)^{\delta/(2+\delta)} \tag{A.1.7}$$

holds uniformly over $z\geq 0$.

Following that, we consider $\mathbb{E}(D_jD_k)$ for each pair (j,k) with $1\leq j\neq k\leq p$. Set $\mathbf{S} = (s_j^{-1}\mathbf{S}_j, s_k^{-1}\mathbf{S}_k)^{\mathrm{T}}$. Let $\mathbf{G}=(\mathbf{G}_1,\mathbf{G}_2)=(G_{11},\ldots,G_{1q},G_{21},\ldots,G_{2q})^{\mathrm{T}}\in\mathbb{R}^{2q}$ be a Gaussian vector with $\mathbb{E}(\mathbf{G})=\mathbf{0}$ and $\operatorname{cov}(\mathbf{G})=\operatorname{cov}(\mathbf{S})$. The block-structured matrix $\operatorname{cov}(\mathbf{S})$ has unit diagonal entries with $\operatorname{cov}(s_j^{-1}\mathbf{S}_j)=\mathbf{I}$. Also, $\operatorname{cov}(s_j^{-1}\mathbf{S}_j, s_k^{-1}\mathbf{S}_k) = (ns_js_k)^{-1}\sum_{i=1}^{n}\operatorname{cov}(\ell'_{\tau_j}(\epsilon_{ij}),\ell'_{\tau_k}(\epsilon_{ik}))\mathbf{I}$ and

$$\left|(ns_js_k)^{-1}\sum_{i=1}^{n}\operatorname{cov}(\ell'_{\tau_j}(\epsilon_{ij}),\ell'_{\tau_k}(\epsilon_{ik}))-r_{\epsilon,jk}\right|\precsim\frac{v_{jk}}{\tau_0^{\delta}}\left(\frac{d+t}{n}\right)^{\delta/(2+\delta)}\tag{A.1.8}$$

by Corollary A.2.1 and Proposition A.2.2, where $\tau_0=\min(\tau_{0j},\tau_{0k})$ and $v_{jk}=\max\{\mathbb{E}(|\epsilon_j|^{2+\delta}),\mathbb{E}(|\epsilon_k|^{2+\delta})\}<\infty$ for some $\delta\in(0,2]$ and $1\le j\ne k\le p$. Putting together (A.1.8), Condition 1 iv), Corollary A.2.1, and Lemmas A.2.2.1 and A.2.2.4 with our choice on τ_j yield

$$\begin{aligned}
&|\,\mathbb{P}(\|\sigma_{\epsilon,jj}^{-1/2}s_j\mathbf{G}_1\|\ge x,\|\sigma_{\epsilon,kk}^{-1/2}s_k\mathbf{G}_2\|\ge x)-\mathbb{P}(\|\mathbf{Z}_1\|\ge x)\,\mathbb{P}(\|\mathbf{Z}_2\|\ge x)|\\
\le&|\,\mathbb{P}(\|\mathbf{G}_1\|\ge\sigma_{\epsilon,jj}^{1/2}s_j^{-1}x,\|\mathbf{G}_2\|\ge\sigma_{\epsilon,kk}^{1/2}s_k^{-1}x)-\mathbb{P}(\|\mathbf{Z}_1\|\ge\sigma_{\epsilon,jj}^{1/2}s_j^{-1}x)\,\mathbb{P}(\|\mathbf{Z}_2\|\ge\sigma_{\epsilon,kk}^{1/2}s_k^{-1}x)|\\
&+|\,\mathbb{P}(\|\sigma_{\epsilon,jj}^{-1/2}s_j\mathbf{Z}_1\|\ge x)\,\mathbb{P}(\|\sigma_{\epsilon,kk}^{-1/2}s_k\mathbf{Z}_2\|\ge x)-\mathbb{P}(\|\mathbf{Z}_1\|\ge x)\,\mathbb{P}(\|\mathbf{Z}_2\|\ge x)|\\
\precsim&\,q^{1/2}|r_{\epsilon,jk}|+q\{n^{-1}(d+t)\}^{\delta/(2+\delta)}.
\end{aligned}$$

It follows that

$$|\,\mathbb{P}(\|\mathbf{G}_1\|\ge s_j^{-1}\sigma_{\epsilon,jj}^{1/2}z,\|\mathbf{G}_2\|\ge s_k^{-1}\sigma_{\epsilon,kk}^{1/2}z)-\{\mathbb{P}(\|\mathbf{Z}\|\ge z)\}^2|\precsim q^{1/2}|r_{\epsilon,jk}|+q\left(\frac{d+t}{n}\right)^{\delta/(2+\delta)}\tag{A.1.9}$$

for $\mathbf{Z}\sim N_q(\mathbf{0},\mathbf{I})$. In addition, Lemma A.2.2.3 gives

$$\sup_{x,y\in\mathbb{R}}|\,\mathbb{P}(\|s_j^{-1}\mathbf{S}_j\|\ge x,\|s_k^{-1}\mathbf{S}_k\|\ge y)-\mathbb{P}(\|\mathbf{G}_1\|\ge x,\|\mathbf{G}_2\|\ge y)|\precsim n^{-1/2}q^{7/4},$$

which implies

$$|\mathbb{E}(D_jD_k)-\mathbb{P}(\|\mathbf{G}_1\|>s_j^{-1}\sigma_{\epsilon,jj}^{1/2}z,\|\mathbf{G}_2\|>s_k^{-1}\sigma_{\epsilon,kk}^{1/2}z)|\precsim n^{-1/2}q^{7/4}.\tag{A.1.10}$$

Putting (A.1.9), (A.1.10), and Lemma A.2.2.1 together, we obtain

$$|\mathbb{E}(D_jD_k)-\{\mathbb{P}(\|\mathbf{Z}_1\|>z)\}^2|\precsim q^{1/2}|r_{\epsilon,jk}|+n^{-1/2}q^{7/4}+q\left(\frac{d+t}{n}\right)^{\delta/(2+\delta)}.\tag{A.1.11}$$

Consequently, it follows from (A.1.6), (A.1.7), (A.1.11), Condition 1, and Lemma A.2.2.5 that

$$\mathbb{E}[\{p_0^{-1}V^\circ(z^2) - \{\mathbb{P}(\chi_q^2 > z^2)\}^2] \precsim q^{1/2}p^{-\kappa_1} + n^{-1/2}q^{7/4} + q\left(\frac{d+t}{n}\right)^{\delta/(2+\delta)} \tag{A.1.12}$$

on $E_1(t)$. Recall that $\mathbb{P}\{E_1(t)\} = 1 - 3p\exp(-t)$ as long as $n \geq C_3(d+t)$. Taking $t = \log(np)$ in (A.1.5) and (A.1.12) proves (A.1.1).

A.1.2 Proof of Proposition 2.3.1

For statistic $\tilde{V}_j = n(\mathbf{C}\widehat{\boldsymbol{\theta}}_j - \mathbf{c}_{0j})^{\mathrm{T}}(\mathbf{C}\widetilde{\boldsymbol{\Sigma}}_j\mathbf{C}^{\mathrm{T}})^{-1}(\mathbf{C}\widehat{\boldsymbol{\theta}}_j - \mathbf{c}_{0j})$, it holds

$$\begin{aligned}
|\tilde{V}_j - V_j^\circ| &= |n(\mathbf{C}\widehat{\boldsymbol{\theta}}_j - \mathbf{c}_{0j})^{\mathrm{T}}(\mathbf{C}\widetilde{\boldsymbol{\Sigma}}_j\mathbf{C}^{\mathrm{T}})^{-1}(\mathbf{C}\widehat{\boldsymbol{\theta}}_j - \mathbf{c}_{0j}) - n(\mathbf{C}\widehat{\boldsymbol{\theta}}_j - \mathbf{c}_{0j})^{\mathrm{T}}(\mathbf{C}\boldsymbol{\Sigma}_j\mathbf{C}^{\mathrm{T}})^{-1}(\mathbf{C}\widehat{\boldsymbol{\theta}}_j - \mathbf{c}_{0j})| \\
&\leq n\|\boldsymbol{\Sigma}_Z^{1/2}(\widehat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j)\|^2\|\boldsymbol{\Sigma}_Z^{-1/2}\mathbf{C}^{\mathrm{T}}\{(\mathbf{C}\widetilde{\boldsymbol{\Sigma}}_j\mathbf{C}^{\mathrm{T}})^{-1} - (\mathbf{C}\boldsymbol{\Sigma}_j\mathbf{C}^{\mathrm{T}})^{-1}\}\mathbf{C}\boldsymbol{\Sigma}_Z^{-1/2}\| \\
&\leq n\|\boldsymbol{\Sigma}_Z^{1/2}(\widehat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j)\|^2\|\mathbf{C}\boldsymbol{\Sigma}_Z^{-1/2}\|^2\|(\mathbf{C}\widetilde{\boldsymbol{\Sigma}}_j\mathbf{C}^{\mathrm{T}})^{-1} - (\mathbf{C}\boldsymbol{\Sigma}_j\mathbf{C}^{\mathrm{T}})^{-1}\| \\
&\precsim n\|\boldsymbol{\Sigma}_Z^{1/2}(\widehat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j)\|^2\|(\mathbf{C}\widetilde{\boldsymbol{\Sigma}}_j\mathbf{C}^{\mathrm{T}})^{-1} - (\mathbf{C}\boldsymbol{\Sigma}_j\mathbf{C}^{\mathrm{T}})^{-1}\|.
\end{aligned}$$

Given $\max_{1\leq j\leq p}\|\widetilde{\boldsymbol{\Sigma}}_j - \boldsymbol{\Sigma}_j\| = o_{\mathbb{P}}[\{\log(np)+d\}^{-1}]$, Lemmas A.2.1.1 and A.2.3.2 imply that

$$\max_{j\in\mathcal{H}_0}\|\tilde{V}_j - V_j^\circ\| \precsim \{\log(np)+d\}\max_{1\leq j\leq p}\|\widetilde{\boldsymbol{\Sigma}}_j - \boldsymbol{\Sigma}_j\| \tag{A.1.13}$$

with probability $1 - 2n^{-1}$ and the right hand side of (A.1.13) is $o_{\mathbb{P}}(1)$. Combining this with the proof of Theorem 2.3.1 and Condition 1, we obtain $p_0^{-1}\widetilde{V}(z) = \mathbb{P}(\chi_q^2 > z) + o_{\mathbb{P}}(1)$. Similarly, $R(z)$ can be replaced by $\widetilde{R}(z)$. Consequently, $|\widetilde{\mathrm{FDP}}(z) - \widetilde{\mathrm{AFDP}}(z)| = o_{\mathbb{P}}(1)$ as $n, p \to \infty$.

A.1.3 Proof of Theorem 2.3.2

Recall that $m_j = n^{-1}\sum_{i=1}^n \mathbb{I}_{\tau_j}^*(e_{ij})$, $\mathbf{W}_j = n^{-1}\sum_{i=1}^n\{\mathbb{I}_{\tau_j}^*(e_{ij})\mathbf{Z}_i\mathbf{Z}_i^{\mathrm{T}}\}$, and $K_j = 1 + (nm_j)^{-1}(d+1)(1-m_j)$. Denote $\mathbf{A}_{jn} := \{\mathbf{W}_j^{-1}(n^{-1}\sum_{i=1}^n \mathbf{Z}_i\mathbf{Z}_i^{\mathrm{T}})\mathbf{W}_j^{-1}\}^{-1}$. For each $j = 1, \ldots, p$,

$$\begin{aligned}\|\widehat{\boldsymbol{\Sigma}}_j - \boldsymbol{\Sigma}_j\| &= \left\| \frac{1}{K_j} \left[\frac{1}{n-d-1} \sum_{i=1}^{n} \{\ell'_{\tau_j}(e_{ij})\}^2 \right] \mathbf{A}_{jn}^{-1} - \sigma_{\epsilon,jj} {\boldsymbol{\Sigma}_Z}^{-1} \right\| \\ &\leq \left| \frac{1}{K_j} \left[\frac{1}{n-d-1} \sum_{i=1}^{n} \{\ell'_{\tau_j}(e_{ij})\}^2 \right] \right| \|\mathbf{A}_{jn}^{-1} - {\boldsymbol{\Sigma}_Z}^{-1}\| \\ &\quad + \left| \frac{1}{K_j} \left[\frac{1}{n-d-1} \sum_{i=1}^{n} \{\ell'_{\tau_j}(e_{ij})\}^2 \right] - \sigma_{\epsilon,jj} \right| \|\boldsymbol{\Sigma}_Z^{-1}\|. \end{aligned} \tag{A.1.14}$$

Note $n/(n-d-1) \to 1$ as $n \to \infty$. Denote $K_j^{-1} n^{-1} \sum_{i=1}^{n} \{\ell'_{\tau_j}(e_{ij})\}^2$ by $f(m_j, y_j)$, where $y_j = n^{-1} \sum_{i=1}^{n} \{\ell'_{\tau_j}(e_{ij})\}^2$. That is, $f[m_j, n^{-1} \sum_{i=1}^{n} \{\ell'_{\tau_j}(e_{ij})\}^2] = K_j^{-1} n^{-1} \sum_{i=1}^{n} \{\ell'_{\tau_j}(e_{ij})\}^2$ with $f(1, \sigma_{\epsilon,jj}) = \sigma_{\epsilon,jj}$. It holds that $f(m_j, y_j)$ is twice differentiable at $(1, \sigma_{\epsilon,jj})$, where

$$\begin{aligned} &\frac{\partial}{\partial y_j} f(m_j, y_j) = K_j^{-1}, \ \frac{\partial}{\partial y_j} f(m_j, y_j)|_{(m_j, y_j) = (1, \sigma_{\epsilon,jj})} = 1, \\ &\frac{\partial}{\partial m_j} f(m_j, y_j) = \frac{n(d+1)}{\{n m_j + (d+1)(1-m_j)\}^2} y_j, \ \frac{\partial}{\partial m_j} f(m_j, y_j)|_{(m_j, y_j) = (1, \sigma_{\epsilon,jj})} = \frac{d+1}{n} \sigma_{\epsilon,jj} \end{aligned}$$

and

$$\begin{aligned} &\frac{\partial^2}{\partial y_j^2} f(m_j, y_j) = 0, \ \frac{\partial^2}{\partial m_j \partial y_j} f(m_j, y_j) = \frac{n(d+1)}{\{n m_j + (d+1)(1-m_j)\}^2}, \\ &\frac{\partial^2}{\partial m_j^2} f(m_j, y_j) = \frac{-2n(d+1)(n-d-1)}{\{n m_j + (d+1)(1-m_j)\}^3} y_j. \end{aligned}$$

Apply Taylor's theorem on $f[m_j, n^{-1} \sum_{i=1}^{n} \{\ell'_{\tau_j}(e_{ij})\}^2]$ with respect to m_j and $n^{-1} \sum \{\ell'_{\tau_j}(e_{ij})\}^2$ at 1 and $\sigma_{\epsilon,jj}$, it follows that

$$\begin{aligned} \frac{1}{K_j} \left[\frac{1}{n} \sum_{i=1}^{n} \{\ell'_{\tau_j}(e_{ij})\}^2 \right] - \sigma_{\epsilon,jj} &= \left[\frac{1}{n} \sum_{i=1}^{n} \{\ell'_{\tau_j}(e_{ij})\}^2 - \sigma_{\epsilon,jj} \right] + \frac{d+1}{n} \sigma_{\epsilon,jj} (m_j - 1) \\ &\quad + R_1(1 + h_1, \sigma_{\epsilon,jj}^2 + h_2), \end{aligned}$$

where $R_1(\cdot, \cdot)$ is the remainder and satisfies $\lim_{\mathbf{h} \to \mathbf{0}} R_1(1 + h_1, \sigma_{\epsilon,jj}^2 + h_2)/\|\mathbf{h}\| = 0$ for $\mathbf{h} = (h_1, h_2) = c\{(m_j, y_j) - (1, \sigma_{\epsilon,jj})\}$ and $c \in (0, 1)$ since f is twice differentiable [156]. On the event A_Δ defined in Lemma A.2.3.3 with $\mathbb{P}(A_\Delta) \geq 1 - 4n^{-1}$, Lemmas A.2.3.4 and A.2.3.5 imply

$$
\begin{aligned}
&\max_{1\le j\le p}\left|\frac{1}{K_j}\left[\frac{1}{n}\sum_{i=1}^{n}\{\ell'_{\tau_j}(e_{ij})\}^2\right]-\sigma_{\epsilon,jj}\right| \\
\le& C_5\left\{\frac{\log(np)+d}{n}\right\}^{\delta/(2+\delta)}+C_6\frac{d+1}{n}\max\left[\left\{\frac{\log(np)+d}{n}\right\}^{1/2},\frac{\Delta}{h_n}\right] \\
&+R_1\left[1+h_1,\sigma^2_{\epsilon,jj}+h_2\right] \\
\le& C\left\{\frac{\log(np)+d}{n}\right\}^{\delta/(2+\delta)}
\end{aligned}
\tag{A.1.15}
$$

with probability at least $1-8n^{-1}$, where C is a constant depending on A_0. Note that the remainder in (A.1.15) is dominated by other terms as long as m_j-1 and $n^{-1}\sum_{i=1}^{n}\{\ell'_{\tau_j}(e_{ij})\}^2-\sigma^2_{\epsilon,jj}$ are small given $d\ll n$.

By Lemma A.2.3.1, it follows that

$$
\begin{aligned}
\|\mathbf{A}_{jn}^{-1}-\mathbf{\Sigma}_Z^{-1}\| &\le \frac{\|\mathbf{\Sigma}_Z^{-1}\|^2}{1-\|\mathbf{\Sigma}_Z^{-1}\|\|\mathbf{A}_{jn}-\mathbf{\Sigma}_Z\|}\|\mathbf{A}_{jn}-\mathbf{\Sigma}_Z\| \\
&= \|\mathbf{\Sigma}_Z^{-1}\|^2\|\mathbf{A}_{jn}-\mathbf{\Sigma}_Z\|\sum_{k=0}^{\infty}(\|\mathbf{\Sigma}_Z^{-1}\|\|\mathbf{A}_{jn}-\mathbf{\Sigma}_Z\|)^k.
\end{aligned}
\tag{A.1.16}
$$

Note that $\|\mathbf{\Sigma}_Z^{-1}\|\|\mathbf{A}_{jn}-\mathbf{\Sigma}_Z\|\ll 1$ as long as $\|\mathbf{A}_{jn}-\mathbf{\Sigma}_Z\|\ll\lambda_{\min}(\mathbf{\Sigma}_Z)$. Hence, we only need to focus on $\|\mathbf{A}_{jn}-\mathbf{\Sigma}_Z\|$. Denote the sample covariance of $\mathbf{Z}_i$ by $\widehat{\mathbf{\Sigma}}_n=n^{-1}\sum_{i=1}^{n}\mathbf{Z}_i\mathbf{Z}_i^{\mathrm{T}}$. Decompose $\mathbf{A}_{jn}$ as

$$
\mathbf{A}_{jn}=\mathbf{W}_j\widehat{\mathbf{\Sigma}}_n^{-1}\mathbf{W}_j=\widehat{\mathbf{\Sigma}}_n+2(\mathbf{W}_j-\widehat{\mathbf{\Sigma}}_n)+(\mathbf{W}_j-\widehat{\mathbf{\Sigma}}_n)\widehat{\mathbf{\Sigma}}_n^{-1}(\mathbf{W}_j-\widehat{\mathbf{\Sigma}}_n)
$$

so that

$$
\begin{aligned}
\mathbf{A}_{jn}-\widehat{\mathbf{\Sigma}}_n &= 2(\mathbf{W}_j-\widehat{\mathbf{\Sigma}}_n)+(\mathbf{W}_j-\widehat{\mathbf{\Sigma}}_n)\widehat{\mathbf{\Sigma}}_n^{-1}(\mathbf{W}_j-\widehat{\mathbf{\Sigma}}_n) \\
&= 2(\mathbf{W}_j-\widehat{\mathbf{\Sigma}}_n)+(\mathbf{W}_j-\widehat{\mathbf{\Sigma}}_n)(\widehat{\mathbf{\Sigma}}_n^{-1}-\mathbf{\Sigma}_Z^{-1})(\mathbf{W}_j-\widehat{\mathbf{\Sigma}}_n) \\
&\quad+(\mathbf{W}_j-\widehat{\mathbf{\Sigma}}_n)\mathbf{\Sigma}_Z^{-1}(\mathbf{W}_j-\widehat{\mathbf{\Sigma}}_n).
\end{aligned}
$$

Hence, the bound for the operator norm of $\mathbf{A}_{jn}-\widehat{\mathbf{\Sigma}}_n$ is obtained by bounding $\|\mathbf{W}_j-\widehat{\mathbf{\Sigma}}_n\|$ and $\|\widehat{\mathbf{\Sigma}}_n^{-1}-\mathbf{\Sigma}_Z^{-1}\|$. By Lemmas A.2.3.1 and A.2.3.6,

$$\begin{aligned}\|\mathbf{A}_{jn} - \widehat{\mathbf{\Sigma}}_n\| &\leq 2\|\mathbf{W}_j - \widehat{\mathbf{\Sigma}}_n\| + \|\mathbf{W}_j - \widehat{\mathbf{\Sigma}}_n\|^2 \left[\|\widehat{\mathbf{\Sigma}}_n^{-1} - \mathbf{\Sigma}_Z^{-1}\| + \{\lambda_{\min}(\mathbf{\Sigma}_Z)\}^{-1}\right] \\ &\leq C \max\left[\left\{\frac{\log(np)+d}{n}\right\}^{1/2}, \frac{\Delta}{h_n}\right],\end{aligned} \tag{A.1.17}$$

where C is a constant depending on A_0, $\lambda_{\max}(\mathbf{\Sigma}_Z)$, and $v_{j,\delta}$ since the first term on the right hand side of the first inequality dominates the others as long as $n \geq C_3\{\log(np)+d\}$. By the triangle inequality, the concentration of sample covariance matrices [81], and (A.1.17), it holds that

$$\|\mathbf{A}_{jn} - \mathbf{\Sigma}_Z\| \leq \|\mathbf{A}_{jn} - \widehat{\mathbf{\Sigma}}_n\| + \|\widehat{\mathbf{\Sigma}}_n - \mathbf{\Sigma}_Z\| \leq C \max\left[\left\{\frac{\log(np)+d}{n}\right\}^{1/2}, \frac{\Delta}{h_n}\right] \tag{A.1.18}$$

with probability $1 - 4n^{-1}$ as long as $n \geq C_3\{\log(np)+d\}$ where C is a constant depending on $\lambda_{\max}(\mathbf{\Sigma}_Z)$, $v_{j,\delta}$, and A_0.

Putting together (A.1.14)-(A.1.18), for $\delta \in (0, 2]$, it follows that

$$\max_{1\leq j\leq p} \|\widehat{\mathbf{\Sigma}}_j - \mathbf{\Sigma}_j\| \leq C \max\left[\left\{\frac{\log(np)+d}{n}\right\}^{\delta/(2+\delta)}, \frac{\Delta}{h_n}\right]$$

with probability at least $1 - 16n^{-1}$ for some positive constant C depending only on $\lambda_{\min}(\mathbf{\Sigma}_Z)$, $\lambda_{\max}(\mathbf{\Sigma}_Z)$, A_0, and $v_{j,\delta}$.

A.2 Auxiliary results

Recall that the first order derivative of the Huber loss is

$$\ell_\tau'(x) = \begin{cases} x & |x| \leq \tau, \\ \tau \operatorname{sgn}(x) & |x| > \tau \end{cases}$$

and its second order derivative is $\ell_\tau''(x) = \mathbb{I}(|x| < \tau)$ when $|x| \neq \tau$.

A.2.1 Some auxiliary lemmas

We first state a few auxiliary lemmas. Proposition A.2.1 is Proposition A.2 from [157]. It quantifies the difference between the first two moments of $\ell'_\tau(\epsilon_j)$ and ϵ_j given the existence of higher moments of ϵ_j.

Proposition A.2.1. *Let z be a real-valued random variable with $\mathbb{E}(z) = 0$ and $\sigma^2 = \mathbb{E}(z^2) > 0$. Assume that $\mathbb{E}(|z|^\kappa) < \infty$ for some $\kappa > 2$. Then*

$$|\mathbb{E}\ell'_\tau(z)| \leq \min\left\{\frac{\sigma^2}{\tau}, \frac{\mathbb{E}(|z|^\kappa)}{\tau^{\kappa-1}})\right\} \text{ and } |\mathbb{E}\{\ell'_\tau(z)\}^2 - \sigma^2| \leq \frac{2\mathbb{E}(|z|^\kappa)}{(\kappa-2)\tau^{\kappa-2}}.$$

The following corollary from [157] reveals the bias of $s_j^2 = \mathbb{E}\{\ell'_{\tau_j}(\epsilon_{ij})^2\}$ with respect to the true error variance $\sigma_{\epsilon,jj}$. It implies that, with the adaptive robustification parameter τ_j, $s_j^2 \to \sigma_{\epsilon,jj}$ as $n \to \infty$.

Corollary A.2.1. *For $1 \leq j \leq p$ and $v_{j,\delta} = \{\mathbb{E}(|\epsilon_j|^{2+\delta})\}^{1/(2+\delta)} < \infty$, it holds that*

$$\sigma_{\epsilon,jj} - \frac{2}{\delta}\frac{v_{j,\delta}^{2+\delta}}{\tau_j^\delta} \leq s_j^2 \leq \sigma_{\epsilon,jj}.$$

Proof. Applying Proposition A.2.1 with $\kappa = 2 + \delta$ for some $\delta > 0$ yields the first inequality. The second inequality follows $\ell'_{\tau_j}(x)^2 \leq x^2$. □

Next, Proposition A.2.2 implies that the covariance of $\ell'_{\tau_j}(\epsilon_j)$ can be approximated by the covariance of true errors. It is employed to prove the main theorems.

Proposition A.2.2. *Assume $\tau = \min(\tau_j, \tau_k)$ and $v_{jk} = \max\{\mathbb{E}(|\epsilon_j|^{2+\delta}), \mathbb{E}(|\epsilon_k|^{2+\delta})\} < \infty$ for $1 \leq j \neq k \leq p$ and $\delta > 0$. Then*

$$|\operatorname{cov}\{\ell'_{\tau_j}(\epsilon_j), \ell'_{\tau_k}(\epsilon_k)\} - \operatorname{cov}(\epsilon_j, \epsilon_k)| \lesssim \max(\tau^{-\delta} v_{jk}, \tau^{-2-2\delta} v_{jk}^2).$$

Proof. By definition,

$$
\begin{aligned}
\operatorname{cov}(\epsilon_j, \epsilon_k) =& \mathbb{E}\{\epsilon_j\epsilon_k \mathbb{I}(|\epsilon_j| \le \tau_j, |\epsilon_k| \le \tau_k)\} + \mathbb{E}[\epsilon_j\epsilon_k\{\mathbb{I}(|\epsilon_j| > \tau_j) + \mathbb{I}(|\epsilon_k| > \tau_k)\}] \\
& - \mathbb{E}\{\epsilon_j\epsilon_k \mathbb{I}(|\epsilon_j| > \tau_j, |\epsilon_k| > \tau_k)\}, \\
\mathbb{E}\{\ell'_{\tau_j}(\epsilon_j)\} =& \mathbb{E}\{\epsilon_j \mathbb{I}(|\epsilon_j| \le \tau_j)\} + \tau_j \mathbb{E}\{\operatorname{sgn}(\epsilon_j)\mathbb{I}(|\epsilon_j| > \tau_j)\},
\end{aligned}
$$

and

$$
\begin{aligned}
\mathbb{E}\{\ell'_{\tau_j}(\epsilon_j)\ell'_{\tau_k}(\epsilon_k)\} =& \mathbb{E}\{\epsilon_j\epsilon_k \mathbb{I}(|\epsilon_j| \le \tau_j, |\epsilon_k| \le \tau_k)\} + \tau_k \mathbb{E}\{\epsilon_j \operatorname{sgn}(\epsilon_k)\mathbb{I}(|\epsilon_j| \le \tau_j, |\epsilon_k| > \tau_k)\} \\
& + \tau_j \mathbb{E}\{\operatorname{sgn}(\epsilon_j)\epsilon_k \mathbb{I}(|\epsilon_j| > \tau_j, |\epsilon_k| \le \tau_k)\} \\
& + \tau_j\tau_k \mathbb{E}\{\operatorname{sgn}(\epsilon_j\epsilon_k)\mathbb{I}(|\epsilon_j| > \tau_j, |\epsilon_k| > \tau_k)\} \\
=& \operatorname{cov}(\epsilon_j, \epsilon_k) - \mathbb{E}[\epsilon_j\epsilon_k\{\mathbb{I}(|\epsilon_j| > \tau_j) + \mathbb{I}(|\epsilon_k| > \tau_k)\}] \\
& + \tau_k \mathbb{E}\{\epsilon_j \operatorname{sgn}(\epsilon_k)\mathbb{I}(|\epsilon_j| \le \tau_j, |\epsilon_k| > \tau_k)\} \\
& + \tau_j \mathbb{E}\{\operatorname{sgn}(\epsilon_j)\epsilon_k \mathbb{I}(|\epsilon_j| > \tau, |\epsilon_k| \le \tau_k)\} \\
& + \mathbb{E}[\{\epsilon_j\epsilon_k + \tau_j\tau_k \operatorname{sgn}(\epsilon_j\epsilon_k)\}\mathbb{I}(|\epsilon_j| > \tau_j, |\epsilon_k| > \tau_k)].
\end{aligned}
$$

Note that

$$
\begin{aligned}
|\mathbb{E}\{\epsilon_j\epsilon_k \mathbb{I}(|\epsilon_j| > \tau_j\}| &= |\mathbb{E}\{\epsilon_j^{1+\delta}\epsilon_k\epsilon_j^{-\delta}\mathbb{I}(|\epsilon_j| > \tau_j)\}| \\
&\le (\mathbb{E}|\epsilon_j^{1+\delta}\epsilon_k|) \max|\epsilon_j^{-\delta}\mathbb{I}(|\epsilon_j| > \tau_j)| \\
&\le \tau_j^{-\delta}\{\mathbb{E}(|\epsilon_j|^{2+\delta})\}^{(1+\delta)/(2+\delta)}\{\mathbb{E}(|\epsilon_k|^{2+\delta})\}^{1/(2+\delta)} \\
&\le \tau_j^{-\delta} v_{jk} \le \tau^{-\delta} v_{jk}.
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
|\mathbb{E}\{\epsilon_j\epsilon_k \mathbb{I}(|\epsilon_j| > \tau_j, |\epsilon_k| > \tau_k\}| &= |\mathbb{E}\{\epsilon_j^{1+\delta/2}\epsilon_k^{1+\delta/2}\epsilon_j^{-\delta/2}\epsilon_k^{-\delta/2}\mathbb{I}(|\epsilon_j| > \tau_j, |\epsilon_k| > \tau_k)\}| \\
&\le \mathbb{E}(|\epsilon_j^{1+\delta/2}\epsilon_k^{1+\delta/2}|) \max|\epsilon_j^{-\delta/2}\epsilon_k^{-\delta/2}\mathbb{I}(|\epsilon_j| > \tau_j, |\epsilon_k| > \tau_k)| \\
&\le \tau_j^{-\delta/2}\tau_k^{-\delta/2}\{\mathbb{E}(|\epsilon_j|^{2+\delta})\}^{(1+\delta)/(2+\delta)}\{\mathbb{E}(|\epsilon_k|^{2+\delta})\}^{1/(2+\delta)} \\
&\le \tau_j^{-\delta/2}\tau_k^{-\delta/2} v_{jk} \le \tau^{-\delta} v_{jk}.
\end{aligned}
$$

$$
\begin{aligned}
|\mathbb{E}\{\epsilon_j \mathbb{I}(|\epsilon_j| \le \tau_j, |\epsilon_k| > \tau_k)\}| &= |\mathbb{E}\{\epsilon_j \epsilon_k^{1+\delta} \epsilon_k^{-1-\delta} \mathbb{I}(|\epsilon_j| \le \tau_j, |\epsilon_k| > \tau_k\}| \\
&\le \mathbb{E}(|\epsilon_j \epsilon_k^{1+\delta}|) \max |\epsilon_k^{-1-\delta} \mathbb{I}(|\epsilon_k| > \tau_k)| \\
&\le \tau_k^{-1-\delta} v_{jk} \le \tau^{-1-\delta} v_{jk},
\end{aligned}
$$

and

$$
\begin{aligned}
|\mathbb{E}\{\mathbb{I}(|\epsilon_j| > \tau_j, |\epsilon_k| > \tau_k\}| &= |\mathbb{E}\{\epsilon_j^{1+\delta/2} \epsilon_k^{1+\delta/2} \epsilon_j^{-1-\delta/2} \epsilon_k^{-1-\delta/2} \mathbb{I}(|\epsilon_j| > \tau_j, |\epsilon_k| > \tau_k)\}| \\
&\le \mathbb{E}(|\epsilon_j^{1+\delta/2} \epsilon_k^{1+\delta/2}|) \max |\epsilon_j^{-1-\delta/2} \epsilon_k^{-1-\delta/2} \mathbb{I}(|\epsilon_j| > \tau_j, |\epsilon_k| > \tau_k)| \\
&\le \tau_j^{-1-\delta/2} \tau_k^{-1-\delta/2} v_{jk} \le \tau^{-2-\delta} v_{jk}.
\end{aligned}
$$

Therefore, we have

$$
|\mathbb{E}\{\ell'_{\tau_j}(\epsilon_j) \ell'_{\tau_k}(\epsilon_k)\} - \operatorname{cov}(\epsilon_j, \epsilon_k)| \lesssim \tau^{-\delta} v_{jk}
$$

as $\max_{1 \le j \le p} v_{j,\delta} \le C_\epsilon$. Hence, it yields

$$
\begin{aligned}
&|\operatorname{cov}\{\ell'_{\tau_j}(\epsilon_j), \ell'_{\tau_k}(\epsilon_k)\} - \operatorname{cov}(\epsilon_j, \epsilon_k)| \\
&\le |\mathbb{E}\{\ell'_{\tau_j}(\epsilon_j) \ell'_{\tau_k}(\epsilon_k)\} - \operatorname{cov}(\epsilon_j, \epsilon_k)| + |\mathbb{E}\{\ell'_{\tau_j}(\epsilon_j)\} \mathbb{E}\{\ell'_{\tau_k}(\epsilon_k)\}| \\
&\lesssim \tau^{-\delta} v_{jk} + \min\left\{ \frac{\sigma_{\epsilon,jj}}{\tau_j}, \frac{\mathbb{E}(|\epsilon_j|^{2+\delta})}{\tau_j^{1+\delta}} \right\} \min\left\{ \frac{\sigma_{\epsilon,kk}}{\tau_k}, \frac{\mathbb{E}(|\epsilon_k|^{2+\delta})}{\tau_k^{1+\delta}} \right\} \\
&\lesssim \max(\tau^{-\delta} v_{jk}, \tau^{-2-2\delta} v_{jk}^2)
\end{aligned}
$$

by Proposition A.2.1. □

The following result on the non-asymptotic bound for the adaptive Huber regression estimator is borrowed from Theorem 7 in [12]. It provides an exponential-type concentration inequalities for $\widehat{\boldsymbol{\theta}}_j$'s with adaptive robustification parameter τ_j, and also gives a non-asymptotic Bahadur representation under the finite moment condition on the errors.

Lemma A.2.1.1. *Under Condition 2, for any $t > 0$, $\tau_{0j} \ge v_{j,\delta} := (\mathbb{E}|\epsilon|^{2+\delta})^{1/(2+\delta)}$, $j = 1, \ldots, p$, $\widehat{\boldsymbol{\theta}}_j$ with $\tau_j = \tau_{0j}\{n(d+t)^{-1}\}^{1/(2+\delta)}$ satisfies*

$$\mathbb{P}[\|\boldsymbol{\Sigma}_Z^{1/2}(\widehat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j)\| \geq C_2\tau_{0j}\{n^{-1}(d+t)\}^{1/2}] \leq 2e^{-t}$$

and

$$\mathbb{P}[\|\boldsymbol{\Sigma}_Z^{1/2}(\widehat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j) - \frac{1}{n}\sum_{i=1}^{n}\{\ell'_\tau(\epsilon_i)\boldsymbol{\Sigma}_Z^{-1/2}\mathbf{Z}_i\}\| \geq C_3\tau_{0j}n^{-1}(d+t)] \leq 3e^{-t}$$

as long as $n \geq C_4(d+t)$, where C_1–C_4 are positive constants depending only on A_0 from Condition 2.

A.2.2 Technical results for proving Theorem 2.3.1

From Lemma A.2.1.1, we expect the following approximation of our testing statistics.

Corollary A.2.2. *For $\mathbf{T}_j^\circ$, $\mathbf{S}_j$, and $\mathbf{R}_j$ in* (A.1.2) *and* (A.1.3)*, respectively, it holds*

$$\|\mathbf{T}_j^\circ - \sigma_{\epsilon,jj}^{-1/2}(\mathbf{S}_j + \mathbf{R}_j)\| \leq C_2\tau_{0j}\frac{d+t}{(n\sigma_{\epsilon,jj})^{1/2}}$$

with probability at least $1 - 2\exp(-t)$ under the random design.

Proof. By Lemma A.2.1.1,

$$\begin{aligned}
&\|\mathbf{T}_j^\circ - \sigma_{\epsilon,jj}^{-1/2}(\mathbf{S}_j + \mathbf{R}_j)\| \\
&\leq (n\sigma_{\epsilon,jj}{}^{-1})^{1/2}\left\|(\mathbf{C}\boldsymbol{\Sigma}_Z^{-1}\mathbf{C}^{\mathrm{T}})^{-1/2}\mathbf{C}\boldsymbol{\Sigma}_Z^{-1/2}\left\{\boldsymbol{\Sigma}_Z^{1/2}(\widehat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j) - \frac{1}{n}\sum_{i=1}^{n}\ell'_\tau(\epsilon_{ij})\boldsymbol{\Sigma}_Z^{-1/2}\mathbf{Z}_i\right\}\right\| \qquad \text{(A.2.1)} \\
&\leq C_2\tau_{0j}\frac{d+t}{(n\sigma_{\epsilon,jj})^{1/2}}
\end{aligned}$$

with probability at least $1 - 3\exp(-t)$. □

The following results show that the distribution of the Bahadur representation in (A.2.1) is close to $N(\mathbf{0}, \mathbf{I})$. We decompose $|\,\mathbb{P}(\|\sigma_{\epsilon,jj}^{-1/2}\mathbf{S}_j\| \geq x) - \mathbb{P}(\|\mathbf{G}\| \geq x)|$ into two parts. Lemma A.2.2.1 quantifies the difference between the cumulative distribution functions of $\|\sigma_{\epsilon,jj}^{-1/2}s_j\mathbf{G}\|$ and $\|\mathbf{G}\|$, and Lemma A.2.2.2 characterizes the difference between the cumulative distribution functions of $\|\sigma_{\epsilon,jj}^{-1/2}s_j\mathbf{G}\|$ and $\|\sigma_{\epsilon,jj}^{-1/2}\mathbf{S}_j\|$.

Lemma A.2.2.1. *Let* $\mathbf{G} \sim N(\mathbf{0}, \mathbf{I}) \in \mathbb{R}^q$. *Let* $\tau_j = \tau_{0j}\{n(d+t)^{-1}\}^{1/(2+\delta)}$ *for some* $\delta > 0$ *where* $\tau_{0j} \geq v_{j,\delta}$. *Then, it holds that*

$$\sup_{x \in \mathbb{R}^+} \left| \mathbb{P}(\|\sigma_{\epsilon,jj}^{-1/2} s_j \mathbf{G}\| \geq x) - \mathbb{P}(\|\mathbf{G}\| \geq x) \right| \leq q^{1/2} \frac{2}{\delta \sigma_{\epsilon,jj}} \frac{v_{j,\delta}^{2+\delta}}{\tau_{0j}^{\delta}} \left(\frac{d+t}{n} \right)^{\delta/(2+\delta)}.$$

Proof. It holds

$$\begin{aligned} \|\sigma_{\epsilon,jj}^{-1} s_j^2 \mathbf{I} - \mathbf{I}\| = |\sigma_{\epsilon,jj}^{-1} s_j^2 - 1| &\leq \frac{2}{\delta \sigma_{\epsilon,jj}} \frac{v_{j,\delta}^{2+\delta}}{\tau_j^{\delta}} \\ \mathrm{tr}\{(\sigma_{\epsilon,jj}^{-1/2} s_j \mathbf{I} - \mathbf{I})^2\} &\leq q \left(\frac{2}{\delta \sigma_{\epsilon,jj}} \frac{v_{j,\delta}^{2+\delta}}{\tau_j^{\delta}} \right)^2. \end{aligned} \tag{A.2.2}$$

With $\tau_j = \tau_{0j}\{n(d+t)^{-1}\}^{1/(2+\delta)}$, (A.2.2) satisfies the conditions of Lemma A.7 in the supplement of [158] whenever $n \geq C_3(d+t)$. Combining Corollary A.2.1 with Lemma A.7 in the supplement of [158], we get the desired result. □

Lemma A.2.2.2. *Let* $\mathbf{G} \sim N\mathbf{0}, \mathbf{I}) \in \mathbb{R}^q$.

$$\sup_{x \in \mathbb{R}^+} \left| \mathbb{P}(\|\sigma_{\epsilon,jj}^{-1/2} \mathbf{S}_j\| \geq x) - \mathbb{P}(\|\sigma_{\epsilon,jj}^{-1/2} s_j \mathbf{G}\| \geq x) \right| \precsim n^{-1/2} q^{7/4}. \tag{A.2.3}$$

Proof. Denote $\mathcal{C}$ the class of convex subsets of $\mathbb{R}^q$. Recall that $\mathrm{cov}(\mathbf{S}_j) = \sigma_{\epsilon,jj}^{-1} s_j^2 \mathbf{I}$ for $\mathbf{S}_j$ in (A.1.3). By Theorem 1.1 in [159],

$$\begin{aligned} \sup_{A \in \mathcal{C}} |\, \mathbb{P}(\sigma_{\epsilon,jj}^{-1/2} \mathbf{S}_j \in A) - \mathbb{P}(\sigma_{\epsilon,jj}^{-1/2} s_j \mathbf{G} \in A)| &\precsim q^{1/4} \sum_{i=1}^{n} \mathbb{E}\|n^{-1/2} \ell'_{\tau_j}(\epsilon_{ij}) \mathbf{A}_j^{-1/2} \mathbf{C} \mathbf{\Sigma}_Z^{-1/2} \widetilde{\mathbf{Z}}_i\|^3 \\ &= \frac{q^{1/4}}{n^{1/2}} \frac{\mathbb{E}\{|\ell'_{\tau_j}(\epsilon_{ij})|^3\}}{\sigma_{\epsilon,jj}} \mathbb{E}\|\mathbf{A}_j^{-1/2} \mathbf{C} \mathbf{\Sigma}_Z^{-1/2} \widetilde{\mathbf{Z}}_i\|^3 \\ &\precsim \frac{q^{7/4}}{n^{1/2}}. \end{aligned}$$

Take $A = \{\mathbf{v} \in \mathbb{R}^{d+1} : \|\mathbf{v}\| \leq x, x > 0\}$, we obtain (A.2.3). □

The following coupling result compares $\mathbb{P}(\|\sigma_{\epsilon,jj}^{-1/2} \mathbf{S}_j\| > x, \|\sigma_{\epsilon,kk}^{-1/2} \mathbf{S}_k\| > y)$ and its Gaussian counterpart for each (j,k) pair.

Lemma A.2.2.3. *Assume Condition 1 holds. Let* $\mathbf{G} = (\mathbf{G}_1, \mathbf{G}_2) \in \mathbb{R}^{2q}$ *be a Gaussian vector with* $\mathbb{E}(\mathbf{G}) = \mathbf{0}$ *and* $\mathrm{cov}(\mathbf{G}) = \mathrm{cov}(\mathbf{S})$*, which is given in Section A.1. It satisfies that*

$$\sup_{x,y\in\mathbb{R}} |\,\mathbb{P}(\|s_j^{-1}\mathbf{S}_j\| > x, \|s_k^{-1}\mathbf{S}_k\| > y) - \mathbb{P}(\|\mathbf{G}_1\| > x, \|\mathbf{G}_2\| > y)| \precsim n^{-1/2}q^{7/4}.$$

Proof. Notice that

$$\begin{aligned}\mathbb{P}(\|s_j^{-1}\mathbf{S}_j\| > x, \|s_k^{-1}\mathbf{S}_k\| > y) = 1 - \mathbb{P}(\|s_j^{-1}\mathbf{S}_j\| \le x) - \mathbb{P}(\|s_k^{-1}\mathbf{S}_k\| \le y)\\ + \mathbb{P}(\|s_j^{-1}\mathbf{S}_j\| \le x, \|s_k^{-1}\mathbf{S}_k\| \le y)\end{aligned}$$

and

$$\begin{aligned}\mathbb{P}(\|\mathbf{G}_1\| > x, \|\mathbf{G}_2\| > y) = 1 - \mathbb{P}(\|\mathbf{G}_1\| \le x) - \mathbb{P}(\|\mathbf{G}_2\| \le y)\\ + \mathbb{P}(\|\mathbf{G}_1\| \le x, \|\mathbf{G}_2\| \le y).\end{aligned}$$

Take $A(x,y) = \{\mathbf{v} = (\mathbf{v}_1, \mathbf{v}_2)^{\mathrm{T}} \in \mathbb{R}^{2q}, \mathbf{v}_1, \mathbf{v}_2 \in \mathbb{R}^q : \|\mathbf{v}_1\| \le x \text{ and } \|\mathbf{v}_2\| \le y \text{ and } x, y \in \mathbb{R}^+ \bigcup \{\infty\}\}$ in Theorem 1.1 in [159], we have

$$\sup_{x,y} |\,\mathbb{P}(\|s_j^{-1}\mathbf{S}_j\| > x, \|s_k^{-1}\mathbf{S}_k\| > y) - \mathbb{P}(\|\mathbf{G}_1\| > x, \|\mathbf{G}_2\| > y)| \precsim \frac{q^{7/4}}{n^{1/2}},$$

which is the desired result. □

Lemma A.2.2.4 below provides a coupling between multivariate normal distributions. We provide two versions of proof. One uses the properties of bivariate chi-square distribution, and the other uses the total variation distance between two multivariate normal distributions.

Lemma A.2.2.4. *Let* $\mathbf{G} = (\mathbf{G}_1, \mathbf{G}_2) \in \mathbb{R}^{2q}$ *be a Gaussian vector with* $\mathbb{E}(\mathbf{G}) = \mathbf{0}$*,* $\mathrm{cov}(\mathbf{G}_i) = \mathbf{I}$ *for* $i = 1, 2$*, and* $\mathrm{corr}(\mathbf{G}_1, \mathbf{G}_2) = r\mathbf{I}$ *where* $|r| \le k_0 < 1$*. Let* $\mathbf{Z}_1, \mathbf{Z}_2 \sim N(\mathbf{0}, \mathbf{I})$ *be independent and identically distributed* q*-dimensional standard normal vectors. Then*

$$\left|\mathbb{P}(\|\mathbf{G}_1\| > z, \|\mathbf{G}_2\| > z) - \mathbb{P}(\|\mathbf{Z}_1\| > z)^2\right| \le C_q|r|$$

for some constant $C_q > 0$ only depending on q.

Proof. For $q = 1$, it holds

$$\begin{aligned}
&\mathbb{P}(|G_1| > x, |G_2| > x) - \mathbb{P}(|Z_1| > x)\,\mathbb{P}(|Z_2| > x) \\
=&\{\mathbb{P}(G_1 < -x, G_2 < -x) - \mathbb{P}(Z_1 < -x)\,\mathbb{P}(Z_2 < -x)\} \\
&- \{\mathbb{P}(G_1 < -x, G_2 < x) - \mathbb{P}(Z_1 < -x)\,\mathbb{P}(Z_2 < x)\} \\
&- \{\mathbb{P}(G_1 < x, G_2 < -x) - \mathbb{P}(Z_1 < x)\,\mathbb{P}(Z_2 < -x)\} \\
&+ \{\mathbb{P}(G_1 < x, G_2 < x) - \mathbb{P}(Z_1 < x)\,\mathbb{P}(Z_2 < x)\}\,,
\end{aligned}$$

so that $|\,\mathbb{P}(|G_1| > x, |G_2| > x) - \mathbb{P}(|Z_1| > x)\,\mathbb{P}(|Z_2| > x)| \le |r|$ follows Corollary 2.1 from [160]. Set $C_1 = 1$.

For $q \ge 2$, notice that $(\|\mathbf{G}_1\|^2, \|\mathbf{G}_2\|^2)$ is a bivariate chi-squared distribution with correlation r^2, and $\|\mathbf{Z}_1\|^2$ and $\|\mathbf{Z}_2\|^2$ are independent χ^2_q distributions. Let $V_1 = \|\mathbf{G}_1\|^2/2 \sim \text{Gamma}(q/2, 1)$ and $V_2 = \|\mathbf{G}_2\|^2/2 \sim \text{Gamma}(q/2, 1)$. Given $|r| < 1$, the joint distribution of V_1 and V_2 is the Wicksell-Kibble's bivariate Gamma distribution with $\text{corr}(V_1, V_2) = r^2$, whose joint probability density function can be represented by an infinite series [161, 162] that

$$f(v_1, v_2; q/2) := f(v_1; q/2) f(v_2; q/2) \sum_{m=0}^{\infty} r^{2m} \frac{m!\Gamma(q/2)}{\Gamma(m + q/2)} L_m^{(q/2-1)}(v_1) L_m^{(q/2-1)}(v_2). \qquad \text{(A.2.4)}$$

Here,

$$L_m^{(q/2-1)}(v) = \frac{(d/dv)^m\{v^m f(v; q/2)\}}{m! f(v; q/2)}$$

is the generalized Laguerre polynomial of order m as defined by Rodrigue's formula, and $f(v; q/2) = v^{q/2-1} \exp(-v)\{\Gamma(q/2)\}^{-1}$ is the probability density function of Gamma$(q/2, 1)$ distribution.

Rewrite (A.2.4) as

$$
\begin{aligned}
&f(v_1, v_2; q/2) - f(v_1; q/2) f(v_2; q/2) \\
&= f(v_1; q/2) f(v_2; q/2) \sum_{m=1}^{\infty} r^{2m} \frac{m! \Gamma(q/2)}{\Gamma(m + q/2)} L_m^{(q/2-1)}(v_1) L_m^{(q/2-1)}(v_2).
\end{aligned}
$$

By Lebesgue's dominated convergence theorem, integrate the above equality from $z^2/2$ to ∞ to yield

$$
\begin{aligned}
&\mathbb{P}(V_1 > z^2/2, V_2 > z^2/2) - \mathbb{P}(V_1 > z^2/2)\, \mathbb{P}(V_2 > z^2/2) \\
&= \int_{z^2/2}^{\infty} \int_{z^2/2}^{\infty} f(v_1; q/2) f(v_2; q/2) \sum_{m=1}^{\infty} r^{2m} \frac{m! \Gamma(q/2)}{\Gamma(m + q/2)} L_m^{(q/2-1)}(v_1) L_m^{(q/2-1)}(v_2) dv_1 dv_2 \qquad \text{(A.2.5)} \\
&= \sum_{m=1}^{\infty} r^{2m} \frac{m! \Gamma(q/2)}{\Gamma(m + q/2)} \left\{ \int_{z^2/2}^{\infty} L_m^{(q/2-1)}(v) f(v; q/2) dv \right\}^2,
\end{aligned}
$$

which is a special case of Theorem 2 in [163] when $\nu = q/2$, $\rho_1^2 = \cdots = \rho_m^2 := r^2$, $c_1 = z^2/2$, and $c_2 = +\infty$. Let $C_m = r^{2m} m! \Gamma(q/2) \{\Gamma(m+q/2)\}^{-1}$ and $T_m(z^2/2) = \int_{z^2/2}^{\infty} L_m^{(q/2-1)}(v) f(v; q/2) dv$ for positive integer m. It follows that

$$
\begin{aligned}
T_m(z^2/2) \; &= \int_{z^2/2}^{\infty} \frac{d}{dv} \frac{d^{m-1}}{dv^{m-1}} \left\{ \frac{v^m f(v; q/2)}{m!} \right\} dv \\
&= \frac{q}{2m} \int_{z^2/2}^{\infty} \frac{d}{dv} \frac{d^{m-1}}{dv^{m-1}} \left\{ \frac{v^{m-1} f(v; q/2+1)}{(m-1)!} \right\} dv \\
&= \frac{q}{2m} \int_{z^2/2}^{\infty} \frac{d}{dv} L_{m-1}^{(q/2)}(v) f(v; q/2+1) dv \\
&= -\frac{q}{2m} L_{m-1}^{(q/2)}(z^2/2) f(z^2/2; q/2+1) \\
&= -\frac{q}{2m} L_{m-1}^{(q/2)}(z^2/2) \frac{(z^2/2)^{q/2}}{\Gamma(q/2+1)} \exp(-z^2/2) \\
&= -\frac{1}{m} \left\{ L_{m-1}^{(q/2)}(z^2/2) \frac{(z^2/2)^{q/2}}{\Gamma(q/2)} \exp(-z^2/2) \right\}.
\end{aligned}
$$

Therefore, in (A.2.5)

$$
\begin{aligned}
C_m\{T_m(z^2/2)\}^2 &= r^{2m}\frac{m!\Gamma(q/2)}{\Gamma(m+q/2)}\frac{1}{m^2}\left\{L_{m-1}^{(q/2)}(z^2/2)\frac{(z^2/2)^{q/2}}{\Gamma(q/2)}\exp(-z^2/2)\right\}^2 \\
&= r^{2m}\left[\frac{1}{m^2}\frac{m!}{\Gamma(m+q/2)}\frac{(z^2/2)^q}{\Gamma(q/2)}\exp(-z^2)\left\{L_{m-1}^{(q/2)}(z^2/2)\right\}^2\right] \\
&= r^{2m}\left[\frac{1}{m^2}\frac{m!}{\Gamma(m+q/2)}f(z^2/2;q/2)\left\{\left(\frac{z^2}{2}\right)^{q/2+1}e^{-z^2/2}\right\}\left\{L_{m-1}^{(q/2)}(z^2/2)\right\}^2\right].
\end{aligned}
\tag{A.2.6}
$$

By Theorem 1 in [164] for the orthonormal Laguerre polynomials,

$$
\frac{\Gamma(m)}{\Gamma(m+q/2)}\left(\frac{z^2}{2}\right)^{q/2+1}e^{-z^2/2}\left\{L_{m-1}^{(q/2)}(z^2/2)\right\}^2 \le 6(m-1)^{1/6}(m+q/2)^{1/2} \tag{A.2.7}
$$

for all $z > 0$ and positive integer m. Putting together (A.2.6), (A.2.7), and $f(z^2/2;q/2) \le 1$ for $q \ge 2$,

$$
\begin{aligned}
C_m\{T_m(z^2/2)\}^2 &\le r^{2m}\frac{1}{m^2}\frac{m!}{\Gamma(m+q/2)}\frac{\Gamma(m+q/2)}{\Gamma(m)}6(m-1)^{1/6}(m+q/2)^{1/2} \\
&= 6r^{2m}\frac{(m-1)^{1/6}(m+q/2)^{1/2}}{m} \\
&\le 3r^{2m}(2+q/2)^{1/2} =: C_q r^{2m},
\end{aligned}
$$

where the inequality in the last line is due to the fact $(m-1)^{1/6}(m+q/2)^{1/2}m^{-1}$ is maximized at $m = 2$ for any q. Therefore,

$$
\sum_{m=1}^{\infty} C_m\{T_m(z^2/2)\}^2 \le \sum_{m=1}^{\infty} C_q r^{2m} = C_q\frac{r^2}{1-r^2} \le C_q\frac{k_0}{1-k_0^2}|r|. \tag{A.2.8}
$$

Combining (A.2.8) with (A.2.5), we prove the desired result with $C_{q,k_0} = C_q k_0/(1-k_0^2)$. In fact $C_{q,k_0} \precsim q^{1/2}$. This completes the first version of proof with bivariate chi-square distribution.

We also provide a simpler proof using the inequality of the total variation distance between two multivariate normal distribution. For $q \ge 2$, let $\mathbf{\Sigma}_1 = \mathbf{I}$ and $\mathbf{\Sigma}_2$ be the covariance matrices of $(\mathbf{Z}_1, \mathbf{Z}_2)$ and $(\mathbf{G}_1, \mathbf{G}_2)$, respectively. By the definition of the total variation distance and Theorem 1.1 in [165], it follows that

$$
\begin{aligned}
&\left|\mathbb{P}(\|\mathbf{G}_1\| > z, \|\mathbf{G}_2\| > z) - \mathbb{P}(\|\mathbf{Z}_1\| > z)^2\right| \\
&\leq \sup_{A:\text{measurable sets}} |\mathbb{P}\{(\mathbf{G}_1, \mathbf{G}_2) \in A\} - \mathbb{P}\{(\mathbf{Z}_1, \mathbf{Z}_2) \in A\}| \\
&\leq \frac{3}{2}\|\boldsymbol{\Sigma}_1^{-1/2}\boldsymbol{\Sigma}_2\boldsymbol{\Sigma}_1^{-1/2} - \mathbf{I}\|_F \\
&= \frac{3}{2}\left\|\begin{bmatrix} \mathbf{0} & r\mathbf{I} \\ r\mathbf{I} & \mathbf{0} \end{bmatrix}\right\|_F = \frac{3}{2}(2q)^{1/2}|r| =: C_q|r| \precsim q^{1/2}|r|
\end{aligned}
$$

where $\|\cdot\|_F$ is the Frobenius norm. □

The following lemma characterizes the distribution of quadratic form of the multivariate normal distribution, which appears in many textbooks (for example, [166]).

Lemma A.2.2.5. *Consider* $\mathbf{X} \sim N_k(\mathbf{0}, \boldsymbol{\Sigma})$. *Let* $\lambda_1, \ldots, \lambda_k$ *be the eigenvalues of* $\boldsymbol{\Sigma}$. *Then* $\|\mathbf{X}\|^2$ *is distributed as* $\sum_{i=1}^k \lambda_i Z_i^2$ *for independent and identically distributed random variables* $Z_i \sim N(0,1)$.

Proof. There exists an orthogonal matrix $\mathbf{A}$ such that $\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^{\mathrm{T}} = \text{diag}(\lambda_1, \ldots, \lambda_k)$. Then $\mathbf{A}\mathbf{X}$ is $N_k(\mathbf{0}, \text{diag}(\lambda_1, \ldots, \lambda_k))$. Hence, $\|\mathbf{A}\mathbf{X}\|^2 = \|\mathbf{X}\|^2$ implies that $\|\mathbf{X}\|^2$ has the same distribution as $\sum_{i=1}^k \lambda_i Z_i^2$. □

A.2.3 Technical results for Section 2.3.2

Lemma A.2.3.1 implies that the operator norm of the difference between two inverse matrices is bounded by a non-decreasing function ($f(x) = x/(1-x)$) of the operator norm of the difference between two matrices. It appears in many linear algebra textbooks (for example, see [167], Chapter 5.8), and is used in the proof of Proposition 2.3.1.

Lemma A.2.3.1. *For* $d \times d$ *invertible matrices* $\mathbf{A}$ *and* $\mathbf{B} = \mathbf{A} + \Delta\mathbf{A}$ *such that* $\rho(\mathbf{A}^{-1}\Delta\mathbf{A}) < 1$ *where* $\rho(\cdot)$ *is the spectral radius, it follows that*

$$
\|\mathbf{A}^{-1} - \mathbf{B}^{-1}\| \leq \frac{\|\mathbf{A}^{-1}\Delta\mathbf{A}\|}{1 - \|\mathbf{A}^{-1}\Delta\mathbf{A}\|}\|\mathbf{A}^{-1}\| \leq \frac{\|\mathbf{A}^{-1}\|^2}{1 - \|\mathbf{A}^{-1}\Delta\mathbf{A}\|}\|\Delta\mathbf{A}\|. \tag{A.2.9}
$$

Moreover, if $\|\mathbf{A}^{-1}\|\|\Delta\mathbf{A}\| < 1$*, it follows that*

$$\|\mathbf{A}^{-1} - \mathbf{B}^{-1}\| \leq \frac{\|\mathbf{A}^{-1}\|}{1 - \|\mathbf{A}^{-1}\|\|\Delta\mathbf{A}\|}\|\mathbf{A}^{-1}\|\|\Delta\mathbf{A}\|. \tag{A.2.10}$$

Proof. Notice that $\mathbf{A}^{-1} - \mathbf{B}^{-1} = \mathbf{A}^{-1}(\mathbf{B} - \mathbf{A})\mathbf{B}^{-1} = \mathbf{A}^{-1}\Delta\mathbf{A}\mathbf{B}^{-1}$, so

$$\|\mathbf{A}^{-1} - \mathbf{B}^{-1}\| = \|\mathbf{A}^{-1}\Delta\mathbf{A}\mathbf{B}^{-1}\| \leq \|\mathbf{A}^{-1}\Delta\mathbf{A}\|\|\mathbf{B}^{-1}\|.$$

Since $\|\mathbf{B}^{-1}\| \leq \|\mathbf{A}^{-1}\| + \|\mathbf{A}^{-1}\Delta\mathbf{A}\|\|\mathbf{B}^{-1}\|$, it holds

$$\|\mathbf{B}^{-1}\| \leq \frac{\|\mathbf{A}^{-1}\|}{1 - \|\mathbf{A}^{-1}\Delta\mathbf{A}\|}.$$

Hence, we prove the first part of (A.2.9). The second inequality of (A.2.9) is straightforward from the first part since $\|\mathbf{A}^{-1}\Delta\mathbf{A}\| \leq \|\mathbf{A}^{-1}\|\|\Delta\mathbf{A}\|$. Inequality (A.2.10) is a direct result from (A.2.9) and $\|\mathbf{A}^{-1}\Delta\mathbf{A}\| \leq \|\mathbf{A}^{-1}\|\|\Delta\mathbf{A}\|$. □

Lemma A.2.3.2. *Under the conditions in Proposition 2.3.1,*

$$\max_{1\leq j\leq p} \|(\mathbf{C}\widetilde{\mathbf{\Sigma}}_j\mathbf{C}^{\mathrm{T}})^{-1} - (\mathbf{C}\mathbf{\Sigma}_j\mathbf{C}^{\mathrm{T}})^{-1}\| \precsim \max_{1\leq j\leq p} \|\widetilde{\mathbf{\Sigma}}_j - \mathbf{\Sigma}_j\|.$$

Proof. Without loss of generality, as $\mathrm{rank}(\mathbf{C}) = q$, we assume $\mathbf{C}\mathbf{C}^{\mathrm{T}} = \mathbf{I}_{q\times q}$ by considering $(\mathbf{C}\mathbf{C}^{\mathrm{T}})^{-1/2}\mathbf{C}$ in (2.2.2). It follows that

$$
\begin{aligned}
\|(\mathbf{C}\boldsymbol{\Sigma}\mathbf{C}^{\mathrm{T}})^{-1}\|\|(\mathbf{C}\widetilde{\boldsymbol{\Sigma}}_j\mathbf{C}^{\mathrm{T}} - \mathbf{C}\boldsymbol{\Sigma}_j\mathbf{C}^{\mathrm{T}})\| &\leq \|(\mathbf{C}\boldsymbol{\Sigma}_j\mathbf{C}^{\mathrm{T}})^{-1}\|\|\widetilde{\boldsymbol{\Sigma}}_j - \boldsymbol{\Sigma}_j\|\|\mathbf{C}\mathbf{C}^{\mathrm{T}}\| \\
&\leq \frac{\|\widetilde{\boldsymbol{\Sigma}}_j - \boldsymbol{\Sigma}_j\|}{\lambda_{\min}(\mathbf{C}\boldsymbol{\Sigma}_j\mathbf{C}^{\mathrm{T}})} \cdot 1 \\
&\leq \frac{\|\widetilde{\boldsymbol{\Sigma}}_j - \boldsymbol{\Sigma}_j\|}{\lambda_{\min}(\boldsymbol{\Sigma}_j)} \\
&\leq \frac{\|\widetilde{\boldsymbol{\Sigma}}_j - \boldsymbol{\Sigma}_j\|}{\sigma_{\epsilon,jj}}\|\boldsymbol{\Sigma}_Z\| \\
&\leq \frac{\|\widetilde{\boldsymbol{\Sigma}}_j - \boldsymbol{\Sigma}_j\|}{c_\epsilon^2}\|\boldsymbol{\Sigma}_Z\|_F \\
&\precsim \|\widetilde{\boldsymbol{\Sigma}}_j - \boldsymbol{\Sigma}_j\|d^{1/2},
\end{aligned}
$$

where $\|\cdot\|_F$ is the Frobenius norm. The last line in the above inequality is less than 1 as long as $\|\widetilde{\boldsymbol{\Sigma}}_j - \boldsymbol{\Sigma}_j\| \precsim d^{-1/2}$, which holds for sufficiently large n and p by the assumptions in Proposition 2.3.1. Therefore, (A.2.10) in Lemma A.2.3.1 holds.

By the assumptions in Proposition 2.3.1,

$$
\|\mathbf{C}\widetilde{\boldsymbol{\Sigma}}_j\mathbf{C}^{\mathrm{T}} - \mathbf{C}\boldsymbol{\Sigma}_j\mathbf{C}^{\mathrm{T}}\| \leq \|\widetilde{\boldsymbol{\Sigma}}_j - \boldsymbol{\Sigma}_j\|. \tag{A.2.11}
$$

Set $\mathbf{A} = \mathbf{C}\boldsymbol{\Sigma}_j\mathbf{C}^{\mathrm{T}}$, $\mathbf{B} = \mathbf{C}\widetilde{\boldsymbol{\Sigma}}_j\mathbf{C}^{\mathrm{T}}$, and $\Delta\mathbf{A} = \mathbf{C}\widetilde{\boldsymbol{\Sigma}}_j\mathbf{C}^{\mathrm{T}} - \mathbf{C}\boldsymbol{\Sigma}_j\mathbf{C}^{\mathrm{T}}$. Note that $\|(\mathbf{C}\boldsymbol{\Sigma}_j\mathbf{C}^{\mathrm{T}})^{-1}\| = 1/\lambda_{\min}(\mathbf{C}\boldsymbol{\Sigma}_j\mathbf{C}^{\mathrm{T}})$ is bounded. Putting Lemma A.2.3.1 and (A.2.11) together,

$$
\begin{aligned}
&\|(\mathbf{C}\widetilde{\boldsymbol{\Sigma}}_j\mathbf{C}^{\mathrm{T}})^{-1} - (\mathbf{C}\boldsymbol{\Sigma}_j\mathbf{C}^{\mathrm{T}})^{-1}\| \\
&\leq \frac{\|(\mathbf{C}\boldsymbol{\Sigma}_j\mathbf{C}^{\mathrm{T}})^{-1}\|^2\|\mathbf{C}\widetilde{\boldsymbol{\Sigma}}_j\mathbf{C}^{\mathrm{T}} - \mathbf{C}\boldsymbol{\Sigma}_j\mathbf{C}^{\mathrm{T}}\|}{1 - \|(\mathbf{C}\boldsymbol{\Sigma}_j\mathbf{C}^{\mathrm{T}})^{-1}(\mathbf{C}\widetilde{\boldsymbol{\Sigma}}_j\mathbf{C}^{\mathrm{T}} - \mathbf{C}\boldsymbol{\Sigma}_j\mathbf{C}^{\mathrm{T}})\|} \\
&\leq \frac{\|(\mathbf{C}\boldsymbol{\Sigma}_j\mathbf{C}^{\mathrm{T}})^{-1}\|^2}{1 - \|(\mathbf{C}\boldsymbol{\Sigma}_j\mathbf{C}^{\mathrm{T}})^{-1}\|\|(\mathbf{C}\widetilde{\boldsymbol{\Sigma}}_j\mathbf{C}^{\mathrm{T}} - \mathbf{C}\boldsymbol{\Sigma}_j\mathbf{C}^{\mathrm{T}})\|}\|\mathbf{C}\widetilde{\boldsymbol{\Sigma}}_j\mathbf{C}^{\mathrm{T}} - \mathbf{C}\boldsymbol{\Sigma}_j\mathbf{C}^{\mathrm{T}}\| \\
&\leq \frac{\|(\mathbf{C}\boldsymbol{\Sigma}_j\mathbf{C}^{\mathrm{T}})^{-1}\|^2}{1 - \|(\mathbf{C}\boldsymbol{\Sigma}_j\mathbf{C}^{\mathrm{T}})^{-1}\|\|\widetilde{\boldsymbol{\Sigma}}_j - \boldsymbol{\Sigma}_j\|}\|\widetilde{\boldsymbol{\Sigma}}_j - \boldsymbol{\Sigma}_j\|,
\end{aligned}
$$

so that the desired result is obtained. □

The following results characterize the non-asymptotic bounds for the proposed covariance estimators. We provide proofs under the random design only as those under the fixed design are

almost identical. First, we study an event on which the maximum of $|e_{ij} - \epsilon_{ij}|$ over all $i = 1, \ldots, n$ and $j = 1, \ldots, p$ is small with an overwhelming probability.

Lemma A.2.3.3. *Let* $\tau_j = \tau_{0j}[n/\{\log(np) + d\}]^{1/(2+\delta)}$ *and* $\tau_{0j} \geq v_{j,\delta}$ *with* $\delta \in (0, 2]$. *Under Conditions 2,*

$$\max_{1\leq i\leq n, 1\leq j\leq p} |e_{ij} - \epsilon_{ij}| \leq C_5 \left\{d^{1/2} + (2\log n)^{1/2}\right\} \left\{\frac{\log(np) + d}{n}\right\}^{1/2}$$

with probability at least $1 - 4n^{-1}$, *where* C_4 *is a constant depending on* $v_{j,\delta}$ *and* A_0 *only.*

Proof. For each i and j, it follows that

$$\begin{aligned}
\max_{1\leq i\leq n, 1\leq j\leq p} |e_{ij} - \epsilon_{ij}| &= \max_{1\leq i\leq n, 1\leq j\leq p} |Z_i^{\mathrm{T}}(\widehat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j)| \\
&\leq \max_{1\leq i\leq n} \|\widetilde{\mathbf{Z}}_i\| \left\{\max_{1\leq j\leq p} \|\boldsymbol{\Sigma}_Z^{1/2}(\widehat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j)\|\right\} \\
&\leq C_4(d^{1/2} + t^{1/2}) \left(\frac{d+s}{n}\right)^{1/2}
\end{aligned} \tag{A.2.12}$$

with probability at least $1 - 2n\exp(-t) - 2p\exp(-s)$ for $t > 0$ and $s > 0$, where C_4 is a constant depending on $v_{j,\delta}$ and A_0. The third inequality above holds by Theorem 3.1 in [81] and Lemma A.2.1.1. Let $t = \log(n^2)$ and $s = \log(np)$, (A.2.12) implies that

$$\max_{1\leq i\leq n, 1\leq j\leq p} |e_{ij} - \epsilon_{ij}| \leq C_4 \left\{d^{1/2} + (2\log n)^{1/2}\right\} \left\{\frac{\log(np) + d}{n}\right\}^{1/2}$$

with probability at least $1 - 4n^{-1}$. □

Denote the event in Lemma A.2.3.3 by A_Δ, where $\Delta = C_4 \left\{d^{1/2} + (2\log n)^{1/2}\right\} [n^{-1}\{\log(np) + d\}]^{1/2}$. Next, we derive the non-asymptotic bound of $\{\ell'_{\tau_j}(\epsilon_{ij})\}^2$.

Lemma A.2.3.4. *Let* $\tau_j = \tau_{0j}[n\{\log(np) + d\}^{-1}]^{1/(2+\delta)}$, *where* $\tau_{0j} \geq v_{j,\delta}$ *for* $\delta \in (0, 2]$. *On the event* A_Δ,

$$\max_{1\leq j\leq p} \left| \frac{1}{n} \sum_{i=1}^{n} \{\ell'_{\tau_j}(e_{ij})\}^2 - \sigma_{\epsilon,jj} \right| \leq C_6 \left\{\frac{\log(np) + d}{n}\right\}^{\delta/(2+\delta)}$$

holds with probability at least $1 - 6n^{-1}$*, where* C_5 *is a constant depending on* A_0*,* $v_{j,\delta}$*, and* v_j*.*

Proof. First, by the triangle inequality,

$$\begin{aligned}\left|\frac{1}{n}\sum_{i=1}^{n}\{\ell'_{\tau_j}(e_{ij})\}^2 - \sigma_{\epsilon,jj}\right| \leq& \left|\frac{1}{n}\sum_{i=1}^{n}\{\ell'_{\tau_j}(e_{ij})\}^2 - \{\ell'_{\tau_j}(\epsilon_{ij})\}^2\right| \\ &+ \left|\frac{1}{n}\sum_{i=1}^{n}\{\ell'_{\tau_j}(\epsilon_{ij})\}^2 - s_j^2\right| + |s_j^2 - \sigma_{\epsilon,jj}|.\end{aligned} \tag{A.2.13}$$

The last term on the right hand side of (A.2.13) is bounded using Corollary A.2.1. That is, $\sigma_{\epsilon,jj} - s_j^2 \leq 2\delta^{-1}\tau_j^{-\delta}v_{j,\delta}^{2+\delta}$.

For the first term on the right hand side of (A.2.13), it holds that

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\{\ell'_{\tau_j}(e_{ij})\}^2 =& \frac{1}{n}\sum_{i=1}^{n}\{\ell'_{\tau_j}(\epsilon_{ij})\}^2 + \frac{2}{n}\sum_{i=1}^{n}\{\epsilon_{ij}\mathbb{I}(|\epsilon_{ij}| \leq \tau_j)\}(e_{ij} - \epsilon_{ij}) \\ &+ \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(|\epsilon_{ij}| \leq \tau_j)(e_{ij} - \epsilon_{ij})^2 + \frac{1}{n}\sum_{i=1}^{n}R_{2j}(e_{ij})\end{aligned}$$

for $\epsilon_{ij} \neq \pm\tau_j$ and $i = 1, \ldots, n$ by Taylor's theorem with the Peano form of remainder

$$R_{2j}(x) = \{\ell'_{\tau_j}(x)\}^2 - \{\ell'_{\tau_j}(\epsilon_{ij})\}^2 - 2\epsilon_{ij}\mathbb{I}(|\epsilon_{ij}| \leq \tau_j)(x - \epsilon_{ij}) - \mathbb{I}(|\epsilon_{ij}| \leq \tau_j)(x - \epsilon_{ij})^2,$$

where $\lim_{x\to\epsilon_{ij}}\{R_{2j}(x)/(x - \epsilon_{ij})^2\} = 0$. Thus, on the event A_Δ, the remainder is $o_{\mathbb{P}}(\Delta^2)$ and dominated by other terms. Then, by Hoeffding's inequality and Proposition A.2.1, on the event A_Δ,

$$\begin{aligned}&\left|\frac{1}{n}\sum_{i=1}^{n}\{\ell'_{\tau_j}(e_{ij})\}^2 - \frac{1}{n}\sum_{i=1}^{n}\{\ell'_{\tau_j}(\epsilon_{ij})\}^2\right| \\ \leq& \left|\frac{2}{n}\sum_{i=1}^{n}\{\epsilon_{ij}\mathbb{I}(|\epsilon_{ij}| \leq \tau_j)\}(e_{ij} - \epsilon_{ij})\right| + \left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(|\epsilon_{ij}| \leq \tau_j)(e_{ij} - \epsilon_{ij})^2\right| + \left|\frac{1}{n}\sum_{i=1}^{n}R_{2j}(e_{ij})\right| \\ \leq& 2\Delta\left|\frac{1}{n}\sum_{i=1}^{n}\{\epsilon_{ij}\mathbb{I}(|\epsilon_{ij}| \leq \tau_j)\}\right| + \Delta^2\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(|\epsilon_{ij}| \leq \tau_j)\right| + \left|\frac{1}{n}\sum_{i=1}^{n}R_{2j}(e_{ij})\right| \\ \leq& 2\Delta\left[|\mathbb{E}\{\epsilon_{ij}\mathbb{I}(|\epsilon_{ij}| \leq \tau_j)\}| + \tau_j\left\{\frac{\log(np)}{2n}\right\}^{1/2}\right]\end{aligned} \tag{A.2.14}$$

$$
\begin{aligned}
&+\Delta^2\left[\mathbb{P}(|\epsilon_{ij}|\le\tau_j)+\left\{\frac{\log(np)}{2n}\right\}^{1/2}\right]+\left|\frac{1}{n}\sum_{i=1}^{n}R_{2j}(e_{ij})\right| \\
\le& 2\Delta\left[\left|\mathbb{E}\{\ell'_{\tau_j}(\epsilon_{ij})\}\right|+\tau_j\left\{\frac{\log(np)}{2n}\right\}^{1/2}\right]+\Delta^2\left[\mathbb{P}(|\epsilon_{ij}|\le\tau_j)+\left\{\frac{\log(np)}{2n}\right\}^{1/2}\right] \\
&+\left|\frac{1}{n}\sum_{i=1}^{n}R_{2j}(e_{ij})\right| \\
\le& C\{d^{1/2}+(2\log n)^{1/2}\}\left\{\frac{\log(np)+d}{n}\right\}^{(1+\delta)/(2+\delta)}
\end{aligned}
\tag{A.2.15}
$$

holds with probability at least $1-4n^{-1}$ as long as $n\ge C_3\{\log(np)+d\}$, where C is a constant depending on A_0, $\sigma_{\epsilon,jj}$, $v_{j,\delta}$, and v_j.

For the second term, let $Q_{ij}=\ell'_{\tau_j}(\epsilon_{ij})/s_j$ with $\mathbb{E}(Q_{ij}^2)=1$. Then,

$$
\mathbb{E}(Q_{ij}^4)=\frac{\mathbb{E}[\{\ell'_{\tau_j}(\epsilon_{ij})\}^4]}{s_j^4}\le\frac{v_j^4}{s_j^4}
$$

and

$$
\mathbb{E}(Q_{ij}^{2k})\le\frac{v_j^4}{s_j^4}\left(\frac{\tau_j^2}{s_j^2}\right)^{k-2}
$$

for all $k\ge 3$. It follows from Bernstein's inequality that for any $t>0$,

$$
\left|\frac{1}{n}\sum_{i=1}^{n}Q_{ij}^2-1\right|\le\frac{(v_j^4)^{1/2}}{s_j^2}\left(\frac{2t}{n}\right)^{1/2}+\frac{\tau_j^2}{s_j^2}\frac{t}{n}
\tag{A.2.16}
$$

with probability at least $1-2\exp(-t)$. Plugging (A.2.14), (A.2.16), and Corollary A.2.1 into (A.2.13) with $\tau_j=\tau_{0j}[n\{\log(np)+d\}^{-1}]^{1/(2+\delta)}$ and $t=\log(np)$, we yield

$$
\begin{aligned}
&\max_{1\le j\le p}\left|\frac{1}{n}\sum_{i=1}^{n}\{\ell'_{\tau_j}(e_{ij})\}^2-\sigma_{\epsilon,jj}\right| \\
&\le C\left\{d^{1/2}+(2\log n)^{1/2}\right\}\left\{\frac{\log(np)+d}{n}\right\}^{(1+\delta)/(2+\delta)} \\
&\quad + v_j^2\left\{\frac{2\log(np)}{n}\right\}^{1/2}+\tau_{0j}^2\frac{t}{n}\left\{\frac{n}{\log(np)+d}\right\}^{2/(2+\delta)}+\frac{2v_{j,\delta}^{2+\delta}}{\delta}\left\{\frac{\log(np)+d}{n}\right\}^{\delta/(2+\delta)} \\
&\le C_5\left\{\frac{\log(np)+d}{n}\right\}^{\delta/(2+\delta)}
\end{aligned}
$$

as long as $n \ge C_3\{\log(np)+d\}$ with probability at least $1-6n^{-1}$ for $\delta \in (0,2]$. □

The next lemma provides the non-asymptotic bound for m_j.

Lemma A.2.3.5. *On the event A_Δ, for $\tau_j = \tau_{0j}[n\{\log(np)+d\}^{-1}]^{1/(2+\delta)}$ where $\tau_{0j} \ge v_{j,\delta}$ for $\delta \in (0,2]$,*

$$
\max_{1\le j\le p}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}^*_{\tau_j}(e_{ij})-1\right| \le C_7\max\left[\left\{\frac{\log(np)+d}{n}\right\}^{1/2},\frac{\Delta}{h_n}\right]
$$

holds with probability at least $1-2n^{-1}$, where C_6 is a constant only depending on $v_{j,\delta}$ and v_j.

Proof. On the event A_Δ, $|\mathbb{I}^*_{\tau_j}(e_{ij})-\mathbb{I}^*_{\tau_j}(\epsilon_{ij})| \le \Delta {h_n}^{-1}$ due to the Lipschitz continuity of $\mathbb{I}^*_{\tau_j}(x)$. It follows that

$$
\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}^*_{\tau_j}(e_{ij})-1\right| \le \left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}^*_{\tau_j}(\epsilon_{ij})-1\right|+\frac{\Delta}{h_n} \tag{A.2.17}
$$

For the first term on the right hand side of (A.2.17), it follow Hoeffding's inequality and Markov's inequality that

$$
\begin{aligned}
\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}^*_{\tau_j}(\epsilon_{ij})-1\right| &\le \mathbb{E}\{1-\mathbb{I}^*_{\tau_j}(\epsilon_{ij})\}+\left(\frac{t}{2n}\right)^{1/2} \\
&\le \mathbb{P}\left(|\epsilon_{ij})| \ge \tau_j\right)+\left(\frac{t}{2n}\right)^{1/2} \\
&\le \frac{v_{j,\delta}^{2+\delta}}{\tau_j^{2+\delta}}+\left(\frac{t}{2n}\right)^{1/2}
\end{aligned}
$$

with probability at least $1-2\exp(-t)$. Therefore,

$$\max_{1\le j\le p}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}^{*}_{\tau_j}(\epsilon_{ij})-1\right|\le\frac{v_{j,\delta}^{2+\delta}}{\tau_{0j}}\frac{\log(np)+d}{n}+\left\{\frac{\log(np)}{2n}\right\}^{1/2} \tag{A.2.18}$$

with probability at least $1-2n^{-1}$. The lemma is therefore proved. □

Lemma A.2.3.6. *Let* $\tau_j=\tau_{0j}[n\{\log(np)+d\}^{-1}]^{1/(2+\delta)}$ *where* $\tau_{0j}\ge v_{j,\delta}$ *for* $\delta\in(0,2]$. *On the event* A_Δ, *we have*

$$\max_{1\le j\le p}\left\|\mathbf{W}_j-\frac{1}{n}\sum_{i=1}^{n}\mathbf{Z}_i\mathbf{Z}_i^{\mathrm{T}}\right\|\le C_8\max\left[\left\{\frac{\log(np)+d}{n}\right\}^{1/2},\frac{\Delta}{h_n}\right]$$

with probability at least $1-2n^{-1}$, *where* $C_7>0$ *is a constant depending only on* $\lambda_{\max}(\mathbf{\Sigma}_Z)$, A_0, *and* $v_{j,\delta}$ *as long as* $n\ge C_3\{\log(np)+d\}$.

Proof. For $\widetilde{\mathbf{Z}}_i=\mathbf{\Sigma}_Z^{-1/2}\mathbf{Z}_i$, we have

$$\begin{aligned}\left\|\mathbf{W}_j-\frac{1}{n}\sum_{i=1}^{n}\mathbf{Z}_i\mathbf{Z}_i^{\mathrm{T}}\right\|&=\left\|\mathbf{\Sigma}_Z^{1/2}\left[\frac{1}{n}\sum_{i=1}^{n}\{\mathbb{I}^{*}_{\tau_j}(e_{ij})-1\}\widetilde{\mathbf{Z}}_i\widetilde{\mathbf{Z}}_i^{\mathrm{T}}\right]\mathbf{\Sigma}_Z^{1/2}\right\|\\&\le\|\mathbf{\Sigma}_Z\|\left\|\frac{1}{n}\sum_{i=1}^{n}\{\mathbb{I}^{*}_{\tau_j}(e_{ij})-1\}\widetilde{\mathbf{Z}}_i\widetilde{\mathbf{Z}}_i^{\mathrm{T}}\right\|.\end{aligned}$$

On the event A_Δ, for each unit vector $\mathbf{u}\in\mathbb{R}^{d+1}$,

$$\begin{aligned}\left|\mathbf{u}^{\mathrm{T}}\left[\frac{1}{n}\sum_{i=1}^{n}\{\mathbb{I}^{*}_{\tau_j}(e_{ij})-1\}\widetilde{\mathbf{Z}}_i\widetilde{\mathbf{Z}}_i^{\mathrm{T}}\right]\mathbf{u}\right|&\le\left|\frac{1}{n}\sum_{i=1}^{n}\left\{\mathbb{I}\left(|\epsilon_{ij}|\ge\tau_j\right)+\frac{\Delta}{h_n}\right\}\langle\mathbf{u},\widetilde{\mathbf{Z}}_i\rangle^2\right|\\&\le\mathbb{E}(\langle\mathbf{u},\widetilde{\mathbf{Z}}_i\rangle^2)\left[\mathbb{E}\left\{\mathbb{I}\left(|\epsilon_{ij}|\ge\tau_j\right)\right\}+\frac{\Delta}{h_n}+C\max(\rho,\rho^2)\right]\end{aligned}$$

with probability at least $1-2\exp(-t)$ where $\rho=\{n^{-1}(d+t)\}^{1/2}$ and $C>0$ is an absolute constant. From properties of the sub-Gaussian random variable [81], $\mathbb{E}(|\mathbf{u}^{\mathrm{T}}\widetilde{\mathbf{Z}}_i|^k)\le A_1^k(ek/2)\Gamma(k/2)$ for all $k\ge1$, where $A_1\ge e^{-1/2}$ is a constant depending only on A_0. Thus,

$$\mathbb{E}(\langle\mathbf{u},\widetilde{\mathbf{Z}}_i\rangle^2)\le A_1^2e$$

and

$$\mathbb{E}\left\{\mathbb{I}\left(|\epsilon_{ij}| \geq \tau_j\right)\right\} = \mathbb{P}\left(|\epsilon_{ij}| \geq \tau_j\right) \leq \frac{v_{j,\delta}^{2+\delta}}{\tau_{0j}^{2+\delta}} \frac{d+t}{n}.$$

Take $t = \log(np)$. Putting together the obtained bounds yields

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \left\{ \mathbb{I}_{\tau_j}^{*}(e_{ij}) - 1 \right\} \mathbf{Z}_i \mathbf{Z}_i^{\mathrm{T}} \right\| \leq C_7 \max \left[\left\{ \frac{\log(np) + d}{n} \right\}^{1/2}, \frac{\Delta}{h_n} \right]$$

with probability at least $1 - 2n^{-1}$ as long as $n \geq C_3\{\log(np) + d\}$, where C_7 is a constant depending on A_0, $\lambda_{\max}(\mathbf{\Sigma}_Z)$, and $v_{j,\delta}$. □

A.3 Testing hypotheses of the linear combinations of $\boldsymbol{\theta}_j$'s

A.3.1 Method

In this section, we briefly discuss testing hypotheses of the linear combinations of regression coefficients, which is a special case of (2.2.2) in the main article with $q = 1$. For $j = 1, \ldots, p$, $\mathbf{c} \in \mathbb{R}^{d+1}$, and $c_{0j} \in \mathbb{R}$, the two-sided and one-sided hypotheses of interest are

$$H_{0j} : \mathbf{c}^{\mathrm{T}} \boldsymbol{\theta}_j = c_{0j} \text{ versus } H_{1j} : \mathbf{c}^{\mathrm{T}} \boldsymbol{\theta}_j \neq c_{0j}, \tag{A.3.1}$$

and

$$H_{0j} : \mathbf{c}^{\mathrm{T}} \boldsymbol{\theta}_j \leq (\geq) c_{0j} \text{ versus } H_{1j} : \mathbf{c}^{\mathrm{T}} \boldsymbol{\theta}_j > (<) c_{0j}, \tag{A.3.2}$$

respectively. For each j, define

$$U_j = n^{1/2} (\mathbf{c}^{\mathrm{T}} \widehat{\mathbf{\Sigma}}_j \mathbf{c})^{-1/2} (\mathbf{c}^{\mathrm{T}} \widehat{\boldsymbol{\theta}}_j - c_{0j}),$$

where $\widehat{\boldsymbol{\theta}}_j$ and $\widehat{\mathbf{\Sigma}}_j$ are estimated by (2.2.3) and (2.2.6) in the main paper. Notice $U_j^2 = V_j$. For threshold $z > 0$, we estimate the number of false discoveries $V(z)$ by

$$\widehat{V}(z) = \begin{cases} 2p\Phi(-z) & \text{(two-sided)}, \\ p\Phi(-z) & \text{(one-sided)}. \end{cases}$$

Let the number of discoveries by $R(z) = \sum_{j=1}^{p} \mathbb{I}(U_j \geq z)$. Then, we compute

$$\widehat{z}_\alpha = \inf \{z \geq 0 : \mathrm{AFDP}(z) \leq \alpha\},$$

where $\mathrm{AFDP}(z) = \widehat{V}(z)/R(z)$. For $j = 1, \ldots, p$, H_{0j} in (A.3.1) or (A.3.2) is rejected whenever $U_j \geq \widehat{z}_\alpha$.

A.3.2 Theoretical guarantees

The following result for testing (A.3.1) is a straightforward corollary of Theorem 2.3.1. Denote $U_j^\circ = n^{1/2}(\mathbf{c}^{\mathrm{T}}\boldsymbol{\Sigma}_j\mathbf{c})^{-1/2}(\mathbf{c}^{\mathrm{T}}\widehat{\boldsymbol{\theta}}_j - c_{0j})$ with known covariance $\boldsymbol{\Sigma}_j$. For $\mathcal{H}_0 = \{j : 1 \leq j \leq p, H_{0j} \text{ is true}\}$, let $V^\circ(z) = \sum_{j \in \mathcal{H}_0} \mathbb{I}(U_j^\circ \geq z)$ and $R^\circ(z) = \sum_{j=1}^{p} \mathbb{I}(U_j^\circ \geq z)$. Define $\mathrm{AFDP}_{c1}^\circ(z) = 2p_0\Phi(-z)/R^\circ(z)$ to be the counterpart of (2.3.1).

Theorem A.3.1. *Consider testing* (A.3.1). *Assume Conditions 1 and 2 hold, and* $p_0 \geq ap$ *for some* $a \in (0, 1)$. *Let* $\tau_j = \tau_{0j} n^{1/(2+\delta)}\{\log(np) + d\}^{-1/(2+\delta)}$ *with* $\tau_{0j} \geq v_{j,\delta}$ *and* $\delta \in (0, 2]$. *Then, for any* $z \geq 0$, $|\mathrm{FDP}^\circ(z) - \mathrm{AFDP}_{c1}^\circ(z)| = o_{\mathbb{P}}(1)$ *as* $n, p \to \infty$.

Next, we provide the corresponding result for testing (A.3.2). Similarly, let $\mathrm{AFDP}_{c2}^\circ(z) = p_0\Phi(-z)/R^\circ(z)$.

Theorem A.3.2. *Consider testing* (A.3.2). *Assume Conditions 1 and 2 hold, and* $p_0 \geq ap$ *for some* $a \in (0, 1)$. *Let* $\tau_j = \tau_{0j} n^{1/(2+\delta)}\{\log(np) + d\}^{-1/(2+\delta)}$ *with* $\tau_{0j} \geq v_{j,\delta}$ *and* $\delta \in (0, 2]$. *Then, for any* $z \geq 0$, $|\mathrm{FDP}^\circ(z) - \mathrm{AFDP}_{c2}^\circ(z)| = o_{\mathbb{P}}(1)$ *as* $n, p \to \infty$.

Proof. The proof is similar to that of Theorem 2.3.1. Let $z \geq 0$. We will show the stronger result that on event $\{p^{-1}R^\circ(z) \geq c\}$ for some $c > 0$,

$$p_0^{-1}V^\circ(z) = \Phi(-z) + O_{\mathbb{P}}(p^{-\kappa_1} + n^{-1/2} + [n^{-1}\{\log(np) + d\}]^{\delta/(2+\delta)}), \tag{A.3.3}$$

which leads to the conclusion immediately.

Let $\sigma_j^2 = \mathbf{c}^{\mathrm{T}}\boldsymbol{\Sigma}_j\mathbf{c} = \sigma_{\epsilon,jj}(\mathbf{c}^{\mathrm{T}}\boldsymbol{\Sigma}_Z^{-1}\mathbf{c}) \in \mathbb{R}$, $U_j^\circ = n^{1/2}\sigma_j^{-1}(\mathbf{c}^{\mathrm{T}}\widehat{\boldsymbol{\theta}}_j - \mathbf{c}^{\mathrm{T}}\boldsymbol{\theta}_j)$, and

$$S_j = n^{-1/2}\|\mathbf{c}^{\mathrm{T}}\boldsymbol{\Sigma}_Z^{-1/2}\|^{-1}\mathbf{c}^{\mathrm{T}}\boldsymbol{\Sigma}_Z^{-1}\sum_{i=1}^{n}[\ell_\tau'(\epsilon_{ij})\mathbf{Z}_i - \mathbb{E}\{\ell_\tau'(\epsilon_{ij})\mathbf{Z}_i\}],$$
$$R_j = n^{-1/2}\|\mathbf{c}^{\mathrm{T}}\boldsymbol{\Sigma}_Z^{-1/2}\|^{-1}\mathbf{c}^{\mathrm{T}}\boldsymbol{\Sigma}_Z^{-1}\sum_{i=1}^{n}\mathbb{E}\{\ell_\tau'(\epsilon_{ij})\mathbf{Z}_i\}].$$

For every $j \in H_{0j}$ and $t \geq 1$, it follows from Lemma A.2.1.1 that

$$\begin{aligned}|U_j^\circ - \sigma_{\epsilon,jj}^{-1/2}(S_j + R_j)| &= \left|n^{1/2}\sigma_j^{-1}(\mathbf{c}^{\mathrm{T}}\widehat{\boldsymbol{\theta}}_j - \mathbf{c}^{\mathrm{T}}\boldsymbol{\theta}_j) - n^{-1/2}\sigma_j^{-1}\mathbf{c}^{\mathrm{T}}\boldsymbol{\Sigma}_Z^{-1}\sum_{i=1}^{n}\ell_\tau'(\epsilon_{ij})\mathbf{Z}_i\right| \\ &\leq n^{1/2}\sigma_{\epsilon,jj}^{-1/2}\left\|\boldsymbol{\Sigma}_Z^{1/2}(\widehat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j) - \frac{1}{n}\sum_{i=1}^{n}\ell_{\tau_j}'(\epsilon_{ij})\boldsymbol{\Sigma}_Z^{-1/2}\mathbf{Z}_i\right\| \\ &\leq C_2\frac{\tau_0(d+t)}{(n\sigma_{\epsilon,jj})^{1/2}}\end{aligned} \tag{A.3.4}$$

with probability greater than $1 - 3\exp(-t)$ as long as $n \geq C_4(d+t)$ with $\tau_j = \tau_{0j}\{n(d+t)^{-1}\}^{1/(2+\delta)}$.

For $j = 1, \ldots, p$, denote $E_{1j}(t)$ on which event (A.3.4) holds, and define $E_1 = \bigcap_{j=1}^{p} E_{1j}(t)$. On E_1,

$$\sum_{j\in H_{0j}}\mathbb{I}\left\{\sigma_{\epsilon,jj}^{-1/2}S_j \geq z + C_2\frac{\tau_0(d+t)}{(n\sigma_{\epsilon,jj})^{1/2}}\right\} \leq V(z) \leq \sum_{j\in H_{0j}}\mathbb{I}\left\{\sigma_{\epsilon,jj}^{-1/2}S_j \geq z - C_2\frac{\tau_0(d+t)}{(n\sigma_{\epsilon,jj})^{1/2}}\right\} \tag{A.3.5}$$

with probability $1 - 3pe^{-t}$. For $x \in \mathbb{R}$, define

$$V^+(x) = \sum_{j\in H_{0j}}\mathbb{I}(\sigma_{\epsilon,jj}^{-1/2}S_j \geq x). \tag{A.3.6}$$

Hence, (A.3.5) can be written as

$$p_0^{-1}V^+\left\{z + C_2\frac{\tau_0(d+t)}{(n\sigma_{\epsilon,jj})^{1/2}}\right\} \leq p_0^{-1}V^\circ(z) \leq p_0^{-1}V^+\left\{z - C_2\frac{\tau_0(d+t)}{(n\sigma_{\epsilon,jj})^{1/2}}\right\}. \tag{A.3.7}$$

Therefore, we only need to derive the orders of $V^+(x)$. The rest of the proof is almost identical to that of Theorem 2.3.1 by replacing Lemma A.2.2.4 by Lemma 2.1 from [160]. We can easily obtain a similar bound for $\mathbb{E}[\{p_0^{-1}V^+(z) - \Phi(-z)\}^2]$ that

$$\mathbb{E}[\{p_0^{-1}V^+(z) - \Phi(-z)\}^2] \precsim p^{-\kappa_1} + n^{-1/2} + \left(\frac{d+t}{n}\right)^{\delta/(2+\delta)}. \tag{A.3.8}$$

Recall that $\mathbb{P}(A_1) \leq 1 - 3pe^{-t}$ whenever $n \gtrsim d + t$. Taking $t = \log(np)$ in (A.3.7) and (A.3.8) proves (A.3.3). □

Remark A.3.1. For testing $H_{0j} : \mathbf{c}^{\mathrm{T}}\boldsymbol{\theta}_j \geq c_0$ versus $H_{1j} : \mathbf{c}^{\mathrm{T}}\boldsymbol{\theta}_j < c_0$, we can use the same argument with (A.3.5) and (A.3.6) replaced by

$$\sum_{j\in\mathcal{H}_{0j}} \mathbb{I}\left\{\sigma_{\epsilon,jj}^{-1/2}S_j \leq -z - C_2\frac{\tau_0(d+t)}{(n\sigma_{\epsilon,jj})^{1/2}}\right\} \leq V^\circ(z) \leq \sum_{j\in\mathcal{H}_{0j}} \mathbb{I}\left\{\sigma_{\epsilon,jj}^{-1/2}S_j \leq -z + C_2\frac{\tau_0(d+t)}{(n\sigma_{\epsilon,jj})^{1/2}}\right\}$$

and $V^-(x) = \sum_{j\in\mathcal{H}_{0j}} \mathbb{I}(\sigma_{\epsilon,jj}^{-1/2}S_j \leq -x)$, respectively.

A.4 Results under the fixed design

In this section, we consider our testing procedure in Section 2.2 for model (2.2.1) under the fixed design. Denote $\mathbf{z}_i^{\mathrm{T}} = \begin{bmatrix} 1 & \mathbf{x}_i^{\mathrm{T}} \end{bmatrix}$ the ith row of design matrix $\mathbf{Z}$. We first impose the following regularity condition, which is similar to that in [12].

Condition 3. *The Gram matrix $\mathbf{S}_n = n^{-1}\sum_{i=1}^n \mathbf{z}_i\mathbf{z}_i^{\mathrm{T}}$ is positive definite and there exist constants c_l and c_u such that $c_l \leq \lambda_{\min}(\mathbf{S}_n) \leq \lambda_{\max}(\mathbf{S}_n) \leq c_u$. As $n \to \infty$, $\mathbf{S}_n \to \boldsymbol{\Sigma}_Z$ which is also positive definite.*

The following condition is similar to the finite fourth order moment condition under the random design.

Condition 4. *There exist constants $\kappa, M > 0$ such that, for $\widetilde{\mathbf{z}}_i = \mathbf{S}_n^{-1/2}\mathbf{z}_i$,*

$$\sup_{\mathbf{u}\in\mathbb{S}^d} \frac{1}{n}\sum_{i=1}^n (\mathbf{u}^{\mathrm{T}}\widetilde{\mathbf{z}}_i)^4 \exp(\kappa|\mathbf{u}^{\mathrm{T}}\widetilde{\mathbf{z}}_i|^2) \leq M.$$

We start with the counterpart of Theorem 2.3.1 to show that our procedure controls the false discovery proportion given the covariances of regression coefficients under the fixed design.

Theorem A.4.1. *Assume Conditions 1, 3, and 4 hold, and* $p_0 \geq ap$ *for some* $a \in (0,1)$. *Let* $\tau_j = \tau_{0j}\{n/\log(np)\}^{1/(2+\delta)}$ *where* $\tau_{0j} \geq v_{j,\delta}$ *for some* $\delta \in (0,2]$. *Then, for any* $z \geq 0$, $|\mathrm{FDP}^\circ(z) - \mathrm{AFDP}^\circ(z)| = o_{\mathbb{P}}(1)$ *as* $n, p \to \infty$.

Proof. For ease of exposition, we use z^2 instead of z in this proof. The proof is similar to that of Theorem 2.3.1. First, define

$$\begin{aligned}\mathbf{s}_j &= n^{-1/2}(\mathbf{C}\mathbf{S}_n^{-1}\mathbf{C}^{\mathrm{T}})^{-1/2}\mathbf{C}\mathbf{S}_n^{-1/2}\sum_{i=1}^n \mathbf{S}_n^{-1/2}\mathbf{Z}_i[\ell'_\tau(\epsilon_{ij}) - \mathbb{E}\{\ell'_\tau(\epsilon_{ij})\}],\\ \mathbf{r}_j &= n^{-1/2}(\mathbf{C}\mathbf{S}_n^{-1}bC^{\mathrm{T}})^{-1/2}\mathbf{C}\mathbf{S}_n^{-1/2}\sum_{i=1}^n \mathbf{S}_n^{-1/2}\mathbf{Z}_i\mathbb{E}\{\ell'_{\tau_j}(\epsilon_{ij})\}.\end{aligned} \tag{A.4.1}$$

By Proposition A.2.1, $\|\mathbf{r}_j\|$ is a small order term. Together with Corollary A.4.1, it implies that

$$\|\mathbf{T}_j^\circ - \sigma_{\epsilon,jj}^{-1/2}(\mathbf{s}_j + \mathbf{r}_j)\| \leq A\tau_{0j}(d+t)^{1/2}\left(\frac{dt}{\sigma_{\epsilon,jj}n}\right)^{1/2} \tag{A.4.2}$$

with probability $1 - 2d\exp(-t)$ as long as $n \geq \max\{32L_\infty^4 d^2 t, 2\kappa^{-2}(2d+t)\}$, where $\Delta_{n,\delta} = n^{-1}\sum_{i=1}^n v_\delta \widetilde{\mathbf{z}}_i\widetilde{\mathbf{z}}_i^{\mathrm{T}}$. The rest of the proof are almost identical to that of Theorem 2.3.1. Define $E_{1j}(t)$ the event on which (A.4.2) holds and let $E_1(t) = \bigcap_{j=1}^p E_{1j}(t)$ where $\mathbb{P}\{E_1(t)\} = 1 - 2dp\exp(-t)$. One can obtain the counterparts of (A.1.5) and (A.1.12),

$$\begin{aligned}&\sum_{j\in\mathcal{H}_0}\mathbb{I}\left[\|\sigma_{\epsilon,jj}^{-1/2}\mathbf{s}_j\| \geq z + A\tau_{0j}(d+t)^{1/2}\left(\frac{dt}{\sigma_{\epsilon,jj}n}\right)^{1/2}\right] \leq V^\circ(z^2)\\ \leq &\sum_{j\in\mathcal{H}_0}\mathbb{I}\left[\|\sigma_{\epsilon,jj}^{-1/2}\mathbf{s}_j\| \geq z - A\tau_{0j}(d+t)^{1/2}\left(\frac{dt}{\sigma_{\epsilon,jj}n}\right)^{1/2}\right]\end{aligned}$$

and

$$\mathbb{E}\{p_0^{-1}V^\circ(z^2) - \mathbb{P}(\chi_q^2 > z^2)\}^2 \precsim q^{1/2}p^{-\kappa_1} + n^{-1/2}q^{7/4} + q(t/n)^{\delta/(2+\delta)}.$$

Using Lemmas A.4.1.4 and A.4.1.5, we can obtain the desired result by taking $t = \log(np)$. □

Counterparts of Theorems A.3.1 and A.3.2 remain true under the assumptions for Theorem A.4.1. Their statements and proofs are identical to those of Theorems A.3.1 and A.3.2 and therefore are omitted.

A.4.1 Technical lemmas under the fixed design

In this subsection, for the sake of completeness, we collect some auxiliary lemmas used for proving Theorem A.4.1. Most proofs, except that for Lemma A.4.1.2, are omitted given their similarities to those in Section A.2.2. We start with three technical lemmas, which are modified from results in [12]. Lemmas A.4.1.1-A.4.1.3 provide general conclusions for the adaptive Huber regression with dimension d under the fixed design, and we suppress index j in their statements for ease of presentation.

Let $\mathcal{L}_\tau(\boldsymbol{\theta}) := n^{-1}\sum_{i=1}^n \ell_\tau(y_i - \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\theta})$. Lemma A.4.1.1 provides the lower bound of $\lambda_{\min}\{\mathbf{S}_n^{-1/2}\nabla^2\mathcal{L}_\tau(\widetilde{\boldsymbol{\theta}})\mathbf{S}_n^{-1/2}\}$, which can be shown by slightly modifying similar arguments in [12] under Condition 3.

Lemma A.4.1.1. *Assume Condition 3 holds and* $v_\delta := \{\mathbb{E}(|\epsilon|^{2+\delta})\}^{1/(2+\delta)} < \infty$ *for* $\delta \in (0.2]$. *Then for any* $t, r > 0$, *the matrix* $\mathbf{S}_n^{-1/2}\nabla^2\mathcal{L}_\tau(\widetilde{\boldsymbol{\theta}})\mathbf{S}_n^{-1/2}$ *with* $\tau > 2L_2 r$ *satisfies that*

$$\min_{\widetilde{\boldsymbol{\theta}}\in\mathbb{R}^{d+1}:\|\mathbf{S}_n^{1/2}(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta})\|\le r} \lambda_{\min}\{\mathbf{S}_n^{-1/2}\nabla^2\mathcal{L}_\tau(\widetilde{\boldsymbol{\theta}})\mathbf{S}_n^{-1/2}\}$$
$$\ge 1 - (2L_2 r/\tau)^2 - L_2^2\{(2v_\delta/\tau)^{2+\delta} + (2n)^{-1/2}t^{1/2}\},$$

with probability at least $1-\exp(-t)$ *where* $L_2 = \max_{1\le i\le n}\|\widetilde{\mathbf{z}}_i\|$.

Proof. It follows that

$$S_n^{-1/2}\nabla^2\mathcal{L}_\tau(\theta)S_n^{-1/2} = n^{-1}\sum_{i=1}^n \widetilde{z}_i\widetilde{z}_i^{\mathrm{T}}\mathbb{I}(|y_i - z_i^{\mathrm{T}}\theta| \le \tau)$$
$$= I - n^{-1}\sum_{i=1}^n \widetilde{z}_i\widetilde{z}_i^{\mathrm{T}}\mathbb{I}(|y_i - z_i^{\mathrm{T}}\theta| > \tau).$$

Define $\widetilde{\theta}_0 = \widetilde{\theta} - \theta$ so that $y_i - z_i^{\mathrm{T}}\widetilde{\theta} = \epsilon_i - z_i^{\mathrm{T}}\widetilde{\theta}_0$., then it follows that

$$\mathbb{I}(|y_i - z_i^{\mathrm{T}}\widetilde{\theta}| > \tau) \leq \mathbb{I}(|\epsilon_i| > \tau/2) + \mathbb{I}(|z_i^{\mathrm{T}}\widetilde{\theta}_0| > \tau/2).$$

For any $u \in \mathbb{S}^d$ and $\widetilde{\theta} \in \mathbb{R}^{d+1}$ satisfying $\|S_n^{1/2}\widetilde{\theta}_0\|_2 \leq r$,

$$\begin{aligned}
&\langle u, S_n^{-1/2}\nabla^2\mathcal{L}_\tau(\theta)S_n^{-1/2}u\rangle \\
&\geq 1 - n^{-1}\sum_{i=1}^n \langle\widetilde{z}_i, u\rangle^2\mathbb{I}(|\epsilon_i| > \tau/2) - n^{-1}\sum_{i=1}^n \langle\widetilde{z}_i, u\rangle^2\mathbb{I}(|z_i^{\mathrm{T}}\widetilde{\theta}_0| > \tau/2) \\
&\geq 1 - \max_{1\leq i\leq n}\|\widetilde{z}_i\|_2^2\left\{n^{-1}\sum_{i=1}^n \mathbb{I}(|\epsilon_i| > \tau/2) + 4\tau^{-2}\|S_n^{1/2}\widetilde{\theta}_0\|_2^2\right\} \\
&\geq 1 - (2L_2 r/\tau)^2 - L_2^2 n^{-1}\sum_{i=1}^n \mathbb{I}(|\epsilon_i| > \tau/2),
\end{aligned}$$

provided that $\tau > 2L_2 r$ where the inequality in the third line holds by

$$\begin{aligned}
n^{-1}\sum_{i=1}^n \langle\widetilde{z}_i, u\rangle^2\mathbb{I}(|z_i^{\mathrm{T}}\widetilde{\theta}_0| > \tau/2) &\leq n^{-1}\sum_{i=1}^n \langle\widetilde{z}_i, u\rangle^2\frac{|z_i^{\mathrm{T}}\widetilde{\theta}_0|^2}{|z_i^{\mathrm{T}}\widetilde{\theta}_0|^2}\mathbb{I}(|z_i^{\mathrm{T}}\widetilde{\theta}_0| > \tau/2) \\
&\leq 4\tau^{-2}\left(n^{-1}\sum_{i=1}^n \langle\widetilde{z}_i, u\rangle^2|\widetilde{z}_i^{\mathrm{T}}S_n^{1/2}\widetilde{\theta}_0|^2\right) \\
&\leq \left(\max_{1\leq i\leq n}\|\widetilde{z}_i\|^2\right)4\tau^{-2}\|S_n^{1/2}\widetilde{\theta}_0\|_2^2\left(n^{-1}\sum_{i=1}^n \langle\widetilde{z}_i, u\rangle^2\right) \\
&= \left(\max_{1\leq i\leq n}\|\widetilde{z}_i\|^2\right)4\tau^{-2}\|S_n^{1/2}\widetilde{\theta}_0\|_2^2\left\{u^{\mathrm{T}}\left(n^{-1}\sum_{i=1}^n \widetilde{z}_i\widetilde{z}_i^{\mathrm{T}}\right)u\right\} \\
&= \left(\max_{1\leq i\leq n}\|\widetilde{z}_i\|^2\right)4\tau^{-2}\|S_n^{1/2}\widetilde{\theta}_0\|_2^2
\end{aligned}$$

By Hoeffding's inequality, for any $z \geq 0$, we have

$$n^{-1}\sum_{i=1}^n \mathbb{I}(|\epsilon_i| > \tau/2) \leq n^{-1}\sum_{i=1}^n \mathbb{P}(|\epsilon_i| > \tau/2) + z$$

with probability at least $1 - \exp(-2nz^2)$. Putting this together with $\mathbb{P}(|\epsilon_i| > \tau/2) \leq (2v_\delta/\tau)^{2+\delta}$ and Condition 3,

$$\langle u, S_n^{-1/2}\nabla^2\mathcal{L}_\tau(\theta)S_n^{-1/2}u\rangle \geq 1-(2L_2r/\tau)^2 - L_2^2\{(2v_\delta/\tau)^{2+\delta}+z\}.$$

Taking $z=\sqrt{t/(2n)}$ gives the desired result. □

The following lemma is a variation of Theorem 1 in [12] under the fixed design. The original theorem assumes finite $(1+\delta)$ order moment of ϵ_i for some $\delta>0$. Using Lemma A.4.1.1, the proof of Lemma A.4.1.2 is similar to that of Theorem 1 in [12], while the major technical challenge focuses on deriving the sharp non-asymptotic rate using our adaptive robustification parameter.

Lemma A.4.1.2. *Assume Conditions 1 and 3 hold and $v_\delta<\infty$ for $\delta\in(0.2]$. Then, for any $t>0$ and $\tau_0\geq v_\delta$, the adaptive Huber regression estimator $\widehat{\boldsymbol{\theta}}=(\widehat{\mu},\widehat{\boldsymbol{\beta}}^{\mathrm{T}})^{\mathrm{T}}\in\mathbb{R}^{d+1}$ in* (2.2.3) *with $\tau=\tau_0(n/t)^{1/(2+\delta)}$ satisfies*

$$\|\mathbf{S}_n^{1/2}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta})\| \leq C(L_\infty,\delta,v_\delta)d^{1/2}\left(\frac{t}{n}\right)^{1/2}$$

with probability at least $1-(2d+3)\exp(-t)$ as long as $n\geq 32L_\infty^4 d^2 t$, where $L_\infty=\max_{1\leq i\leq n}\|\widetilde{\mathbf{z}}_i\|_\infty$ and $C(L_\infty,\delta,v_\delta)$ is a constant only depending on L_∞, δ, and v_δ.

Proof. Recall that $\tau=\tau_0(n/t)^{1/(2+\delta)}$. Let $\widehat{\boldsymbol{\theta}}_\eta=\boldsymbol{\theta}+\eta(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta})$ with $\eta\in(0,1]$ so that $\|\mathbf{S}_n^{1/2}(\widehat{\boldsymbol{\theta}}_\eta-\boldsymbol{\theta})\|\leq r$. Lemma 2 from [12] gives

$$\langle\nabla\mathcal{L}_\tau(\widehat{\boldsymbol{\theta}}_\eta)-\nabla\mathcal{L}_\tau(\boldsymbol{\theta}),\widehat{\boldsymbol{\theta}}_\eta-\boldsymbol{\theta}\rangle \leq \eta\langle\nabla\mathcal{L}_\tau(\widehat{\boldsymbol{\theta}})-\nabla\mathcal{L}_\tau(\boldsymbol{\theta}),\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}\rangle,$$

where $\nabla\mathcal{L}_\tau(\widehat{\boldsymbol{\theta}})=0$ by the Karush-Kuhn-Tucker condition. By the mean value theorem for vector-valued functions, the equality

$$\nabla\mathcal{L}_\tau(\widehat{\boldsymbol{\theta}}_\eta)-\nabla\mathcal{L}_\tau(\boldsymbol{\theta})=\left[\int_0^1\nabla^2\mathcal{L}_\tau\{(1-t)\boldsymbol{\theta}+t\widehat{\boldsymbol{\theta}}_\eta\}dt\right](\widehat{\boldsymbol{\theta}}_\eta-\boldsymbol{\theta})$$

holds, where the integral of a matrix is component-wise integrals. If there exists a constant $a_0>0$ such that

$$\min_{\widetilde{\theta}\in\mathbb{R}^{d+1}:\|\mathbf{S}_n^{1/2}(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta})\|\le r} \lambda_{\min}\{\mathbf{S}_n^{-1/2}\nabla^2\mathcal{L}_\tau(\widetilde{\boldsymbol{\theta}})\mathbf{S}_n^{-1/2}\} \ge a_0, \tag{A.4.3}$$

then

$$\begin{aligned}
&a_0\|\mathbf{S}_n^{1/2}(\widehat{\boldsymbol{\theta}}_\eta-\boldsymbol{\theta})\|_2^2\\
&\le \lambda_{\min}\left[\mathbf{S}_n^{-1/2}\nabla^2\mathcal{L}_\tau\{(1-t)\boldsymbol{\theta}+t\widehat{\boldsymbol{\theta}}_\eta\}\mathbf{S}_n^{-1/2}\right]\|\mathbf{S}_n^{1/2}(\widehat{\boldsymbol{\theta}}_\eta-\boldsymbol{\theta})\|_2^2\\
&=\frac{(\widehat{\boldsymbol{\theta}}_\eta-\boldsymbol{\theta})^{\mathrm{T}}\mathbf{S}_n^{1/2}}{\|\mathbf{S}_n^{1/2}(\widehat{\boldsymbol{\theta}}_\eta-\boldsymbol{\theta})\|_2}\left[\mathbf{S}_n^{-1/2}\nabla^2\mathcal{L}_\tau\{(1-t)\boldsymbol{\theta}+t\widehat{\boldsymbol{\theta}}_\eta\}S_n^{-1/2}\right]\frac{\mathbf{S}_n^{1/2}(\widehat{\boldsymbol{\theta}}_\eta-\boldsymbol{\theta})}{\|\mathbf{S}_n^{1/2}(\widehat{\boldsymbol{\theta}}_\eta-\boldsymbol{\theta})\|_2}\|\mathbf{S}_n^{1/2}(\widehat{\boldsymbol{\theta}}_\eta-\boldsymbol{\theta})\|_2^2\\
&=(\widehat{\boldsymbol{\theta}}_\eta-\boldsymbol{\theta})^{\mathrm{T}}\left[\nabla^2\mathcal{L}_\tau\{(1-t)\boldsymbol{\theta}+t\widehat{\boldsymbol{\theta}}_\eta\}\right](\widehat{\boldsymbol{\theta}}_\eta-\boldsymbol{\theta})
\end{aligned}$$

and

$$\begin{aligned}
a_0\|\mathbf{S}_n^{1/2}(\widehat{\boldsymbol{\theta}}_\eta-\boldsymbol{\theta})\|_2^2 &\le \int_0^1(\widehat{\boldsymbol{\theta}}_\eta-\boldsymbol{\theta})^{\mathrm{T}}\left[\nabla^2\mathcal{L}_\tau\{(1-t)\boldsymbol{\theta}+t\widehat{\boldsymbol{\theta}}_\eta\}\right](\widehat{\boldsymbol{\theta}}_\eta-\boldsymbol{\theta})dt\\
&=(\widehat{\boldsymbol{\theta}}_\eta-\boldsymbol{\theta})^{\mathrm{T}}\left\{\nabla\mathcal{L}_\tau(\widehat{\boldsymbol{\theta}}_\eta)-\nabla\mathcal{L}_\tau(\boldsymbol{\theta})\right\}\\
&\le \eta(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta})^{\mathrm{T}}\{-\nabla\mathcal{L}_\tau(\boldsymbol{\theta})\}\\
&\le \eta\|\mathbf{S}_n^{-1/2}\nabla\mathcal{L}_\tau(\boldsymbol{\theta})\|\|\mathbf{S}_n^{1/2}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta})\|
\end{aligned}$$

by putting together all the results above. Setting $\eta=1$ yields

$$a_0\|\mathbf{S}_n^{1/2}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta})\|_2 \le \|\mathbf{S}_n^{-1/2}\nabla\mathcal{L}_\tau(\boldsymbol{\theta})\|_2. \tag{A.4.4}$$

Denote the kth entry of $\xi=\mathbf{S}_n^{-1/2}\nabla\mathcal{L}_\tau(\boldsymbol{\theta})$ by $\xi_k=-n^{-1}\sum_{i=1}^n\ell_\tau'(\epsilon_i)\widetilde{z}_{ik}$. By the triangle inequality, $|\xi_k|\le|\xi_k-\mathbb{E}(\xi_k)|+|\mathbb{E}(\xi_k)|$. By Proposition A.2.1, as long as $n\ge(\sigma^{-2}v_\delta^{2+\delta})^{(2+\delta)/\delta}t$, it follows that $|\mathbb{E}\ell_\tau'(\epsilon)|\le\tau^{-(1+\delta)}v_\delta^{2+\delta}$. By the definition of $\ell_\tau'(\cdot)$,

$$\begin{aligned}
&\left|\mathbb{E}\{n^{-1}\ell_\tau'(\epsilon_i)\widetilde{z}_{ik}\}\right|\le\frac{v_\delta^{2+\delta}}{n\tau^{1+\delta}}L_{\infty},\\
&\left|\frac{1}{n}\ell_\tau'(\epsilon_i)\widetilde{z}_{ik}-\mathbb{E}\left\{\frac{1}{n}\ell_\tau'(\epsilon_i)\widetilde{z}_{ik}\right\}\right|\le|\widetilde{z}_{ik}|\left(\frac{\tau}{n}+\frac{v_\delta^{2+\delta}}{n\tau^{1+\delta}}\right)\le L_\infty\left(\frac{\tau}{n}+\frac{v_\delta^{2+\delta}}{n\tau^{1+\delta}}\right),\text{ and}\\
&\mathbb{E}\left[\{n^{-1}\ell_\tau'(\epsilon_i)\widetilde{z}_{ik}-n^{-1}\mathbb{E}\{\ell_\tau'(\epsilon_i)\}\widetilde{z}_{ik}\}^2\right]\le n^{-2}s^2\widetilde{z}_{ik}^2
\end{aligned}$$

By Bernstein's inequality [81],

$$|\xi_k| \leq f(n,t) + \frac{v_\delta^{2+\delta}}{\tau^{1+\delta}} L_\infty \leq C \left(\frac{t}{n}\right)^{1/2}$$

with probability at least $1 - 2\exp(-t)$ as long as $\tau \geq v_\delta (n/t)^{1/(2+\delta)}$ and $n > t$ where C is a constant depending on L_∞, δ and v_δ, and

$$\begin{aligned} f(n,t) &= \frac{L_\infty}{3}\frac{t}{n}\left(\tau + \frac{v_\delta^{2+\delta}}{\tau^{1+\delta}}\right) + \frac{1}{3}\left\{\frac{L_\infty^2}{9}\frac{t^2}{n^2}\left(\tau + \frac{v_\delta^{2+\delta}}{\tau^{1+\delta}}\right)^2 + 18s^2\frac{t}{n}\frac{1}{n}\sum_{i=1}^{n}\widetilde{z}_{ik}^2\right\}^{1/2} \\ &\leq \frac{L_\infty}{3}\frac{t}{n}\left(\tau + \frac{v_\delta^{2+\delta}}{\tau^{1+\delta}}\right) + \frac{L_\infty}{3}\left\{\frac{1}{9}\frac{t^2}{n^2}\left(\tau + \frac{v_\delta^{2+\delta}}{\tau^{1+\delta}}\right)^2 + 18s^2\frac{t}{n}\right\}^{1/2} \\ &\precsim \left(\frac{t}{n}\right)^{1/2}. \end{aligned}$$

Then, for any $t > 0$,

$$\begin{aligned} \mathbb{P}(\|\xi\|_2 \geq C(d+1)^{1/2} n^{-1/2} t^{1/2}) &\leq \mathbb{P}(\|\xi\|_\infty \geq C n^{-1/2} t^{1/2}) \\ &\leq \sum_{k=1}^{d+1} \mathbb{P}(|\xi_k| \geq C n^{-1/2} t^{1/2}) \\ &\leq 2(d+1)\exp(-t). \end{aligned} \tag{A.4.5}$$

By Lemma A.4.1.1, (A.4.3) holds for $a_0 = 1/2$ and $r = \tau/(4L_2)$ with probability at least $1 - \exp(-t)$ since

$$\begin{aligned} \min_{\|\mathbf{S}_n^{1/2}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta})\| \leq \tau/(4L_2)} \lambda_{\min}\{\mathbf{S}_n^{-1/2} \nabla^2 \mathcal{L}_\tau(\widetilde{\boldsymbol{\theta}}) \mathbf{S}_n^{-1/2}\} &\geq \frac{3}{4} - L_2^2 \left\{\left(\frac{2v_\delta}{\tau_0}\right)^{2+\delta}\frac{t}{n} + \left(\frac{t}{2n}\right)^{1/2}\right\} \\ &\geq \frac{1}{2}, \end{aligned}$$

holds as long as $n \geq \max(32L_2^4, 2^{5+\delta}L_2^2)t = 32\max(L_2^4, 2^\delta L_2^2)t$. By (A.4.4) and (A.4.5), we have

$$\|\mathbf{S}_n^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})\| \leq 2Cd^{1/2}n^{-1/2}t^{1/2}$$

with probability at least $1 - (2d+3)\exp(-t)$. □

Lemma A.4.1.3 provides a nonasymptotic Bahadur representation under the fixed design, and it implies that $\sqrt{n}\mathbf{S}_n^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ can be approximated by the multivariate normal distribution. It is a variation of Theorem 3.3 in the first version of [12], which is available on ArXiv:1706.06991v1. It can be proved using Lemma A.4.1.2 with $\tau = \tau_0(n/t)^{1/(2+\delta)}$.

Lemma A.4.1.3. *Assume that Conditions 3 and 4 hold, and that $v_\delta < \infty$ for $\delta \in (0.2]$. Then, for any $t > 0$ and $\tau_0 \geq v_\delta$, the estimator $\widehat{\boldsymbol{\theta}}$ given in (2.2.3) with $\tau = \tau_0(n/t)^{1/(2+\delta)}$ satisfies that*

$$\mathbb{P}\left\{\left\|\mathbf{S}_n^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) - \frac{1}{n}\sum_{i=1}^n \ell'_\tau(\epsilon_i)\mathbf{S}_n^{-1/2}\mathbf{Z}_i\right\| \geq A\tau_0(d+t)^{1/2}\frac{(dt)^{1/2}}{n}\right\} \leq 2(d+2)e^{-t},$$

whenever $n \geq \max\{32L_\infty^4 d^2 t, 2\kappa^{-2}(2d+t)\}$, where $A > 0$ is a constant depending only on M in Condition 4, $C(L_\infty, \delta, v_\delta)$ from Lemma A.4.1.2, and $\tau_0^{-2}\|\Delta_{n,\delta}\|$ with $\Delta_{n,\delta} = n^{-1}\sum_{i=1}^n v_\delta^2 \widetilde{z}_i \widetilde{z}_i^{\mathrm{T}}$.

We conclude this subsection with the counterparts of results in Section A.2.2. From Lemma A.4.1.3, the adaptive Huber regression estimator is expected to be approximated by a Bahadur representation under the fixed design.

Corollary A.4.1. *For $\mathbf{T}_j^\circ$ and its Bahadur representation in* (A.4.1)*, it holds*

$$\|\mathbf{T}_j^\circ - \sigma_{\epsilon,jj}^{-1/2}(\mathbf{s}_j + \mathbf{r}_j)\| \leq A\tau_{0j}(d+t)^{1/2}\frac{dt}{(n\sigma_{\epsilon,jj})^{1/2}}$$

with probability at least $1 - 2(d+2)\exp(-t)$.

The following lemmas show that the distribution of the Bahadur representation in (A.4.1) is close to $N(\mathbf{0}, \sigma_{\epsilon,jj}^2\mathbf{I})$. We decompose $|\,\mathbb{P}(\|\sigma_{\epsilon,jj}^{-1/2}\mathbf{s}_j\| \geq x) - \mathbb{P}(\|\mathbf{G}\| \geq x)|$ into two parts. Lemma A.4.1.4 quantifies the difference between the cumulative distribution functions of $\|\sigma_{\epsilon,jj}^{-1/2}s_j\mathbf{G}\|$ and $\|\mathbf{G}\|$, and Lemma A.4.1.5 quantifies the distinction between the cumulative distribution functions of $\|\sigma_{\epsilon,jj}^{-1/2}s_j\mathbf{G}\|$ and $\|\sigma_{\epsilon,jj}^{-1/2}\mathbf{s}_j\|$. Their proofs are similar to those of Lemmas A.2.2.1-A.2.2.2 and therefore are omitted.

Lemma A.4.1.4. *Let* $\mathbf{G} \sim N(\mathbf{0}, \mathbf{I}) \in \mathbb{R}^q$. *For* $\tau_j = \tau_{0j}(n/t)^{1/(2+\delta)}$ *for some* $\delta \in (0, 2]$ *where* $\tau_{0j} \geq v_{j,\delta}$,

$$\sup_{x \in \mathbb{R}^+} \left| \mathbb{P}(\|\sigma_{\epsilon,jj}^{-1/2} s_j \mathbf{G}\| \geq x) - \mathbb{P}(\|\mathbf{G}\| \geq x) \right| \leq q^{1/2} \frac{v_\delta^{2+\delta}}{\delta \tau_{0j}^\delta \sigma_{\epsilon,jj}} \left(\frac{t}{n}\right)^{\delta/(2+\delta)}.$$

Lemma A.4.1.5. *Let* $\mathbf{G} \sim N(\mathbf{0}, \mathbf{I}) \in \mathbb{R}^q$.

$$\sup_{x \in \mathbb{R}^+} |\, \mathbb{P}(\|\sigma_{\epsilon,jj}^{-1/2} \mathbf{s}_j\| \geq x) - \mathbb{P}(\|\sigma_{\epsilon,jj}^{-1/2} s_j \mathbf{G}\| \geq x)| \precsim n^{-1/2} q^{7/4}.$$

A.5 Additional results from simulation studies

In this section, we report additional numerical results for simulations detailed in Section 2.4 in the main paper. For ease of presentation, we revisit the simulation settings. We generate data from (2.2.1) in the main paper for $n = 85, 120, 150$, $p = 1000, 2000$, $p_1 = 50$, and $d = 6, 8$. We consider three heavy-tailed error distributions:

(a) Pareto distribution with shape parameter 4 and scale parameter 1,

(b) log-normal distribution with $\mu = 0$ and $\sigma = 1$, and

(c) a mixture of the log-normal distribution in (b) and the t_2 distribution with proportion 0.7 and 0.3 respectively.

To incorporate dependence, we set $\mathbf{\Xi} = 100\mathbf{R}_\epsilon^{1/2}\mathbf{E}$, where the correlation matrix $\mathbf{R}_\epsilon$ has one of the following three structures:

- *Model* 1, $\mathbf{R}_\epsilon$ is the identity matrix;
- *Model* 2, $\mathbf{R}_\epsilon = (r_{\epsilon,jk})_{1 \leq j,k \leq p}$ is sparse with $r_{\epsilon,jj} = 1$ and $r_{\epsilon,ij} = r_{\epsilon,ji}$ independently drawn from $0.3 \times \text{Bernoulli}(0.1)$ for $i \neq j$; and
- *Model* 3, $\mathbf{R}_\epsilon = (r_{\epsilon,jk})_{1 \leq j,k \leq p}$ with $r_{\epsilon,jj} = 1$, $r_{\epsilon,j,j+1} = r_{\epsilon,j+1,j} = 0.3$, $r_{\epsilon,j,j+2} = r_{\epsilon,j+2,j} = 0.1$, and $r_{\epsilon,j,j+k} = r_{\epsilon,j+k,j} = 0$ for $k \geq 3$.

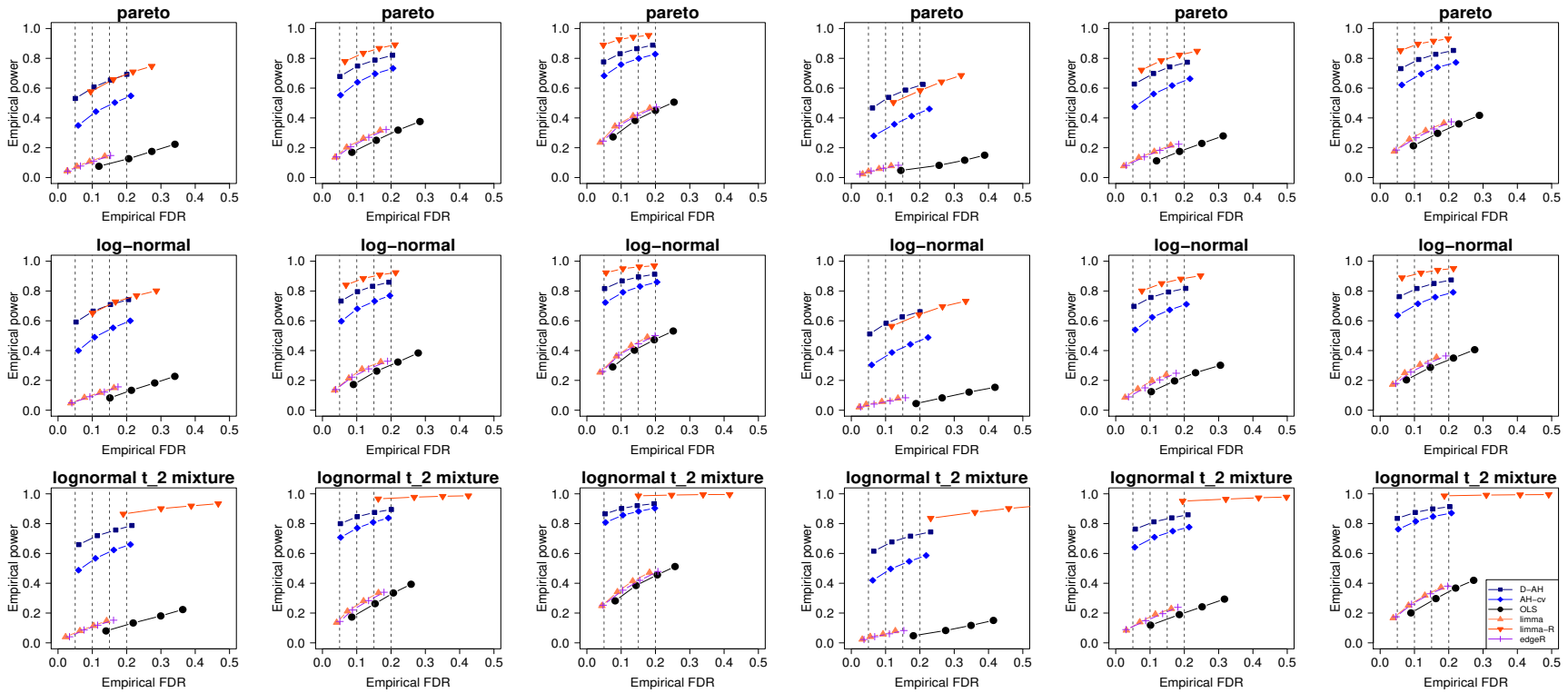
For each $j = 1, \ldots, p$, we set $\mu_j = 5000$ and consider two hypotheses:

- *Hypothesis* 1, $H_{0j} : \mathbf{1}^{\mathrm{T}}\boldsymbol{\beta}_j = 0$ versus $H_{aj} : \mathbf{1}^{\mathrm{T}}\boldsymbol{\beta}_j \neq 0$, where $q = 1$, and
- *Hypothesis* 2, $H_{0j} : \boldsymbol{\beta}_j = 0 \in \mathbb{R}^d$ versus $H_{aj} : \boldsymbol{\beta}_j \neq 0$ $(j = 1, \ldots, p)$, where $q = d$.

For *Hypothesis* 1, we let $\beta_{jk} \sim \mathrm{Unif}(-150, 150)$ for $1 \leq j \leq p$ and $1 \leq k \leq d-1$, $\beta_{jd} = -\sum_{k=1}^{d-1} \beta_{jk}$ for $1 \leq j \leq p - p_1$ so that $\mathbf{1}^{\mathrm{T}}\boldsymbol{\beta}_j = 0$, and $\beta_{jd} = \delta d^{1/2} W_j - \sum_{k=1}^{d-1} \beta_{jk}$ for $p - p_1 + 1 \leq j \leq p$, where W_j are Rademacher random variables. For *Hypothesis* 2, let $\boldsymbol{\beta}_j = 0$ for $1 \leq j \leq p - p_1$, and $\beta_{jk} = (2d^{-1})^{1/2} \delta W_{jk}$ for $p - p_1 + 1 \leq j \leq p$ and $1 \leq k \leq d$, where W_{jk} are Rademacher random variables. We take $\delta = 22.5$ for results in Figures A.1 to A.11.

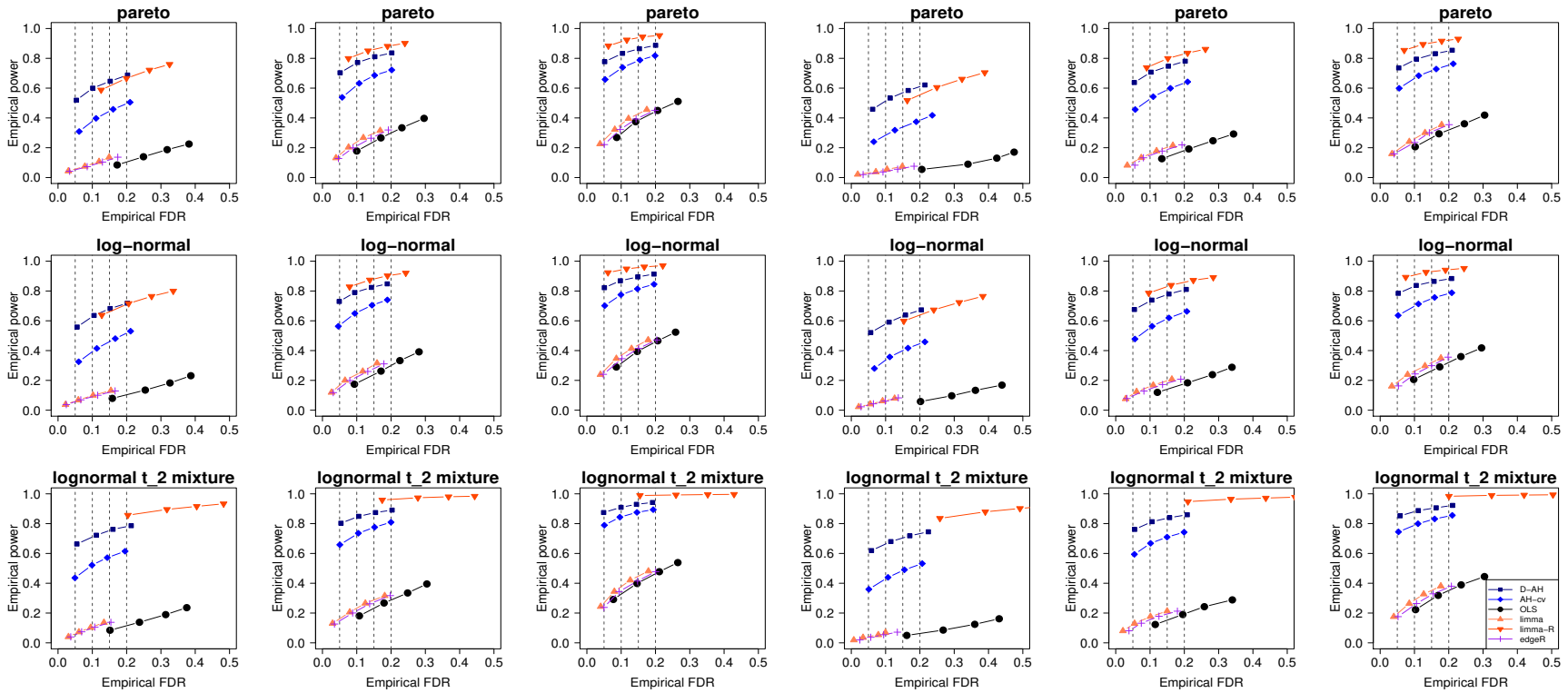
Results for testing different hypotheses under Model 1 are presented in Figures A.1-A.4, and those under Model 3 are depicted in Figures A.8-A.11. The simulation results for Model 2 with different d's are displayed in Figures A.5-A.7. Similar observations to Section 2.4 are made from these extra numerical results. The proposed method that employs data-adaptive Huber regression or selects τ_j via five-fold cross-validation outperforms other competing methods in general with satisfactory control of the empirical false discovery rate and good powers. When n is small and p is large (as $p = 2000$), the control of empirical false discovery rate is challenging for all methods. However, as n increases, our method preserves the nominal level of false discovery rate and is more powerful than `edgeR` and `limma`. Similar observations are made when the dependence is strong (Model 3) and d is large.

Similar to Figure 2.3 in the main paper, Figure A.12 compares the powers of different methods for testing *Hypothesis 1*, the linear contrast, with varying signal strengths as defined in Section 2.4. Similar to Figure 2.3 in the main paper, the proposed method with either adaptive Huber regression or cross validation-selected τ_j's outperforms `limma` and `edgeR` for all error settings.



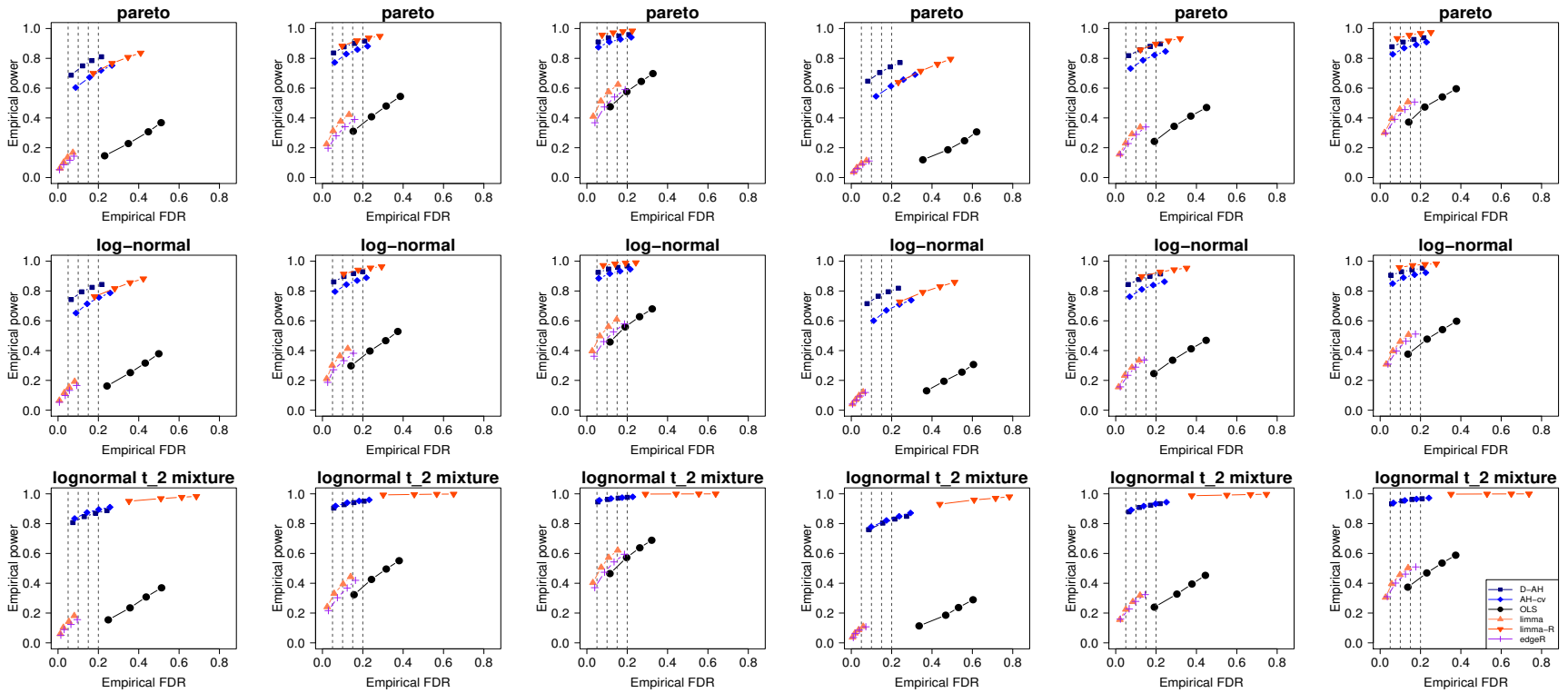
(a) $(n, p) = (85, 1000)$ **(b)** $(n, p) = (120, 1000)$ **(c)** $(n, p) = (150, 1000)$ **(d)** $(n, p) = (85, 2000)$ **(e)** $(n, p) = (120, 2000)$ **(f)** $(n, p) = (150, 2000)$

Figure A.1: Empirical false discovery rate (FDR) and power for testing *Hypothesis* 1, a single contrast, under Model 1 (independent and identically distributed errors) with $d = 6$ by our procedure with the fully data adaptive Huber regression (D-AH, ■); our procedure with the adaptive Huber regression and the five-fold cross-validation (AH-cv, ♦); the ordinary least square estimator (OLS, •); `limma` (▲); `limma` with the robust regression (`limma`-R, ▼); and `edgeR` (+). Each point on the figures displays the empirical false discovery rates and power of the corresponding method at a nominal false discovery rate, which is marked as a vertical gray dashed line. The pre-specified false discovery rates are $0.05, 0.1, 0.15, 0.2$. Error distributions are displayed in the plot captions.



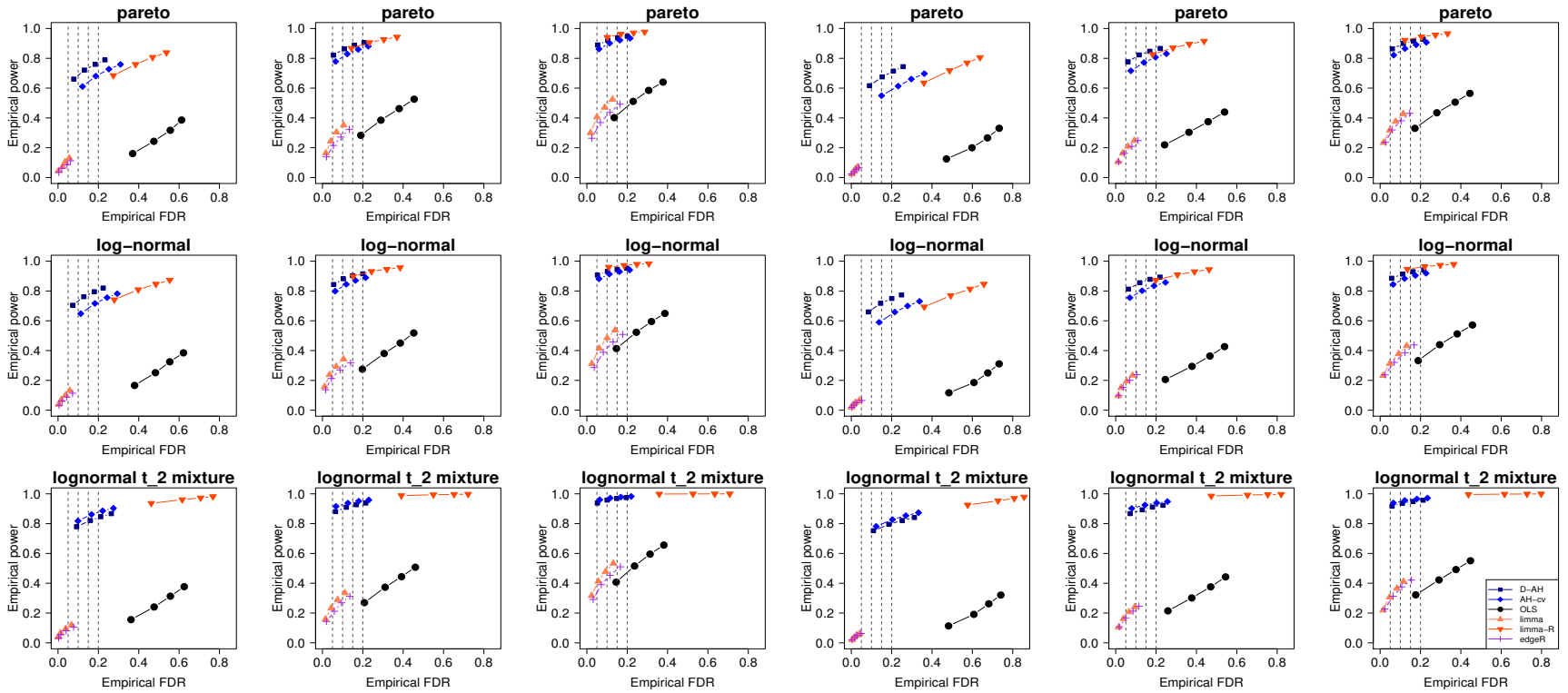
(a) $(n, p) = (85, 1000)$ **(b)** $(n, p) = (120, 1000)$ **(c)** $(n, p) = (150, 1000)$ **(d)** $(n, p) = (85, 2000)$ **(e)** $(n, p) = (120, 2000)$ **(f)** $(n, p) = (150, 2000)$

Figure A.2: Empirical false discovery rate (FDR) and power for testing *Hypothesis* 1, a single contrast, under Model 1 (independent and identically distributed errors) with $d = 8$ by our procedure with the fully data adaptive Huber regression (D-AH, ■); our procedure with the adaptive Huber regression and the five-fold cross-validation (AH-cv, ♦); the ordinary least square estimator (OLS, •); `limma` (▲); `limma` with the robust regression (`limma`-R, ▼); and `edgeR` (+). Each point on the figures displays the empirical false discovery rates and power of the corresponding method at a nominal false discovery rate, which is marked as a vertical gray dashed line. The pre-specified false discovery rates are $0.05, 0.1, 0.15, 0.2$. Error distributions are displayed in the plot captions.



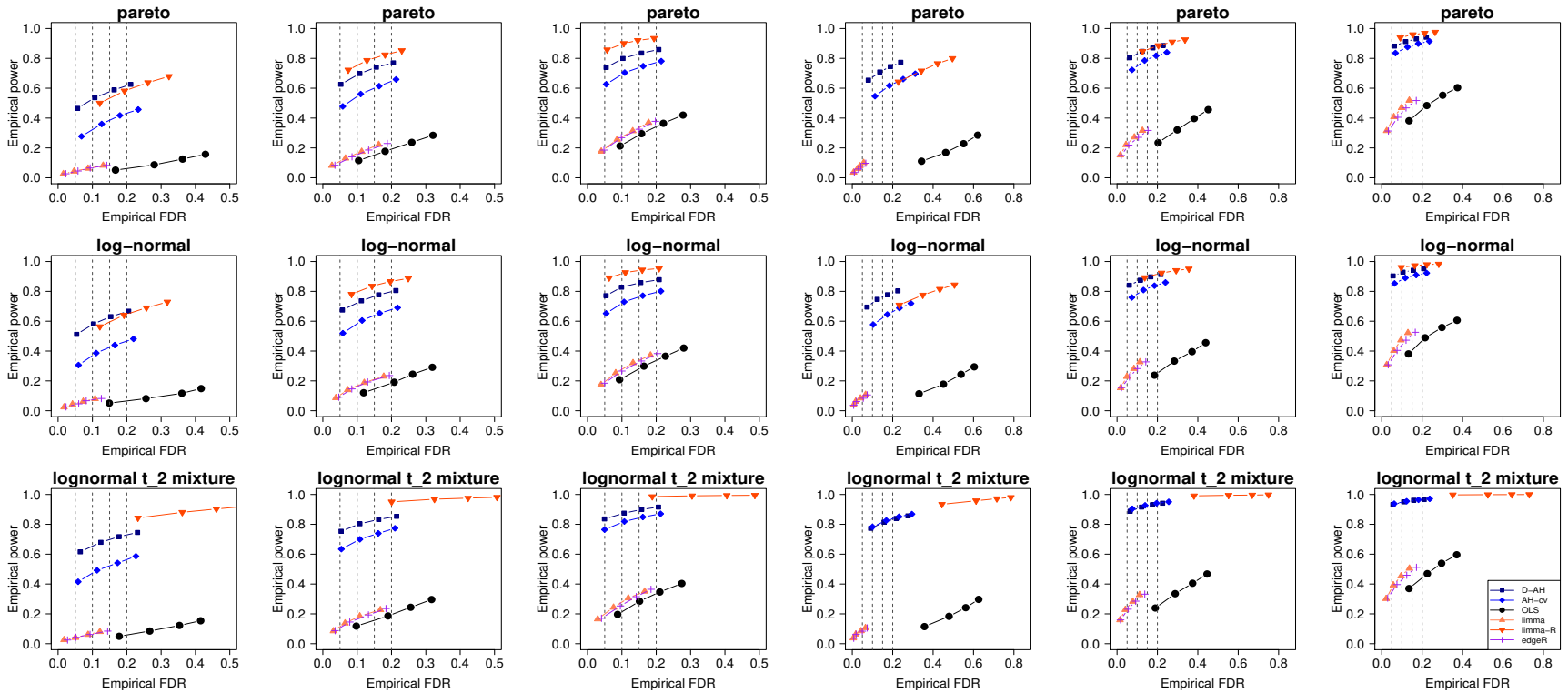
(a) $(n, p) = (85, 1000)$ **(b)** $(n, p) = (120, 1000)$ **(c)** $(n, p) = (150, 1000)$ **(d)** $(n, p) = (85, 2000)$ **(e)** $(n, p) = (120, 2000)$ **(f)** $(n, p) = (150, 2000)$

Figure A.3: Empirical false discovery rate (FDR) and power for testing *Hypothesis* 2 under Model 1 (independent and identically distributed errors) with $d = 6$ by our procedure with the fully data adaptive Huber regression (D-AH, ■); our procedure with the adaptive Huber regression and the five-fold cross-validation (AH-cv, ♦); the ordinary least square estimator (OLS, ●); `limma` (▲); `limma` with the robust regression (`limma`-R, ▼); and `edgeR` (+). Each point on the figures displays the empirical false discovery rates and power of the corresponding method at a nominal false discovery rate, which is marked as a vertical gray dashed line. The pre-specified false discovery rates are $0.05, 0.1, 0.15, 0.2$. Error distributions are displayed in the plot captions.



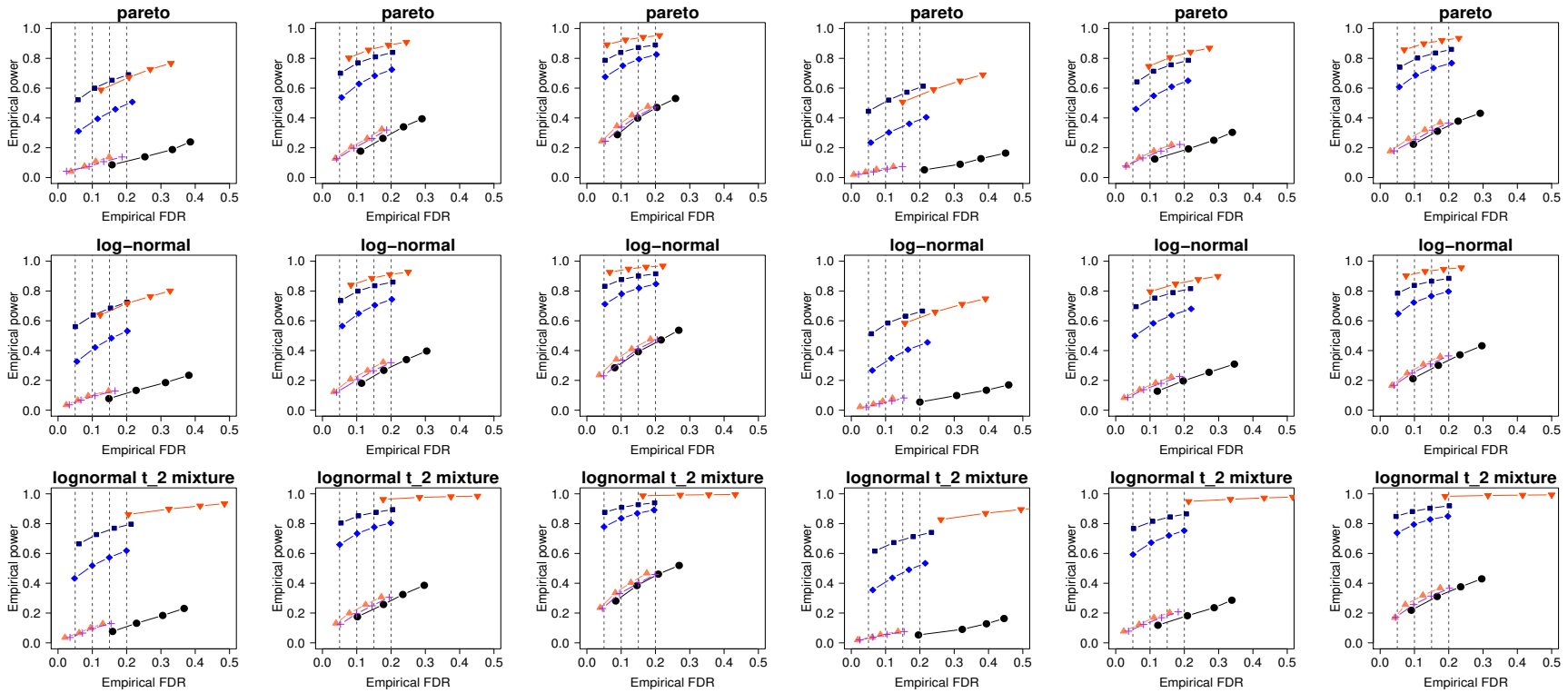
(a) $(n, p) = (85, 1000)$ **(b)** $(n, p) = (120, 1000)$ **(c)** $(n, p) = (150, 1000)$ **(d)** $(n, p) = (85, 2000)$ **(e)** $(n, p) = (120, 2000)$ **(f)** $(n, p) = (150, 2000)$

Figure A.4: Empirical false discovery rate (FDR) and power for testing *Hypothesis* 2 under Model 1 (independent and identically distributed errors) with $d = 8$ by our procedure with the fully data adaptive Huber regression (D-AH, ■); our procedure with the adaptive Huber regression and the five-fold cross-validation (AH-cv, ♦); the ordinary least square estimator (OLS, ●); `limma` (▲); `limma` with the robust regression (`limma`-R, ▼); and `edgeR` (+). Each point on the figures displays the empirical false discovery rates and power of the corresponding method at a nominal false discovery rate, which is marked as a vertical gray dashed line. The pre-specified false discovery rates are $0.05, 0.1, 0.15, 0.2$. Error distributions are displayed in the plot captions.



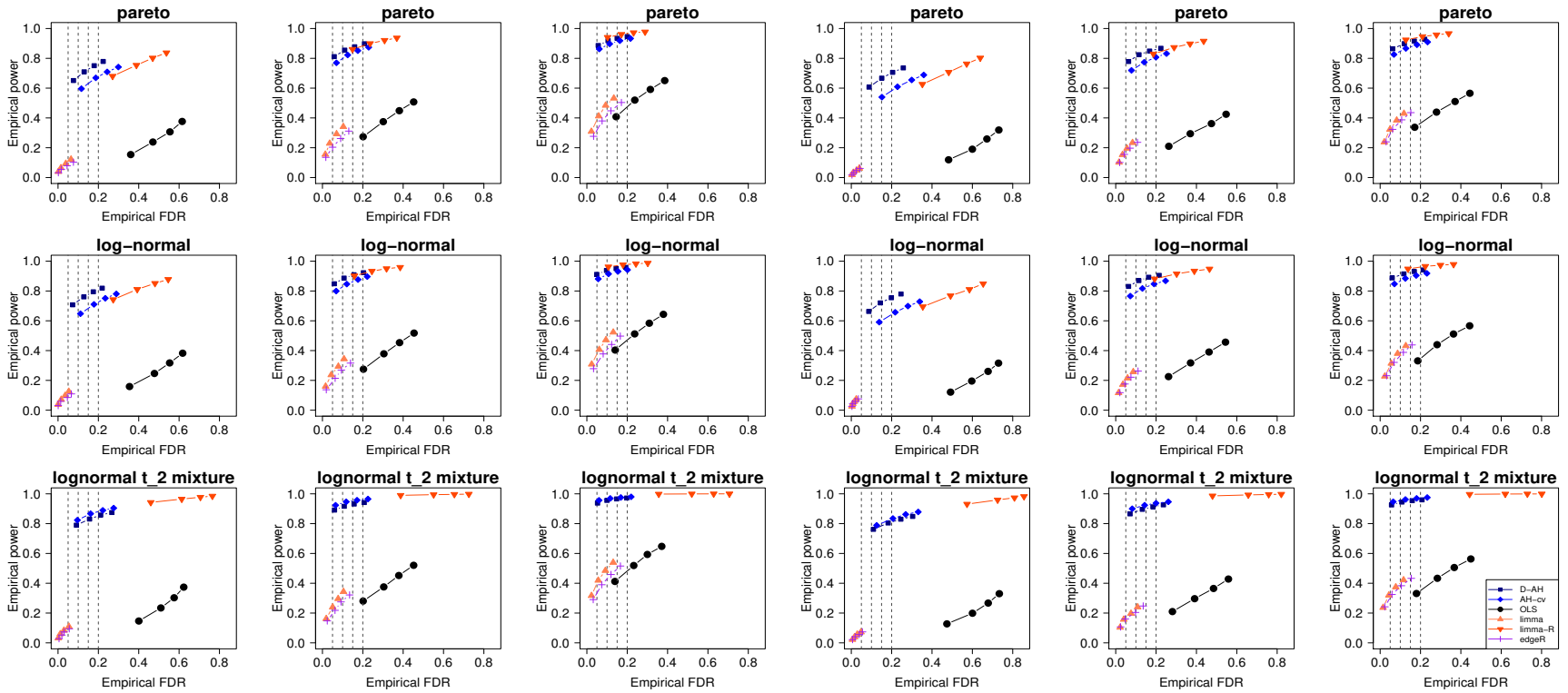
(a) H1, $n = 85$ **(b)** H1, $n = 120$ **(c)** H1, $n = 150$ **(d)** H2, $n = 85$ **(e)** H2, $n = 120$ **(f)** H2, $n = 150$

Figure A.5: Empirical false discovery rate (FDR) and power for testing *Hypothesis* 1 (H1) and *Hypothesis* 2 (H2) under Model 2 (sparsely dependent errors) with $p = 2000$ and $d = 6$ by our procedure with the fully data adaptive Huber regression (D-AH, ■); our procedure with the adaptive Huber regression and the five-fold cross-validation (AH-cv, ♦); the ordinary least square estimator (OLS, •); `limma` (▲); `limma` with the robust regression (`limma`-R, ▼); and `edgeR` (+). Each point on the figures displays the empirical false discovery rates and power of the corresponding method at a nominal false discovery rate, which is marked as a vertical gray dashed line. The pre-specified false discovery rates are $0.05, 0.1, 0.15, 0.2$. Error distributions are displayed in the plot captions.



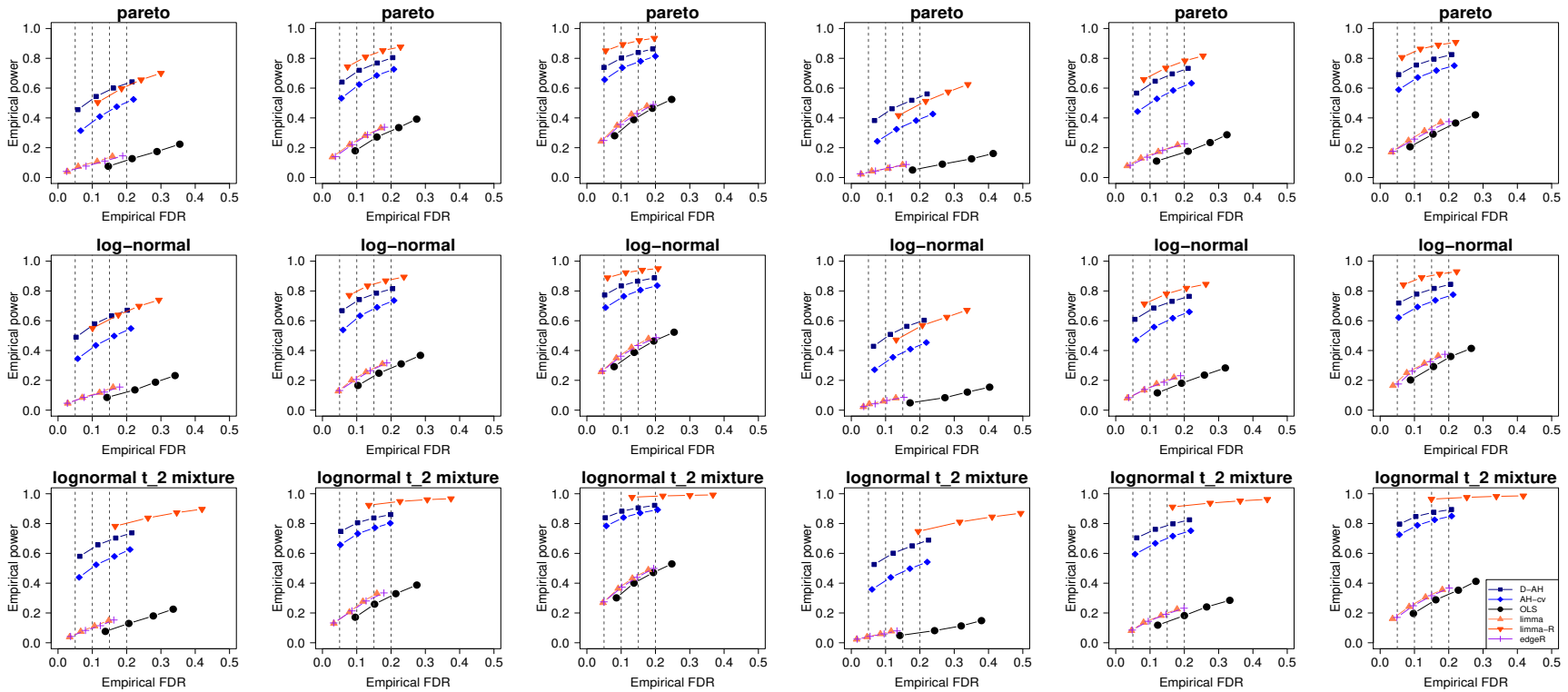
(a) $(n, p) = (85, 1000)$ **(b)** $(n, p) = (120, 1000)$ **(c)** $(n, p) = (150, 1000)$ **(d)** $(n, p) = (85, 2000)$ **(e)** $(n, p) = (120, 2000)$ **(f)** $(n, p) = (150, 2000)$

Figure A.6: Empirical false discovery rate (FDR) and power for testing *Hypothesis* 1, a single contrast, under Model 2 (sparsely dependent errors) with $d = 8$ by our procedure with the fully data adaptive Huber regression (D-AH, ■); our procedure with the adaptive Huber regression and the five-fold cross-validation (AH-cv, ♦); the ordinary least square estimator (OLS, ●); `limma` (▲); `limma` with the robust regression (`limma`-R, ▼); and `edgeR` (+). Each point on the figures displays the empirical false discovery rates and power of the corresponding method at a nominal false discovery rate, which is marked as a vertical gray dashed line. The pre-specified false discovery rates are $0.05, 0.1, 0.15, 0.2$. Error distributions are displayed in the plot captions.



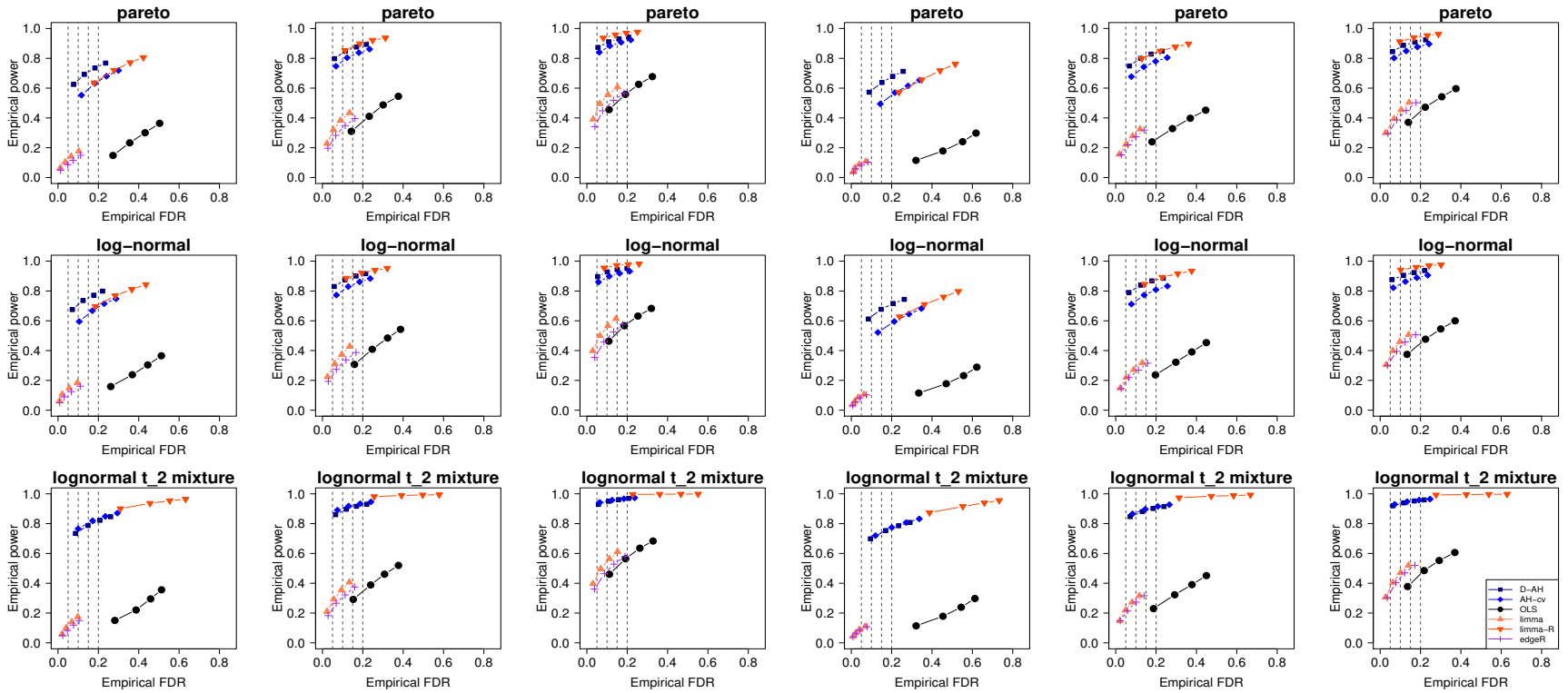
(a) $(n, p) = (85, 1000)$ **(b)** $(n, p) = (120, 1000)$ **(c)** $(n, p) = (150, 1000)$ **(d)** $(n, p) = (85, 2000)$ **(e)** $(n, p) = (120, 2000)$ **(f)** $(n, p) = (150, 2000)$

Figure A.7: Empirical false discovery rate (FDR) and power for testing *Hypothesis* 2 under Model 2 (sparsely dependent errors) with $d = 8$ by our procedure with the fully data adaptive Huber regression (D-AH, ■); our procedure with the adaptive Huber regression and the five-fold cross-validation (AH-cv, ♦); the ordinary least square estimator (OLS, •); `limma` (▲); `limma` with the robust regression (`limma`-R, ▼); and `edgeR` (+). Each point on the figures displays the empirical false discovery rates and power of the corresponding method at a nominal false discovery rate, which is marked as a vertical gray dashed line. The pre-specified false discovery rates are $0.05, 0.1, 0.15, 0.2$. Error distributions are displayed in the plot captions.



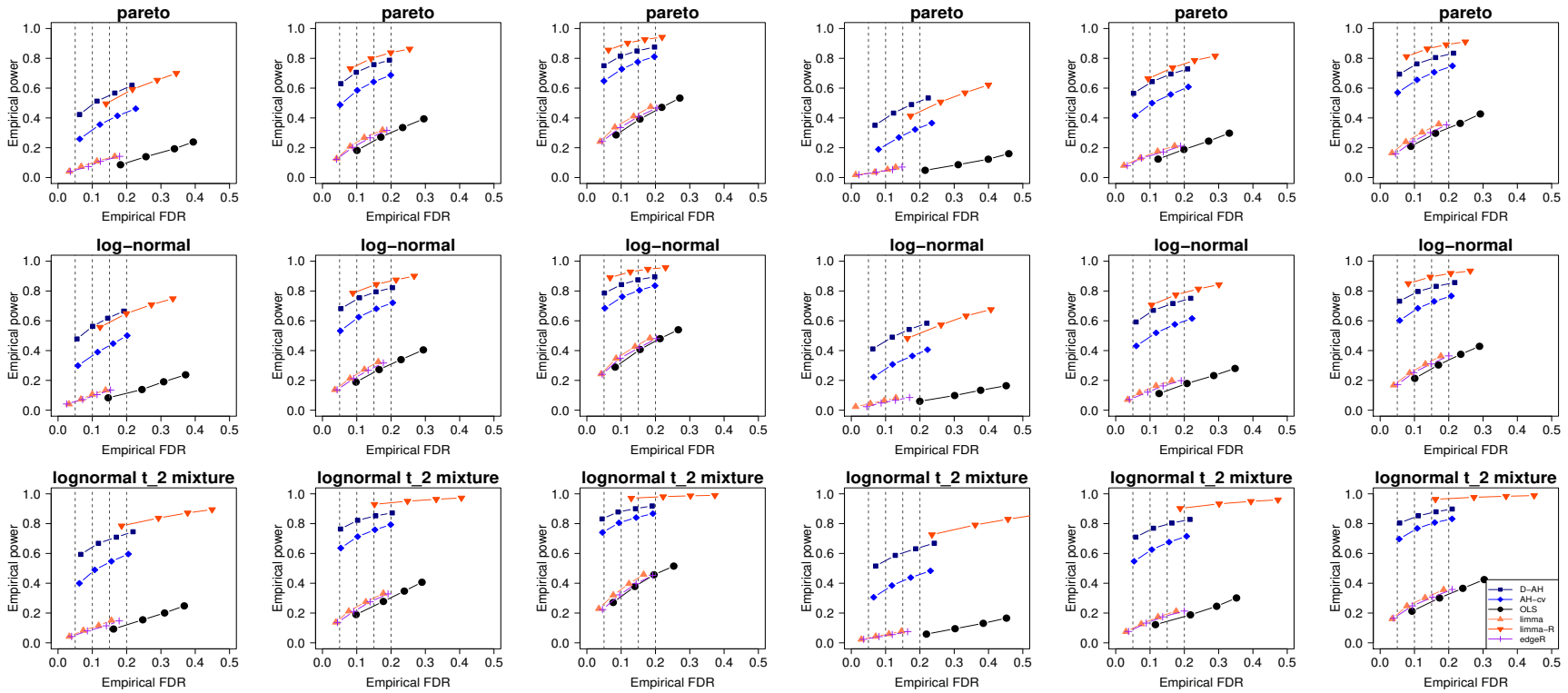
(a) $(n, p) = (85, 1000)$ **(b)** $(n, p) = (120, 1000)$ **(c)** $(n, p) = (150, 1000)$ **(d)** $(n, p) = (85, 2000)$ **(e)** $(n, p) = (120, 2000)$ **(f)** $(n, p) = (150, 2000)$

Figure A.8: Empirical false discovery rate (FDR) and power for testing *Hypothesis* 1, a single contrast, under Model 3 (banding dependence in errors) with $d = 6$ by our procedure with the fully data adaptive Huber regression (D-AH, ■); our procedure with the adaptive Huber regression and the five-fold cross-validation (AH-cv, ♦); the ordinary least square estimator (OLS, ●); `limma` (▲); `limma` with the robust regression (`limma`-R, ▼); and `edgeR` (+). Each point on the figures displays the empirical false discovery rates and power of the corresponding method at a nominal false discovery rate, which is marked as a vertical gray dashed line. The pre-specified false discovery rates are $0.05, 0.1, 0.15, 0.2$. Error distributions are displayed in the plot captions.



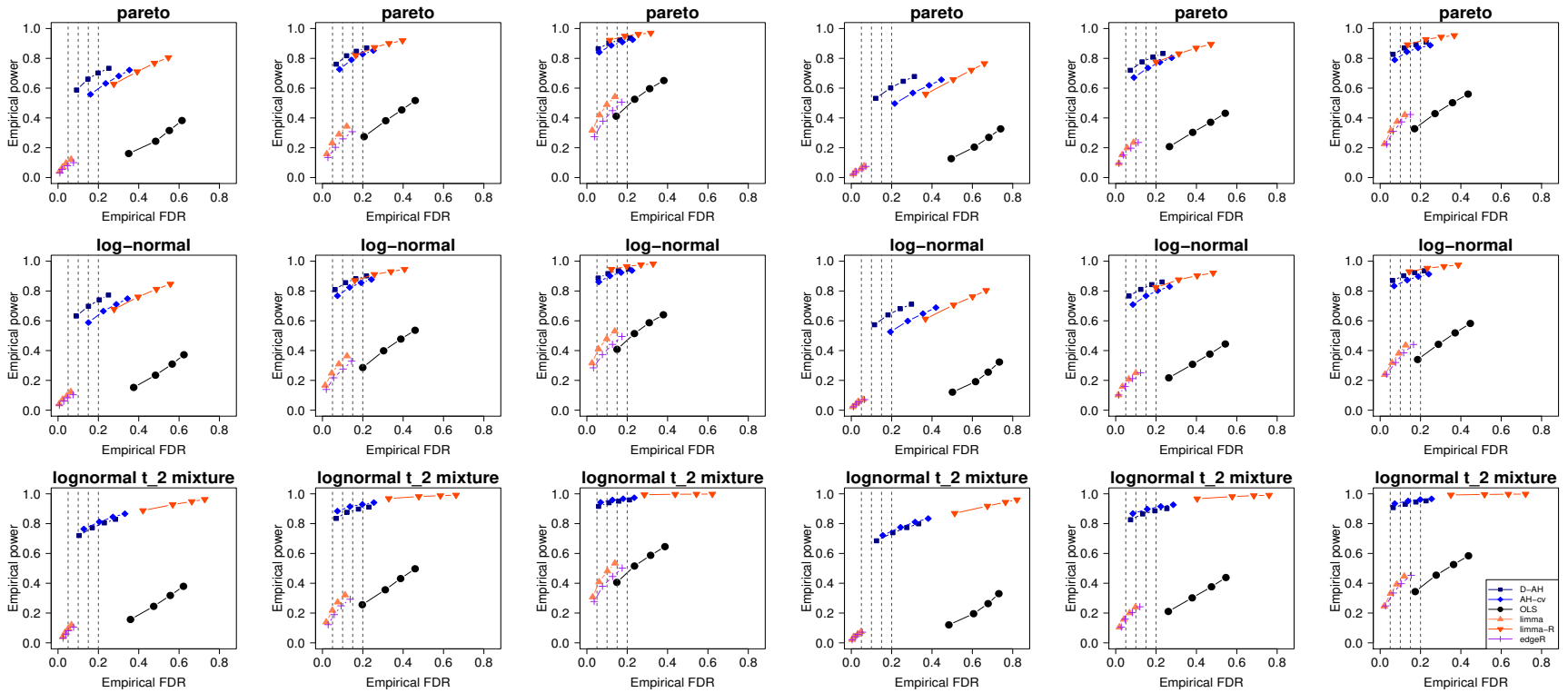
(a) $(n, p) = (85, 1000)$ **(b)** $(n, p) = (120, 1000)$ **(c)** $(n, p) = (150, 1000)$ **(d)** $(n, p) = (85, 2000)$ **(e)** $(n, p) = (120, 2000)$ **(f)** $(n, p) = (150, 2000)$

Figure A.9: Empirical false discovery rate (FDR) and power for testing *Hypothesis* 2 under Model 3 (banding dependence in errors) with $d = 6$ by our procedure with the fully data adaptive Huber regression (D-AH, ■); our procedure with the adaptive Huber regression and the five-fold cross-validation (AH-cv, ⧫); the ordinary least square estimator (OLS, •); `limma` (▲); `limma` with the robust regression (`limma`-R, ▼); and `edgeR` (+). Each point on the figures displays the empirical false discovery rates and power of the corresponding method at a nominal false discovery rate, which is marked as a vertical gray dashed line. The pre-specified false discovery rates are $0.05, 0.1, 0.15, 0.2$. Error distributions are displayed in the plot captions.



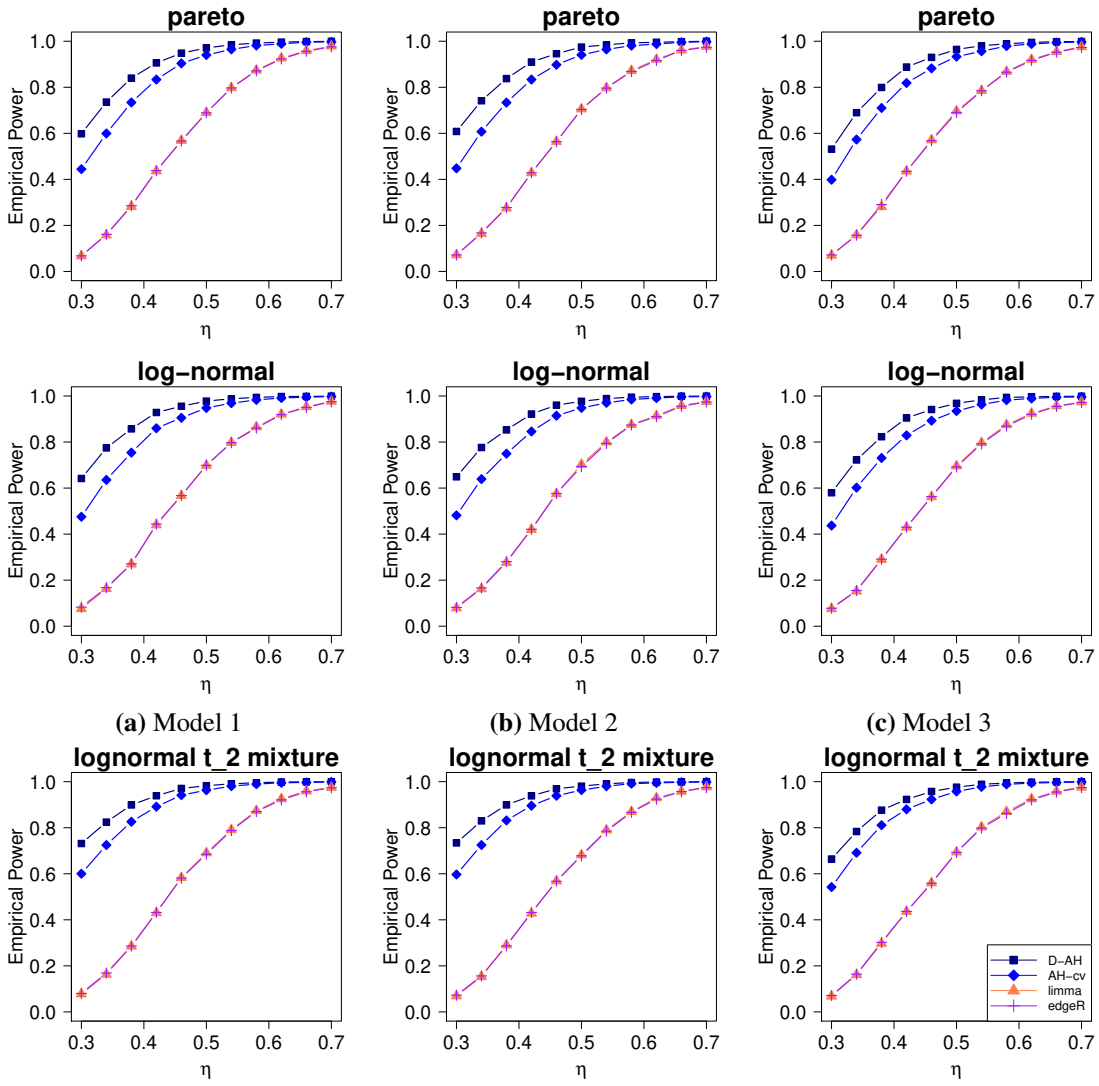
(a) $(n, p) = (85, 1000)$ **(b)** $(n, p) = (120, 1000)$ **(c)** $(n, p) = (150, 1000)$ **(d)** $(n, p) = (85, 2000)$ **(e)** $(n, p) = (120, 2000)$ **(f)** $(n, p) = (150, 2000)$

Figure A.10: Empirical false discovery rate (FDR) and power for testing *Hypothesis* 1, a single contrast, under Model 3 (banding dependence in errors) with $d = 8$ by our procedure with the fully data adaptive Huber regression (D-AH, ■); our procedure with the adaptive Huber regression and the five-fold cross-validation (AH-cv, ♦); the ordinary least square estimator (OLS, ●); `limma` (▲); `limma` with the robust regression (`limma`-R, ▼); and `edgeR` (+). Each point on the figures displays the empirical false discovery rates and power of the corresponding method at a nominal false discovery rate, which is marked as a vertical gray dashed line. The pre-specified false discovery rates are $0.05, 0.1, 0.15, 0.2$. Error distributions are displayed in the plot captions.



(a) $(n, p) = (85, 1000)$ **(b)** $(n, p) = (120, 1000)$ **(c)** $(n, p) = (150, 1000)$ **(d)** $(n, p) = (85, 2000)$ **(e)** $(n, p) = (120, 2000)$ **(f)** $(n, p) = (150, 2000)$

Figure A.11: Empirical false discovery rate (FDR) and power for testing *Hypothesis* 2 under Model 3 (banding dependence in errors) with $d = 8$ by our procedure with the fully data adaptive Huber regression (D-AH, ■); our procedure with the adaptive Huber regression and the five-fold cross-validation (AH-cv, ♦); the ordinary least square estimator (OLS, •); `limma` (▲); `limma` with the robust regression (`limma`-R, ▼); and `edgeR` (+). Each point on the figures displays the empirical false discovery rates and power of the corresponding method at a nominal false discovery rate, which is marked as a vertical gray dashed line. The pre-specified false discovery rates are $0.05, 0.1, 0.15, 0.2$. Error distributions are displayed in the plot captions.



(a) Model 1 **(b)** Model 2 **(c)** Model 3

Figure A.12: Empirical powers for testing *Hypothesis* 1, a single contrast, with $\eta = \{0.30, 0.34, \ldots, 0.66, 0.7\}$, $n = 100$, $d = 6$, and $p = 1000$ by our procedure with the fully data adaptive Huber regression (D-AH, ■); our procedure with the adaptive Huber regression and the five-fold cross-validation (AH-cv, ⧫); `limma` (▲); and `edgeR` (+). Columns (a)-(c) are results for Models 1-3, respectively.

A.6 Additional results for the analysis on Project Gutenberg

In this section, we present addition details for analyzing data from the Standardized Project Gutenberg Corpus (SPGC) as described in Section 2.5 in the main article. Table A.1 displays a

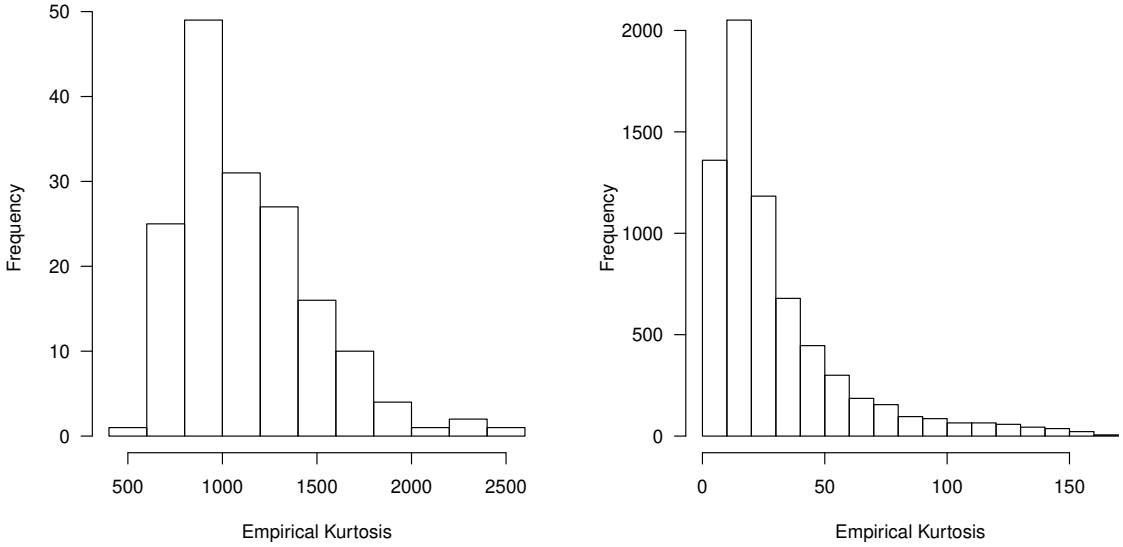
Table A.1: Snapshot of the raw data in SPGC. PG id's represent different books: *Alice's Adventures in Wonderland* (PG19033) by Lewis Carroll, *Oliver Twist* (PG730), *Great Expectations* (PG1400), and *A Christmas Carol* (PG24022) by Charles Dickens, and *A Study in Scarlet* (PG244), *The Sign of the Four* (PG2097), and *The Hound of the Baskervilles* (PG2852) by Arthur Conan Doyle.

word/PG id	PG19033	PG730	PG1400	PG24022	PG244	PG2097	PG2852
the	636	9493	8143	1595	2569	2335	3330
and	337	5239	7071	1046	1368	1179	1628
of	201	3852	4433	696	1215	1122	1594
to	249	3852	5071	676	1093	1079	1408
a	277	3702	4040	709	1005	1092	1306
i	160	1357	6632	280	938	1219	1497
jolly	0	7	20	3	0	0	0
king	25	4	14	0	0	0	0
loss	0	20	12	1	4	3	4
colour	0	5	0	4	5	0	6
oliver	0	859	0	0	0	0	1
shaded	0	0	5	0	0	2	1
alice	172	0	0	0	5	0	0
murder	1	19	19	0	18	10	6
christmas	0	1	9	85	0	0	0

snapshot of the raw word count data. The empirical kurtosis of the normalized data is reported in Figure A.13, which provides the evidence of heavy tailedness of the data.

The word counts displayed in Table A.1 agrees with the Zipf's law [70], that is the frequency of a word in a corpus is inversely proportional to its rank in the frequency table. A few topic-related words or proper nouns are more frequently encountered in certain works. For example, *A Christmas Carol* has more "Christmas" than other books, and *Oliver Twist* has a substantially higher frequency of "Oliver" than others. In addition, the raw word count data matrix is sparse and consists of 62751 unique English words. Most of the words have zero counts, 89% of them are removed by the filtering process in Section 2.5 in the main paper accordingly. Upon filtering, 51% of all the entries in the normalized count matrix are zero, and 82% of them are below 5.

Figure A.14 (a) displays a hierarchical clustering result for 23 authors from U.K. and U.S. in the original SPGC data. We observe that Charles Darwin and Thomas H. Huxley were closely related and, as a matter of fact, they are both English biologists in the nineteenth century who focused on the evolution theory. Hence, in terms of the word count distributions, their writings



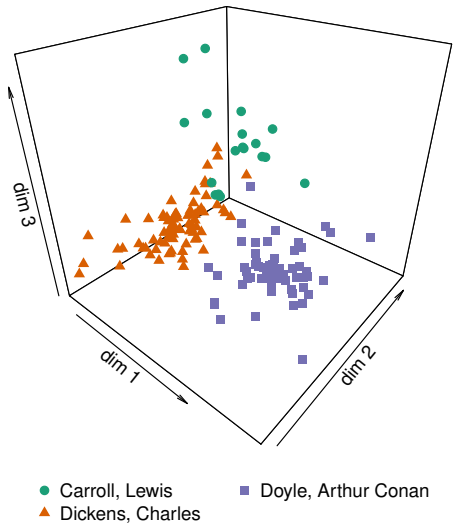
(a) The empirical kurtosis by books. **(b)** The empirical kurtosis by words.

Figure A.13: The empirical kurtosis of words counts for 167 books (panel (a)) and the empirical kurtosis of counts for 6839 words (panel (b)) from the works of Lewis Carroll, Charles Dickens, and Arthur Conan Doyle. The normalized counts are used.

are more similar to each other and distinguishable compared to other authors. In addition, Lewis Carroll, Arthur Conan Doyle, and Charles Dickens are closely related from Figure A.14 (a). From Figure A.14 (b), we notice that the works among Lewis Carroll, Arthur Conan Doyle, and Charles Dickens are separated in general.

The Venn diagram in Figure A.15 displays the number of differentially represented words for hypotheses considered in the first application in Section 2.5 in the main paper. For example, Dickens has 949 differentially represented words that distinguish him from the other two authors. Among those 949 words, “catch" and “curious" appear to be the most significant whereas “clock", “horseback", and “present" are the least significant ones. Further quantitative linguistic or literature investigations are required to uncover more insights on these identified differentially represented words.

(a) Hierarchical clustering for splitting 23 authors into four groups. Each color represent a group.



(b) Multidimensional scaling plot of works by Lewis Carroll, Arthur Conan Doyle, and Charles Dickens. Points represent books and symbols stands for authors. Data are post-processing.

Figure A.14: Exploratory displays of the data.

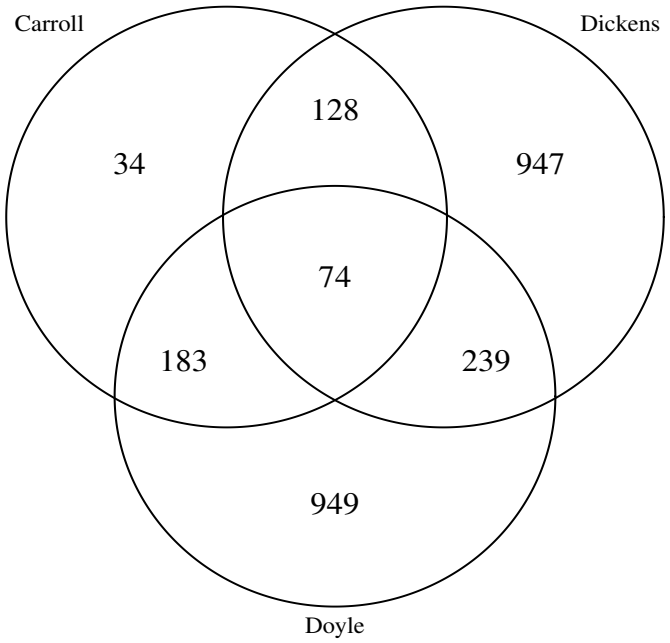


Figure A.15: Comparing word counts of books of Lewis Carroll, Charles Dickens, and Arthur Conan Doyle by our method with the nominal false discovery rate controlled at 0.5%. The Venn diagram displays the number of differentially represented words for Hypothesis CDD2 (Carroll), Hypothesis CDD3 (Dickens), and Hypothesis CDD4 (Doyle).

Appendix B

Supplementary materials for Chapter 4

B.1 Derivation of DRAGON

In this subsection, we provide the derivation of (4.2.23). We first check the relationship between (4.2.14) and (4.2.22). This argument is also employed to verify the relationship between scaled Lasso and square-root Lasso [102].

Lemma B.1.0.1. (4.2.14) *has the same* $\widehat{\boldsymbol{\beta}}_j$ *as* (4.2.22) *when* $\sigma_j = n^{-1/2}\|\mathbf{X}_j - \mathbf{X}_{-j}\widehat{\boldsymbol{\beta}}_j\|_2$.

Proof. It holds by the inequality of arithmetic-geometric mean:

$$\frac{\|\mathbf{X}_j - \mathbf{X}_{-j}\boldsymbol{\beta}_j\|_2^2}{2n\sigma_j} + \frac{\sigma_j}{2} \geq \frac{\|\mathbf{X}_j - \mathbf{X}_{-j}\boldsymbol{\beta}_j\|_2}{n},$$

where the equality holds when $\sigma_j^2 = n^{-1}\|\mathbf{X}_j - \mathbf{X}_{-j}\boldsymbol{\beta}_j\|_2^2$. Hence, $\sigma_j = n^{-1/2}\|\mathbf{X}_j - \mathbf{X}_{-j}\widehat{\boldsymbol{\beta}}_j\|_2$ holds for the common solution $\widehat{\boldsymbol{\beta}}_j$. □

Hence, we can use (4.2.22) for the algorithm derivation, which provides simplicity in algebra.

Following that, we derive (4.2.23) from (4.2.22) by using similar argument in [147]. By the Karush-Kuhn-Tucker condition, for $k \neq j$,

$$-\frac{1}{n\sigma_j}\langle \mathbf{X}_k, \mathbf{X}_j - \mathbf{X}_{-j}\boldsymbol{\beta}_j\rangle + \lambda\,\mathrm{sgn}(\beta_{j,k}) + \rho\frac{\beta_{j,k}}{\|(1, -\boldsymbol{\beta}_j)\|_2} = 0,\ (\boldsymbol{\beta} \neq \mathbf{0}) \tag{B.1.1}$$

where $\partial|\beta_{j,k}|$ is a sub-gradient of $|\beta_{j,k}|$.

If $\beta_{j,k} \neq 0$, it holds that

$$-\frac{1}{n\sigma_j}\langle \mathbf{X}_k, \mathbf{X}_j - \mathbf{X}_{-j}\boldsymbol{\beta}_j\rangle + \lambda\,\mathrm{sgn}(\beta_{j,k}) + \rho\frac{\beta_{j,k}}{\|(1, -\boldsymbol{\beta}_j)\|_2} = 0,$$

then it holds that

$$\left(\frac{\|\mathbf{X}_k\|_2^2}{n\sigma_j}+\frac{\lambda}{|\beta_{j,k}|}+\frac{\rho}{\|(1,-\boldsymbol{\beta}_j)\|_2}\right)\beta_{j,k}=\frac{1}{n\sigma_j}\langle\mathbf{X}_k,\mathbf{X}_j-\sum_{l=1,l\neq j}^{p}\beta_{j,l}\mathbf{X}_l\rangle. \tag{B.1.2}$$

By taking the absolute value of the second equation above, we obtain

$$|\beta_{j,k}|=\left(\frac{\|\mathbf{X}_k\|_2^2}{n\sigma_j}+\frac{\rho}{\|(1,-\boldsymbol{\beta}_j)\|_2}\right)^{-1}\left(\frac{1}{n\sigma_j}\langle\mathbf{X}_k,\mathbf{X}_j-\sum_{l=1,l\neq j}^{p}\beta_{j,l}\mathbf{X}_l\rangle-\lambda\right)_+. \tag{B.1.3}$$

where $(x)_+=\max(0,x)$. Plugging (B.1.3) into (B.1.2), we have

$$\beta_{j,k}=\left(\frac{\|\mathbf{X}_k\|_2^2}{n\sigma_j}+\frac{\rho}{\|(1,-\boldsymbol{\beta}_j)\|_2}\right)^{-1}S_\lambda\left(\frac{1}{n\sigma_j}\langle\mathbf{X}_k,\mathbf{X}_j-\sum_{l=1,l\neq j}^{p}\beta_{j,l}\mathbf{X}_l\rangle\right). \tag{B.1.4}$$

We conclude this subsection with code implementation. Our implementation is based on `RcppArmadillo`, which allows our implementation is faster than pure Rimplementation. Given $\boldsymbol{\beta}_j^{(t-1)}$, we compute and store $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ in the initialization step and $\|(1,-\boldsymbol{\beta}_j^{(t-1)})\|_2$ for updating $\beta_{j,k}^{(t)}$ to reduce computation time.

B.2 Proofs of main theorems

B.2.1 Proof of Proposition 4.2.4

When $\rho=0$, it holds by the relationship between LASSO and square-root LASSO [168].

Assuming $\rho>0$. For notation simplicity, we use $\widetilde{\boldsymbol{\beta}}:=(1,-\boldsymbol{\beta})$. Let λ^* is the tuning parameter of (4.2.13). The KKT condition of (4.2.13) is

$$\begin{aligned}
&2\left(\frac{1}{\sqrt{n}}\|\mathbf{X}\widetilde{\boldsymbol{\beta}}\|_2+\rho\|\widetilde{\boldsymbol{\beta}}\|_2\right)\left(-\frac{1}{\sqrt{n}}\frac{\mathbf{X}_{-j}^{\mathrm{T}}\mathbf{X}\widetilde{\boldsymbol{\beta}}}{\|\mathbf{X}\widetilde{\boldsymbol{\beta}}\|_2}+\rho\frac{\boldsymbol{\beta}}{\|\widetilde{\boldsymbol{\beta}}\|_2}\right)+\lambda^*\nabla\|\boldsymbol{\beta}\|_1=\mathbf{0};\\
&\|\boldsymbol{\beta}\|_1\leq c \text{ for some } c>0;\\
&\lambda^*(\|\boldsymbol{\beta}\|_1-c)=0\\
&\lambda^*\geq 0.
\end{aligned} \tag{B.2.1}$$

Let λ^* is the tuning parameter of (4.2.13). The KKT condition of (4.2.14) is

$$
\begin{aligned}
&-\frac{1}{\sqrt{n}}\frac{\mathbf{X}_{-j}^{\mathrm{T}}\mathbf{X}\widetilde{\boldsymbol{\beta}}}{\|\mathbf{X}\widetilde{\boldsymbol{\beta}}\|_2}+\rho\frac{\boldsymbol{\beta}}{\|\widetilde{\boldsymbol{\beta}}\|_2}+\lambda^{**}\nabla\|\boldsymbol{\beta}\|_1=\mathbf{0};\\
&\|\boldsymbol{\beta}\|_1\leq c \text{ for some } c>0;\\
&\lambda^{**}(\|\boldsymbol{\beta}\|_1-c)=0\\
&\lambda^{**}\geq 0.
\end{aligned}
\tag{B.2.2}
$$

By Lemma B.3.2.2, for the unique solution of (4.2.13) $\widehat{\boldsymbol{\beta}}$, the two optimization problems have the same solution when

$$
\lambda^{**}=\left(\frac{1}{\sqrt{n}}\|\mathbf{X}_j-\mathbf{X}_{-j}\widehat{\boldsymbol{\beta}}\|_2+\rho\|(1,-\widehat{\boldsymbol{\beta}})\|_2\right)^{-1}\frac{\lambda^*}{2}
$$

holds.

B.2.2 Proof of Proposition 4.2.5

Let $\widehat{\sigma}_j=n^{-1/2}\|\mathbf{X}_j-\mathbf{X}_{-j}\boldsymbol{\beta}\|_2$ and $\widehat{\gamma}_{jk}:=n^{-1}\langle\mathbf{X}_k,\mathbf{X}_j-\mathbf{X}_{-j}\boldsymbol{\beta}\rangle$. i) and ii) can be proved by using the argument of Lemmas 4.2.3.1 and 4.2.3.2. We provide the proof of i) only for readers' convenience. For $\rho>0$ and $\widehat{\beta}_{j,k}\neq 0$, the solution of the optimization problem (4.2.14) satisfies

$$
\begin{aligned}
&-\frac{\widehat{\gamma}_{jk}}{\widehat{\sigma}_j}+\rho\frac{\widehat{\beta}_{jk}}{\|(1,-\widehat{\boldsymbol{\beta}}_j)\|_2}+\lambda\operatorname{sgn}(\widehat{\beta}_{jk})=0,\\
&\rho=\frac{\|(1,-\widehat{\boldsymbol{\beta}}_j)\|_2}{\widehat{\beta}_{jk}}\left(\frac{\widehat{\gamma}_{jk}}{\widehat{\sigma}_j}-\lambda\operatorname{sgn}(\widehat{\beta}_{jk})\right).
\end{aligned}
$$

This implies $\operatorname{sgn}\{\widehat{\gamma}_{jk}\widehat{\sigma}_j^{-1}-\lambda\operatorname{sgn}(\widehat{\beta}_{jk})\}=\operatorname{sgn}(\widehat{\beta}_{jk})$. Then, the partial derivative,

$$
\begin{aligned}
\frac{\partial\rho}{\partial\widehat{\beta}_{jk}}=&\frac{\widehat{\beta}_{jk}^2-\|(1,-\widehat{\boldsymbol{\beta}}_j)\|_2^2}{\widehat{\beta}_{jk}^2\|(1,-\widehat{\boldsymbol{\beta}}_j)\|_2}\left(\frac{\widehat{\gamma}_{jk}}{\widehat{\sigma}_j}-\lambda\operatorname{sgn}(\widehat{\beta}_{jk})\right)\\
&-\frac{\|(1,-\widehat{\boldsymbol{\beta}}_j)\|_2(n^{-1}\|\mathbf{X}_k\|_2^2\widehat{\sigma}_j^2-\widehat{\gamma}_{jk}^2)}{\widehat{\sigma}_j^3}\frac{1}{\widehat{\beta}_{jk}},
\end{aligned}
$$

has the same sign as $-\widehat{\beta}_{jk}$ since $\|n^{-1/2}\mathbf{X}_k\|_2^2\widehat{\sigma}_j^2 \geq \widehat{\gamma}_{jk}^2$ by Cauchy-Schwartz inequality and $\widehat{\beta}_{jk}^2 - \|(1, -\widehat{\boldsymbol{\beta}}_j)\|_2^2 < 0$. This proves i).

B.3 Auxiliary results

B.3.1 Proofs of Lemmas for the bivariate example

In this subsection, we provide proofs for Lemmas 4.2.3.1 and 4.2.3.2 from the bivariate example to study properties of (4.2.14). We first present the proof of Lemma 4.2.3.1, which is about the population version optimization problem (4.2.19).

proof of Lemma 4.2.3.1. The minimizer $\beta_{*,\rho}$ satisfies

$$\frac{\beta_{*,\rho} - r}{\{1 - r^2 + (\beta_{*,\rho} - r)^2\}^{1/2}} + \rho\frac{\beta_{*,\rho}}{(1 + \beta_*^2)^{1/2}} = 0. \tag{B.3.1}$$

When $\rho = 0$, $\beta_{*,0} = r$. For the rest of the proof, we assume $\rho > 0$. If $\beta_{*,\rho} > 0$, then $\beta_{*,\rho} - r < 0$. If $\beta_{*,\rho} < 0$, then $\beta_{*,\rho} - r > 0$. This proves (i) and (ii). When $\rho = 1$, it holds that $r^2\beta_{*,1}^2 - 2r\beta_{*,1} + r^2 = 0$. Combining this with (ii), we prove (iii).

From (B.3.1), we obtain

$$\begin{aligned}
\rho &= \frac{r - \beta_{*,\rho}}{\beta_{*,\rho}}\left\{\frac{1 + \beta_{*,\rho}^2}{1 - r^2 + (\beta_{*,\rho} - r)^2}\right\}^{1/2}, \\
\frac{\partial\rho}{\partial\beta_{*,\rho}} &= \frac{1}{\beta_{*,\rho}^2\{1 - r^2 + (\beta_{*,\rho} - r)^2\}^2}\left\{\frac{1 - r^2 + (\beta_{*,\rho} - r)^2}{1 + \beta_{*,\rho}^2}\right\}^{1/2} f(\beta_{*,\rho}), \\
f(\beta_{*,\rho}) &:= -(r - \beta_{*,\rho})\{1 - r^2 + (\beta_{*,\rho} - r)^2\} - \beta_{*,\rho}(1 - r^2)(1 + \beta_{*,\rho}^2).
\end{aligned} \tag{B.3.2}$$

Since the sign of $f(\beta_{*,\rho})$ is opposite to the one of $\beta_{*,\rho}$, and it has the same sign as $\partial\rho/\partial\beta_{*,\rho}$ and $\partial\beta_{*,\rho}/\partial\rho$. This proves (iv).

By (iv), $|\beta_{*,\rho}|$ shrinks as ρ increases. If $\beta_{*,\rho} \to c \neq 0$ as $\rho \to \infty$, the first equation in (B.3.2) does not hold. Hence, $\beta_{*,\rho} \to 0$ as $\rho \to \infty$. □

Following this, we present the proof of Lemma 4.2.3.2, which is the empirical version of (4.2.19).

proof of Lemma 4.2.3.2. Let $\widehat{\sigma}^2 = n^{-1}\sum_{i=1}^n (X_{1i} - X_{2i}\beta)^2$ and $\widehat{\gamma} := n^{-1}\sum_{i=1}^n \{(X_{1i} - X_{2i}\beta)X_{2i}\}$. The minimizer satisfies

$$-\frac{\widehat{\gamma}}{\widehat{\sigma}} + \rho\frac{\beta}{(1+\beta^2)^{1/2}} = 0. \tag{B.3.3}$$

It is easy to prove (i) and (ii) from (B.3.3). $\widehat{\beta}_0 = \widehat{\beta}_{OLS}$ is straightforward. ii) is proved by the similar argument as Lemma 4.2.3.1.

To prove (iii) and (iv), we rearrange (B.3.3),

$$\rho = \frac{(1+\beta^2)^{1/2}}{\widehat{\sigma}}\frac{\widehat{\gamma}}{\beta}. \tag{B.3.4}$$

For $\rho > 0$, (B.3.4) implies the followings:

$$\begin{aligned}
\operatorname{sgn}(\beta) =& \operatorname{sgn}(\widehat{\gamma}), \\
\frac{\partial\rho}{\partial\beta} =& \frac{1}{\widehat{\sigma}\beta^2(1+\beta^2)^{1/2}}\left\{-\widehat{\gamma} - (1+\beta^2)\left(\frac{1}{n}\sum_{i=1}^n X_{2i}^2 - \frac{\widehat{\gamma}^2}{\widehat{\sigma}^2}\right)\beta\right\}.
\end{aligned}$$

Note that $\widehat{\sigma}^2 n^{-1}\sum X_{2i}^2 \geq \widehat{\gamma}^2$ by Cauchy-Schwartz inequality. This proves (iii) by $\operatorname{sgn}(\partial\beta/\partial\rho) = \operatorname{sgn}(\partial\rho/\partial\beta) = -\operatorname{sgn}(\beta)$. We can prove (iv) using the same argument as Lemma 4.2.3.1. □

B.3.2 Technical lemmas for Proposition 4.2.4

We provide lemmas as building blocks for the main proofs.

Lemma B.3.2.1. *For $\widetilde{\boldsymbol{\beta}} = (1, -\boldsymbol{\beta}) \in \mathbb{R}^p$, it follows that*

1. *$\|\boldsymbol{\beta}\|_1$ is convex but not strictly convex;*
2. *$\|\widetilde{\boldsymbol{\beta}}\|_2$ is strictly convex;*
3. *$\|\mathbf{X}_j - \mathbf{X}_{-j}\boldsymbol{\beta}\|_2 = \|\mathbf{X}\widetilde{\boldsymbol{\beta}}\|_2$ is convex;*

for all $\boldsymbol{\beta}$.

Proof. Through out the proofs, set $\alpha \in [0,1]$ and $\boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_2$.

By the definition of vector norms, it is straightforward that $\|\boldsymbol{\beta}\|_1$ is convex. The equality $\alpha\|\boldsymbol{\beta}_1\|_1 + (1-\alpha)\|\boldsymbol{\beta}_2\|_1 = \|\alpha\boldsymbol{\beta}_1 + (1-\alpha)\boldsymbol{\beta}_2\|_1$ holds when $\boldsymbol{\beta}_1 = c\boldsymbol{\beta}_2$ for any $c \in \mathbb{R}^+$, which proves the first one.

For $\boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_2$, it follows that

$$\begin{aligned}\|\alpha(1,-\boldsymbol{\beta}_1) + (1-\alpha)(1,-\boldsymbol{\beta}_2)\|_2 &\leq \|\alpha(1,-\boldsymbol{\beta}_1)\|_2 + \|(1-\alpha)(1,-\boldsymbol{\beta}_2)\|_2 \\ &= \alpha\|\widetilde{\boldsymbol{\beta}}_1\|_2 + (1-\alpha)\|\widetilde{\boldsymbol{\beta}}_2\|_2\end{aligned}$$

thus $\|\widetilde{\boldsymbol{\beta}}\|_2$ is convex. The equality holds when $(1,-\boldsymbol{\beta}_1) = c(1,-\boldsymbol{\beta}_2)$ for any $c \in \mathbb{R}^+$, which implies $c = 1$ and $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$. Therefore, $\|\widetilde{\boldsymbol{\beta}}\|_2$ is strictly convex.

Lastly, $\|\mathbf{X}_j - \mathbf{X}_{-j}\boldsymbol{\beta}\|_2$ is convex by the triangle inequality:

$$\begin{aligned}\|\mathbf{X}_j - \mathbf{X}_{-j}\{\alpha\boldsymbol{\beta}_1 + (1-\alpha)\boldsymbol{\beta}_2\}\|_2 &= \|\alpha(\mathbf{X}_j - \mathbf{X}_{-j}\boldsymbol{\beta}_1) + (1-\alpha)(\mathbf{X}_j - \mathbf{X}_{-j}\boldsymbol{\beta}_2)\|_2 \\ &\leq \alpha\|\mathbf{X}_j - \mathbf{X}_{-j}\boldsymbol{\beta}_1\|_2 + (1-\alpha)\|\mathbf{X}_j - \mathbf{X}_{-j}\boldsymbol{\beta}_2\|_2.\end{aligned}$$

□

In the following result, we prove the uniqueness of (4.2.13), which is similar to the uniqueness of the elastic net solution [169].

Lemma B.3.2.2. (4.2.13) *has a unique solution.*

Proof. Since (4.2.13) is a convex problem, there exists a solution. By Lemma B.3.2.1, $f(\boldsymbol{\beta}) := \left(n^{-1/2}\|\mathbf{X}_j - \mathbf{X}_{-j}\boldsymbol{\beta}\|_2 + \rho\|\widetilde{\boldsymbol{\beta}}\|_2\right)^2 + \lambda\|\boldsymbol{\beta}\|_1$ is strictly convex. This implies the uniqueness the solution.

□

We can also prove the uniqueness of the solution of (4.2.14).

Lemma B.3.2.3. (4.2.14) *has a unique solution.*

Proof. Since (4.2.14) is a convex problem, there exists a solution. By Lemma B.3.2.1, $f(\boldsymbol{\beta}) := n^{-1/2}\|\mathbf{X}_j - \mathbf{X}_{-j}\boldsymbol{\beta}\|_2 + \lambda\|\boldsymbol{\beta}\|_1 + \rho\|\widetilde{\boldsymbol{\beta}}\|_2$ is strictly convex. This implies the uniqueness the solution. □

B.4 More on Algorithm

B.4.1 ADMM algorithm for (4.2.14)

In addition to Algorithm in Section 4.2.4, we introduce another algorithm by the alternating direction methods of multipliers (ADMM) algorithm [170]. ADMM algorithm is an algorithm well fitted for distributed convex optimization.

We reformulate the optimization problem (4.2.14) for ADMM algorithm by introducing an equality constraint:

$$\begin{aligned} &\underset{\boldsymbol{\beta},\boldsymbol{\theta}}{\operatorname{argmin}} \left\{ \frac{1}{n^{1/2}}\|\boldsymbol{\theta}\|_2 + \lambda\|\boldsymbol{\beta}\|_1 + \rho\|\widetilde{\boldsymbol{\beta}}\|_2 \right\} \\ &\text{subject to } \boldsymbol{\theta} = \mathbf{X}\widetilde{\boldsymbol{\beta}}, \end{aligned} \tag{B.4.1}$$

which can be solved by

$$\begin{aligned} \theta_i^{(t+1)} &= \left(1 + \frac{1}{n^{1/2}\gamma}\frac{1}{\|\boldsymbol{\theta}_i^*\|_2}\right)^{-1} (\mu_i^{(t)} + X_{ij} - \mathbf{X}_{i,-j}^T\boldsymbol{\beta}^{(t)}) \\ \beta_k^{(t+1)} &= \left(\frac{\rho}{\|(1, -\boldsymbol{\beta}_k^*)\|_2} + \gamma\|\mathbf{X}_k\|_2^2\right)^{-1} S_\lambda\left(\gamma\left\langle \mathbf{X}_k, \mathbf{v}_{jk}\right\rangle\right) \\ \boldsymbol{\mu}^{(t+1)} &= \boldsymbol{\mu}^{(t)} + (\mathbf{X}_j - \mathbf{X}_{-j}\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\theta}^{(t+1)}) \end{aligned} \tag{B.4.2}$$

for $i = 1, \ldots, n$ and $k \in \{1, \ldots, j-1, j+1, \ldots, d\}$ where

$$\begin{aligned} \boldsymbol{\theta}_i^* &:= (\theta_1^{(t+1)}, \ldots, \theta_{i-1}^{(t+1)}, \theta_i^{(t)}, \ldots, \theta_n^{(t)}), \\ \boldsymbol{\beta}_k^* &= (\beta_1^{(t+1)}, \ldots, \beta_{k-1}^{(t+1)}, \beta_k^{(t)}, \ldots, \beta_p^{(t)}), \\ \mathbf{v}_{jk} &:= \boldsymbol{\mu}^{(t)} + \mathbf{X}_j - \sum_{l=1, l\neq k}^{j-1} \beta_l^{(t+1)}\mathbf{X}_l - \sum_{l=j+1, l\neq k}^{d} \beta_l^{(t)}\mathbf{X}_l - \boldsymbol{\theta}^{(t+1)}, \end{aligned}$$

$S_\lambda(\cdot)$ is the soft-thresholding operator. We call this algorithm DRAGON-ADMM.

We provide derivation of (B.4.2). ADMM algorithm for (B.4.1) is the iterating scheme by solving the scaled form [170]:

$$\begin{aligned}
\boldsymbol{\theta}^{(t+1)} &= \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \left\{ n^{-1/2}\|\boldsymbol{\theta}\|_2 + \frac{\gamma}{2}\|\boldsymbol{\mu}^{(t)} + \mathbf{X}_j - \mathbf{X}_{-j}\boldsymbol{\beta}^{(t)} - \boldsymbol{\theta}\|_2^2 \right\} \\
\boldsymbol{\beta}^{(t+1)} &= \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \lambda\|\boldsymbol{\beta}\|_1 + \rho\|\widetilde{\boldsymbol{\beta}}\|_2 + \frac{\gamma}{2}\|\boldsymbol{\mu}^{(t)} + \mathbf{X}_j - \mathbf{X}_{-j}\boldsymbol{\beta} - \boldsymbol{\theta}^{(t+1)}\|_2^2 \right\} \\
\boldsymbol{\mu}^{(t+1)} &= \boldsymbol{\mu}^{(t)} + (\mathbf{X}_j - \mathbf{X}_{-j}\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\theta}^{(t+1)})
\end{aligned} \tag{B.4.3}$$

for some $\gamma > 0$. The minimizer of the first problem in (B.4.3) satisfies

$$\begin{aligned}
& n^{-1/2}\frac{\boldsymbol{\theta}}{\|\boldsymbol{\theta}\|_2} + \gamma(\boldsymbol{\theta} - \boldsymbol{\mu}^{(t)} - \mathbf{X}_j + \mathbf{X}_{-j}\boldsymbol{\beta}^{(t)}) = \mathbf{0} \\
\Rightarrow & \left(n^{-1/2}\frac{1}{\|\boldsymbol{\theta}\|_2} + \gamma \right) \boldsymbol{\theta} = \gamma(\boldsymbol{\mu}^{(t)} + \mathbf{X}_j - \mathbf{X}_{-j}\boldsymbol{\beta}^{(t)}).
\end{aligned}$$

Hence, we update $\boldsymbol{\theta}$ by solving iterating equations

$$\theta_i^{(t+1)} = \left(1 + \frac{1}{n^{1/2}\gamma}\frac{1}{\|\boldsymbol{\theta}_i^*\|_2} \right)^{-1} (\mu_i^{(t)} + X_{ij} - \mathbf{X}_{i,-j}^T\boldsymbol{\beta}^{(t)}) \tag{B.4.4}$$

for $i = 1, \ldots, n$ where $\boldsymbol{\theta}_i^* := (\theta_1^{(t+1)}, \ldots, \theta_{i-1}^{(t+1)}, \theta_i^{(t)}, \ldots, \theta_n^{(t)})$. The minimizer of the second problem in (B.4.3) satisfies

$$\begin{aligned}
& \lambda \cdot \operatorname{sign}(\beta_k) + \rho\frac{\beta_k}{\|\widetilde{\boldsymbol{\beta}}\|_2} - \gamma\mathbf{X}_k^{\mathrm{T}}(\boldsymbol{\mu}^{(t)} + \mathbf{X}_j - \mathbf{X}_{-j}\boldsymbol{\beta} - \boldsymbol{\theta}^{(t+1)}) = 0 \\
\Rightarrow & \left(\frac{\rho}{\|\widetilde{\boldsymbol{\beta}}\|_2} + \gamma\|\mathbf{X}_k\|_2^2 \right) \beta_k = \gamma\langle \mathbf{X}_k, \boldsymbol{\mu}^{(t)} + \mathbf{X}_j - \mathbf{X}_{-(j,k)}\boldsymbol{\beta}_{-k} - \boldsymbol{\theta}^{(t+1)} \rangle - \lambda \cdot \operatorname{sign}(\beta_k).
\end{aligned}$$

Thus, we update $\boldsymbol{\beta}$ by

$$\begin{aligned}
\beta_k^{(t+1)} &= \left(\frac{\rho}{\|\widetilde{\boldsymbol{\beta}}_k^*\|_2} + \gamma\|\mathbf{X}_k\|_2^2 \right)^{-1} S_\lambda\left(\gamma \langle \mathbf{X}_k, \mathbf{v}_{jk} \rangle\right) \\
\mathbf{v}_{jk} &:= \boldsymbol{\mu}^{(t)} + \mathbf{X}_j - \sum_{l=1, l\neq k}^{j-1} \beta_l^{(t+1)}\mathbf{X}_l - \sum_{l=j+1, l\neq k}^{d} \beta_l^{(t)}\mathbf{X}_l - \boldsymbol{\theta}^{(t+1)}
\end{aligned} \tag{B.4.5}$$

for all $k \in \{1, \ldots, j-1, j+1, \ldots, p\}$ where $S_\lambda(\cdot)$ is the soft-thresholding operator, $\boldsymbol{\beta}_k^* = (\beta_1^{(t+1)}, \ldots, \beta_{k-1}^{(t+1)}, \beta_k^{(t)}, \ldots, \beta_p^{(t)})$, and $\widetilde{\boldsymbol{\beta}}_k^* = (1, -\boldsymbol{\beta}_k^*)$.

DRAGON-ADMM is derived from (B.4.2) directly. However, it is slower than DRAGON. First and foremost, ADMM algorithm uses augmentation parameter, which also need to be selected by user and affect the convergence speed. Second, it requires more computations from vector and matrix operations.

B.5 Additional results from simulation studies

In this section, we report additional numerical results for simulations detailed in Section 4.3 in the main paper. We report the results from $(n, p) = \{(200, 150), (200, 300)\}$. For ease of presentation, we revisit the precision matrix structure settings and contamination settings.

We consider three precision matrix structures for the first stage of data generation procedure. Construct the precision matrix by $\boldsymbol{\Omega} = \mathbf{D}\widetilde{\boldsymbol{\Omega}}\mathbf{D}$ where $\widetilde{\boldsymbol{\Omega}} = \{\widetilde{\omega}_{j,k}\}_{1 \le j,k \le p}$ and $\mathbf{D}$ is a diagonal matrix with elements $d_{j,j}$:

1. *Model 1* (Banded) Set $\widetilde{\omega}_{j,j} = 1$, $\widetilde{\omega}_{j,j+1} = \widetilde{\omega}_{j+1,j} = 0.6$, $\widetilde{\omega}_{j,j+2} = \widetilde{\omega}_{j+2,j} = 0.3$, $\widetilde{\omega}_{j,k} = 0$ for $|j-k| \ge 3$. Generate $d_{j,j} \sim uniform(1, 5)$.

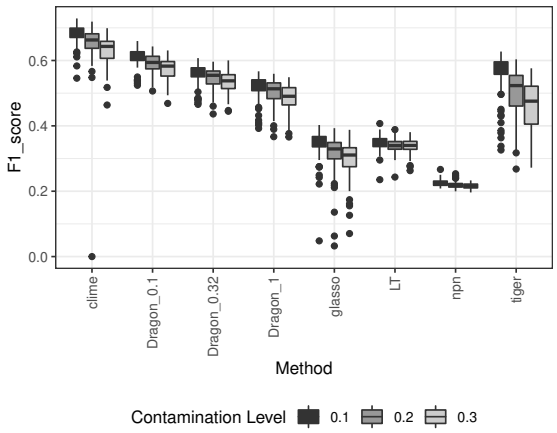
2. *Model 2* (Block diagonal) Set a block diagonal matrix with block size $p/10$ such that the diagonal entries are 1 and the off-diagonal entries are 0.5, then we permute the matrix by rows/columns to get $\widetilde{\boldsymbol{\Omega}}$. We use $d_{j,j} = 1$ for $j = 1, \ldots, p/2$ and $d_{j,j} = 1.5$ for $j = p/2+1, \ldots, p$ to obtain the final product $\boldsymbol{\Omega}$.

3. *Model 3* (Erdös-Rényi) Generate $\widetilde{\boldsymbol{\Omega}} = \{\widetilde{\omega}_{jk}\}_{1 \le j,k \le p}$ $\widetilde{\omega}_{1,jj} = 1$, $\widetilde{\omega}_{1,jk} = \delta_{jk} u_{jk}$ for $j < k$ where $\delta_{jk} \sim Ber(0.05)$ and $u_{jk} \sim uniform(0.4, 0.8)$, and $\widetilde{\omega}_{1,kj} = \widetilde{\omega}_{1,jk}$. Generate $d_{j,j} \sim uniform(1, 5)$. Then set $\boldsymbol{\Omega} = \mathbf{D}\{\widetilde{\boldsymbol{\Omega}} + (|\lambda_{\min}(\widetilde{\boldsymbol{\Omega}})| + 0.05)\mathbf{I}\}\mathbf{D}$ where $\lambda_{\min}(\cdot)$ is the smallest eigenvalue of the matrix.

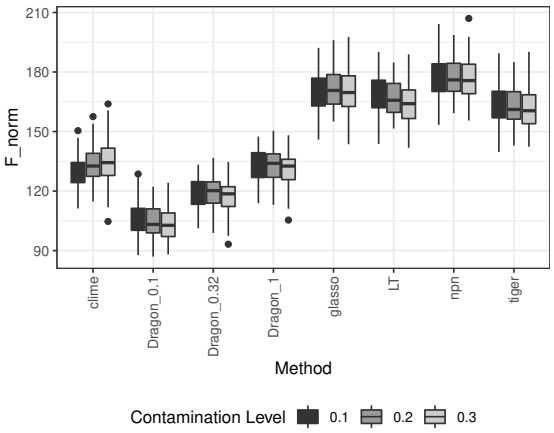
We consider three distributional perturbation scenarios on the data generation. For each pair of (n, p), (i) Row-wise contamination- We Generate $n_1 = \lceil 100 \times (1-\alpha) \rceil$ samples from $N(\mathbf{0}, \boldsymbol{\Omega}^{-1})$

where Ω is the precision matrix specified above and $\alpha \in \{0.1, 0.2, 0.3\}$. Then, we generate $n - n_1$ samples from the multivariate t_3-distribution with the true precision matrix $\mathbf{\Omega}$; (ii) Cell-wise contamination- Generate n samples from $N(\mathbf{0}, \mathbf{\Omega}^{-1})$. We randomly select αnp indices to add cell-wise contaminants drawn from $N(0, 1)$ where $1 \leq i \leq n$, $1 \leq j \leq p$, and $\alpha \in \{0.1, 0.2, 0.3\}$; (iii) Tail deviation; we draw n samples from the multivariate t_3-distribution to have its true precision matrix be $\mathbf{\Omega}$.

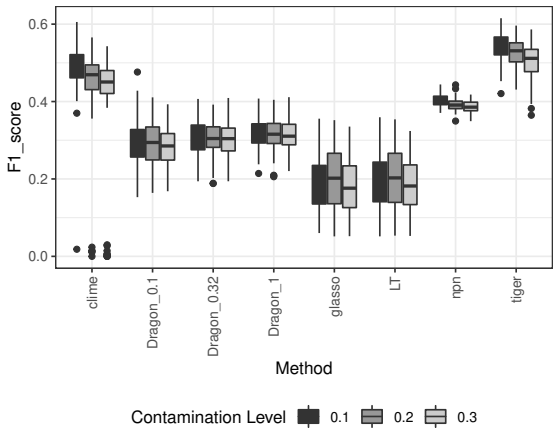
Figures B.1–B.3 displays results when $(n, p) = (200, 150)$, and Figures B.4–B.6 displays results when $(n, p) = (200, 300)$. We observe the similar phenomenon as the result from $(n, p) = (100, 150)$. DRAGON provides good selection performance and the smallest Frobenius norm in the most of the settings. Erdös-Rényi graph is the most difficult structure for DRAGON in terms of selection performance. On the other hand, DRAGON outperforms the competing methods in block diagonal structure. DRAGON is not dominated by one of the competing methods over all settings.



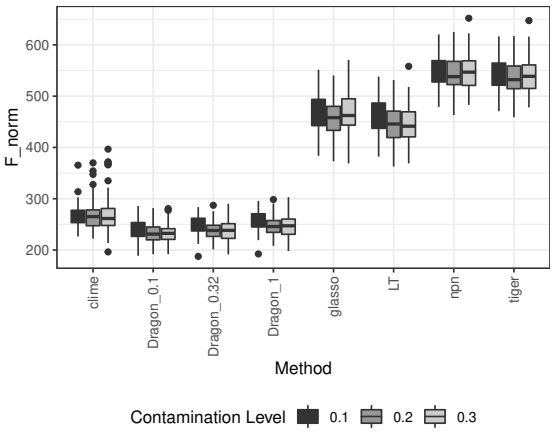
(a) Banded, F1 score



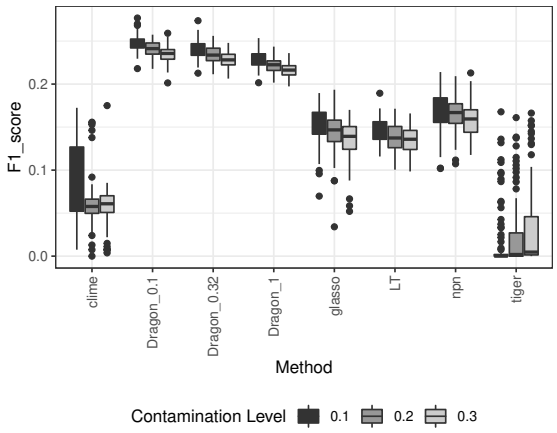
(b) Banded, Frobenius norm



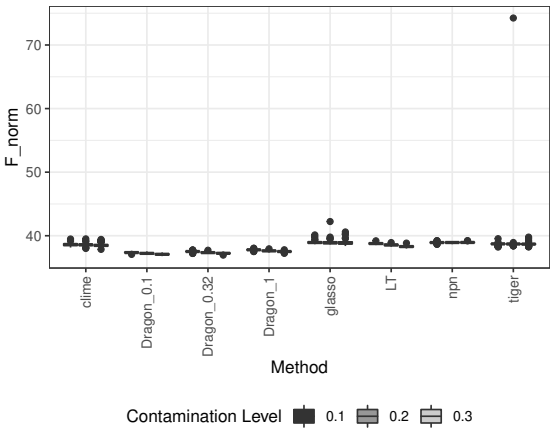
(c) Erdös-Rényi, F1 score



(d) Erdös-Rényi, Frobenius norm

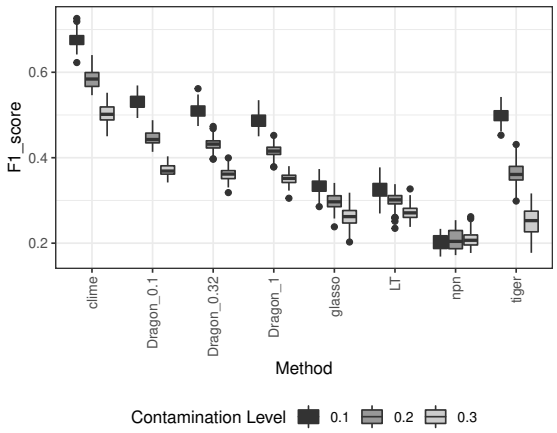


(e) Block diagonal, F1 score

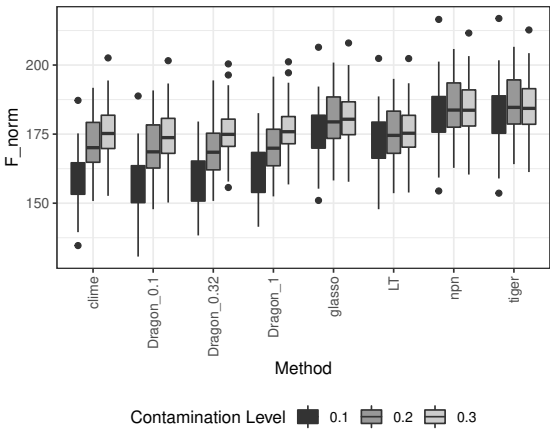


(f) Block diagonal, Frobenius norm

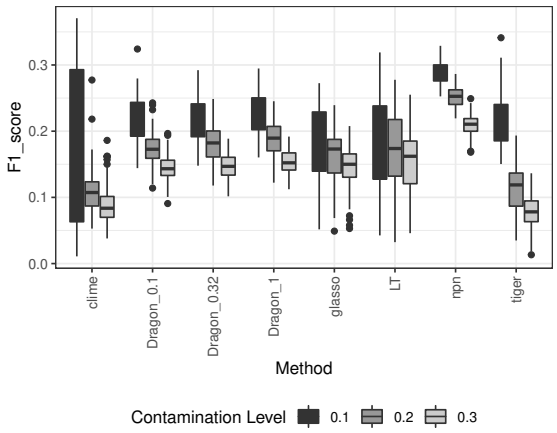
Figure B.1: F1 score (left) and Frobenius norm (right) under the rowwise contamination setting when $(n, p) = (200, 150)$. Each boxplot summarizes the results from 100 repetitions of experiment.



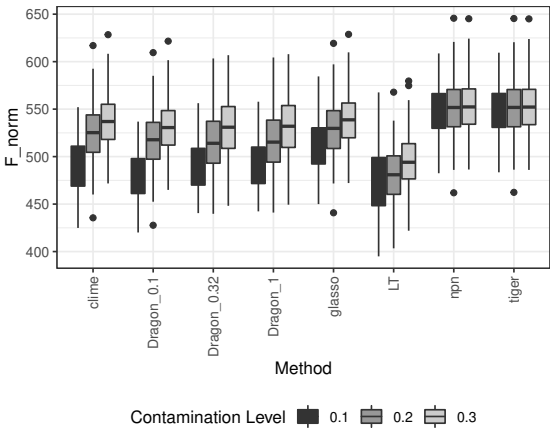
(a) Banded, F1 score



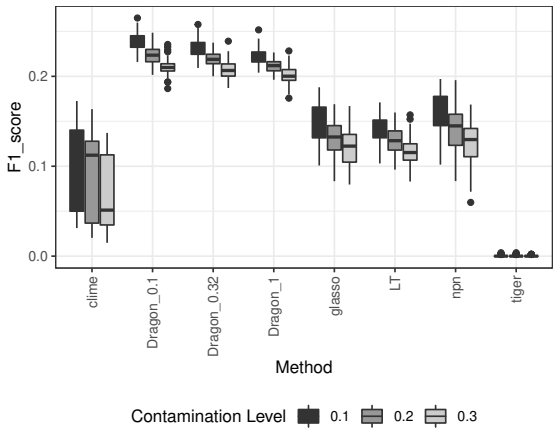
(b) Banded, Frobenius norm



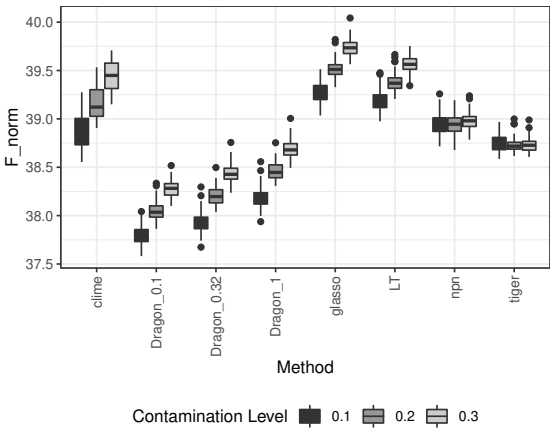
(c) Erdös-Rényi, F1 score



(d) Erdös-Rényi, Frobenius norm

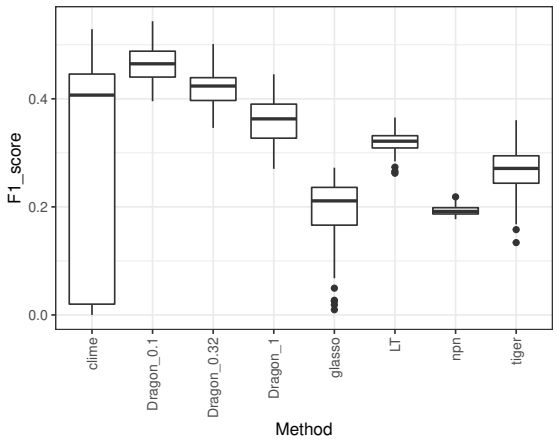


(e) Block diagonal, F1 score

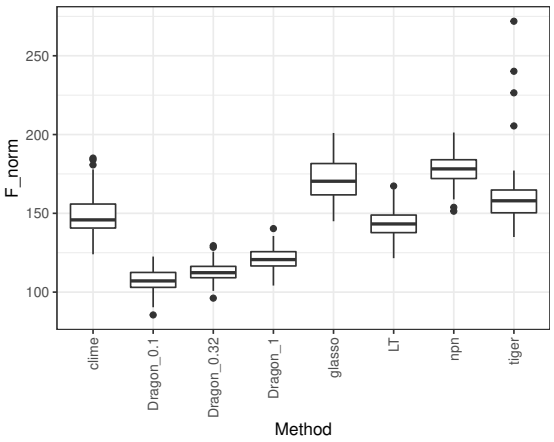


(f) Block diagonal, Frobenius norm

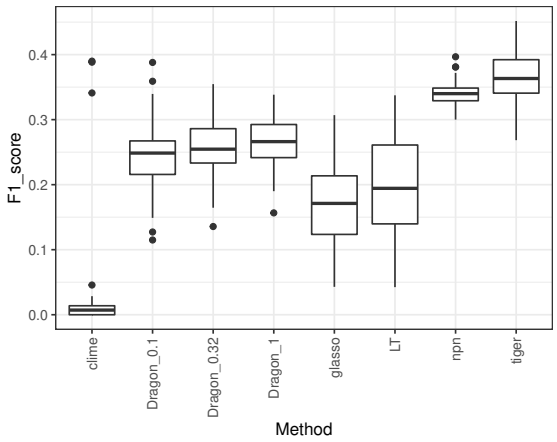
Figure B.2: F1 score (left) and Frobenius norm (right) under the cellwise contamination setting when $(n, p) = (200, 150)$. Each boxplot summarizes the results from 100 repetitions of experiment.



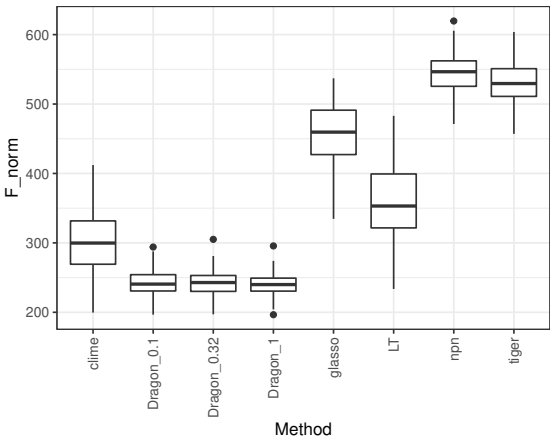
(a) Banded, F1 score



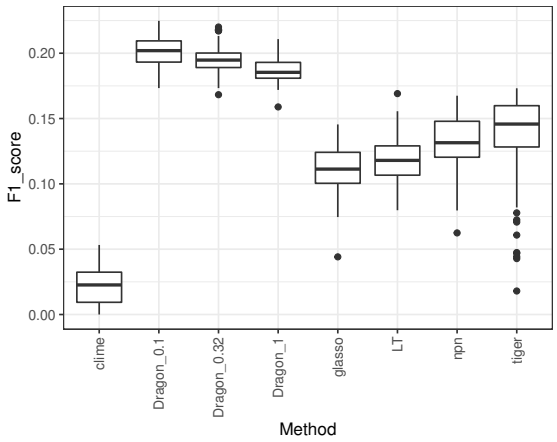
(b) Banded, Frobenius norm



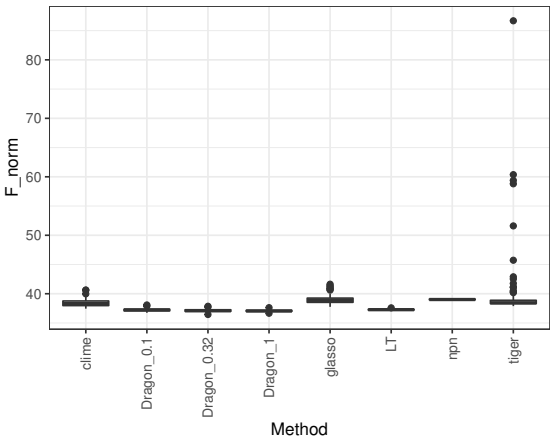
(c) Erdös-Rényi, F1 score



(d) Erdös-Rényi, Frobenius norm

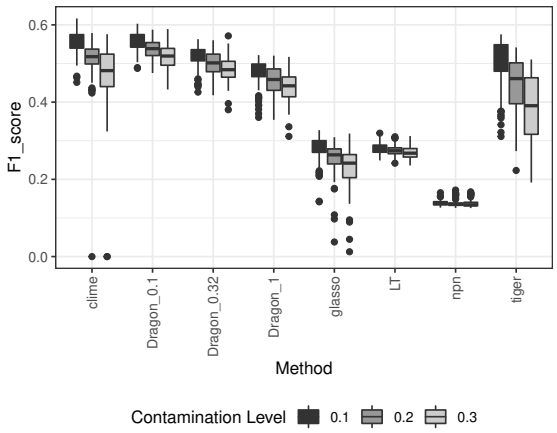


(e) Block diagonal, F1 score

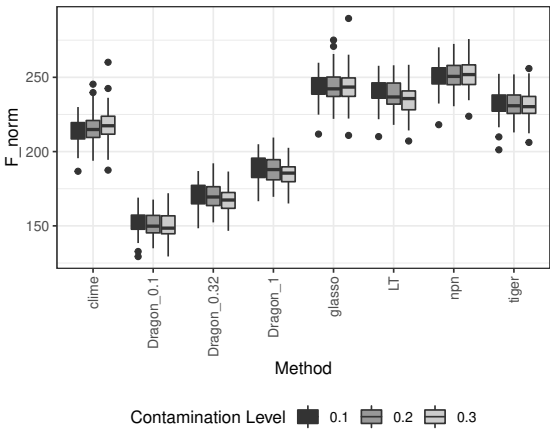


(f) Block diagonal, Frobenius norm

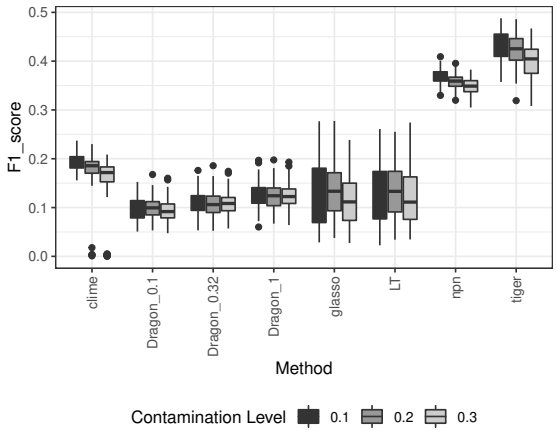
Figure B.3: F1 score (left) and Frobenius norm (right) under the tail deviation setting when $(n, p) = (200, 150)$. Each boxplot summarizes the results from 100 repetitions of experiment.



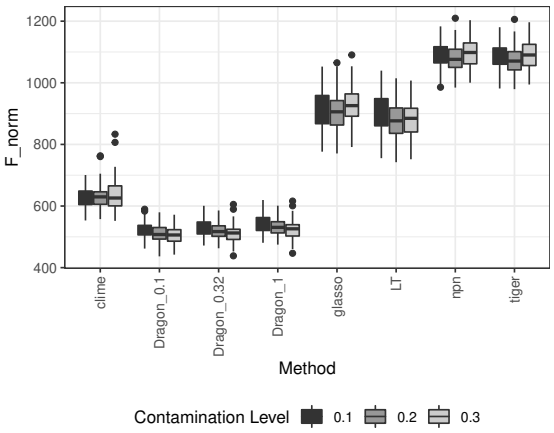
(a) Banded, F1 score



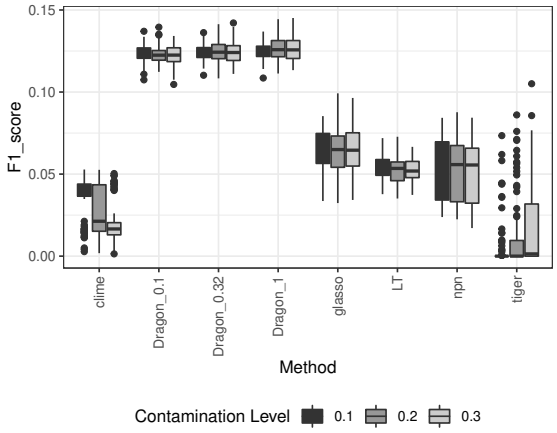
(b) Banded, Frobenius norm



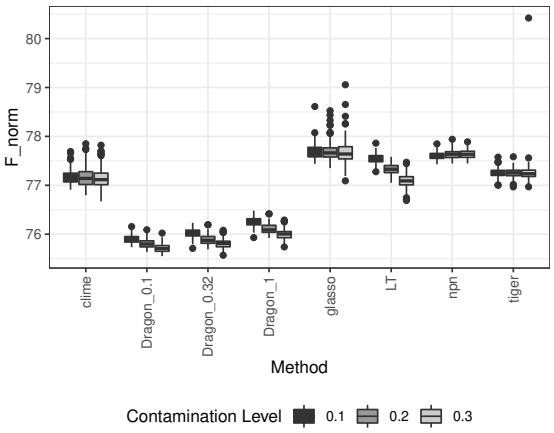
(c) Erdös-Rényi, F1 score



(d) Erdös-Rényi, Frobenius norm

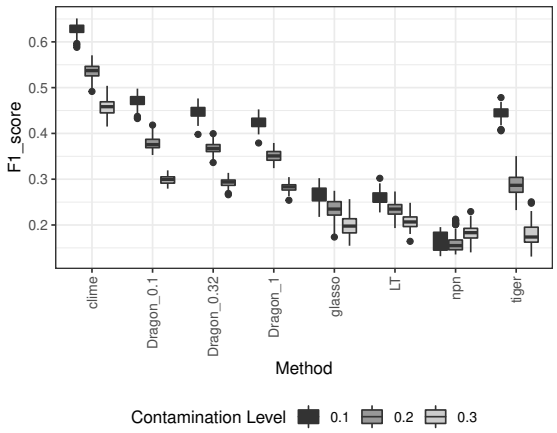


(e) Block diagonal, F1 score

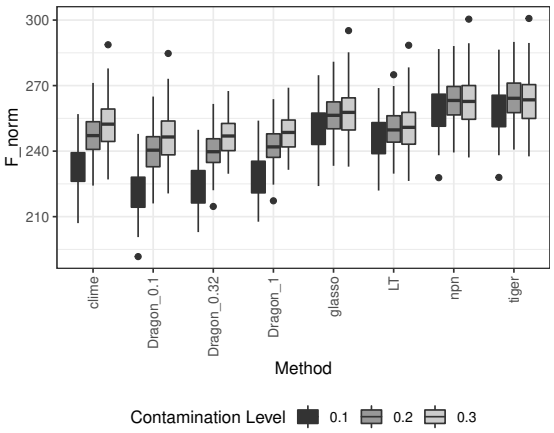


(f) Block diagonal, Frobenius norm

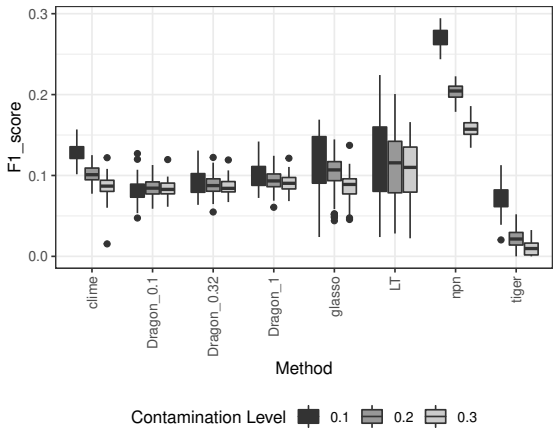
Figure B.4: F1 score (left) and Frobenius norm (right) under the rowwise contamination setting when $(n, p) = (200, 300)$. Each boxplot summarizes the results from 100 repetitions of experiment.



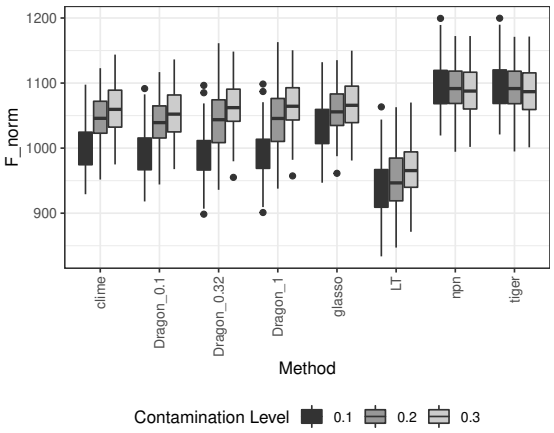
(a) Banded, F1 score



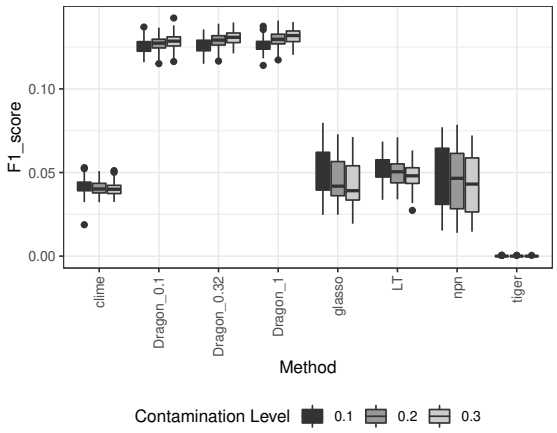
(b) Banded, Frobenius norm



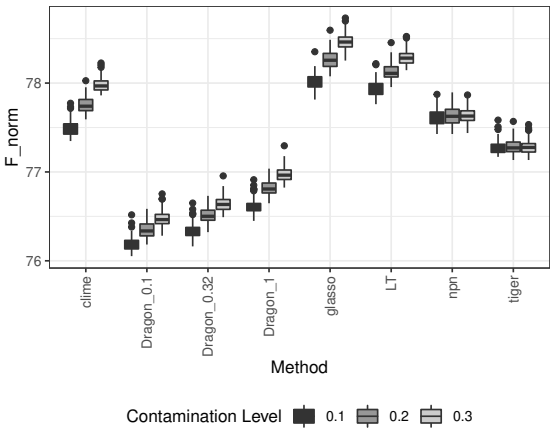
(c) Erdös-Rényi, F1 score



(d) Erdös-Rényi, Frobenius norm

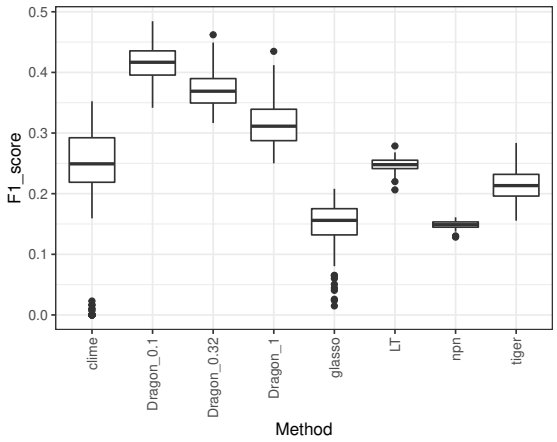


(e) Block diagonal, F1 score

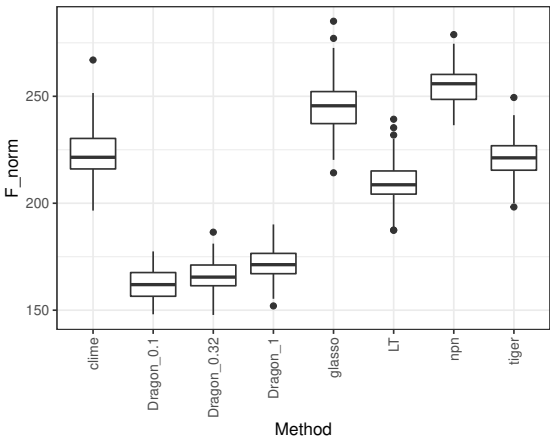


(f) Block diagonal, Frobenius norm

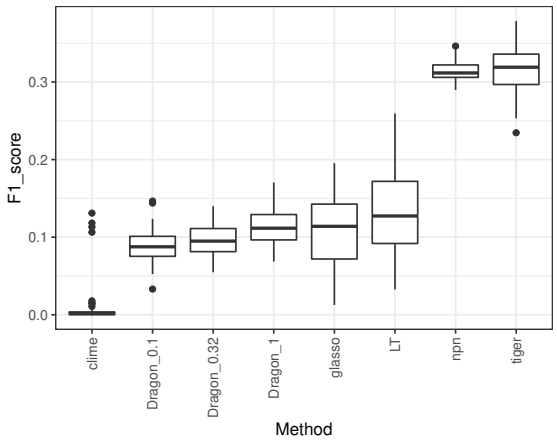
Figure B.5: F1 score (left) and Frobenius norm (right) under the cellwise contamination setting when $(n, p) = (200, 300)$. Each boxplot summarizes the results from 100 repetitions of experiment.



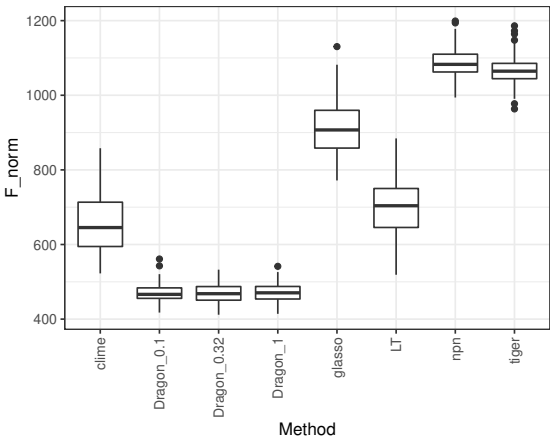
(a) Banded, F1 score



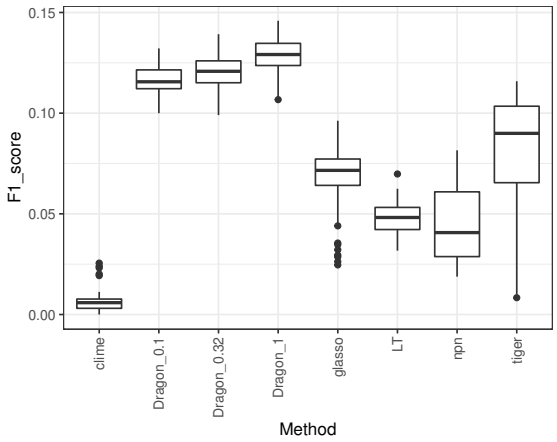
(b) Banded, Frobenius norm



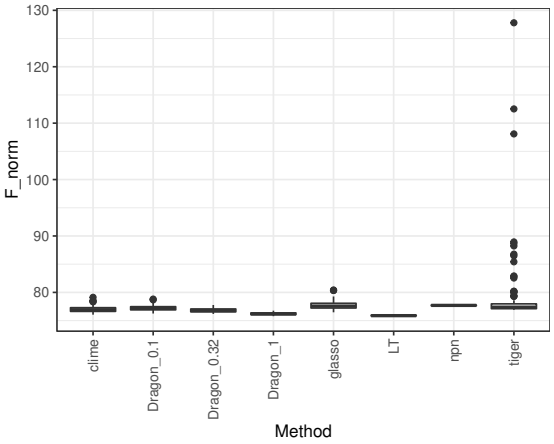
(c) Erdös-Rényi, F1 score



(d) Erdös-Rényi, Frobenius norm



(e) Block diagonal, F1 score



(f) Block diagonal, Frobenius norm

Figure B.6: F1 score (left) and Frobenius norm (right) under the tail deviation setting when $(n, p) = (200, 300)$. Each boxplot summarizes the results from 100 repetitions of experiment.