Title
Modern Statistical Inference for Classical Statistical Problems

Author
Lei, Lihua

Publication Date
2019

Peer reviewed|Thesis/dissertation

Modern Statistical Inference for Classical Statistical Problems

by

Lihua Lei

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Statistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Peter J. Bickel, Co-chair
Professor Michael I. Jordan, Co-chair
Professor Venkatachalam Anantharam
Assistant Professor William Fithian

Summer 2019

Modern Statistical Inference for Classical Statistical Problems

Abstract

Modern Statistical Inference for Classical Statistical Problems

by

Lihua Lei

Doctor of Philosophy in Statistics

University of California, Berkeley

Professor Peter J. Bickel, Co-chair

Professor Michael I. Jordan, Co-chair

This dissertation addresses three classical statistics inference problems with novel ideas and techniques driven by modern statistics. My purpose is to highlight the fact that even the most fundamental problems in statistics are not fully understood and the unexplored parts may be handled by advances in modern statistics. Pouring new wine into old bottles may generate new perspectives and methodologies for more complicated problems. On the other hand, re-investigating classical problems help us understand the historical development of statistics and pick up the scattered pearls forgotten over the course of history.

Chapter 2 discusses my work supervised by Professor Noureddine El Karoui and Professor Peter J. Bickel on regression M-estimates in moderate dimensions. In this work, we investigate the asymptotic distributions of coordinates of regression M-estimates in the moderate p/n regime, where the number of covariates p grows proportionally with the sample size n. Under appropriate regularity conditions, we establish the coordinate-wise asymptotic normality of regression M-estimates assuming a fixed-design matrix. Our proof is based on the second-order Poincaré inequality (Chatterjee 2009) and leave-one-out analysis (El Karoui et al. 2011). Some relevant examples are indicated to show that our regularity conditions are satisfied by a broad class of design matrices. We also show a counterexample, namely the ANOVA-type design, to emphasize that the technical assumptions are not just artifacts of the proof. Finally, the numerical experiments confirm and complement our theoretical results.

Chapter 3 discusses my joint work with Professor Peter J. Bickel on exact inference for linear models. We propose the cyclic permutation test (CPT) for testing general linear hypotheses for linear models. This test is non-randomized and valid in finite samples with exact type-I error α for arbitrary fixed design matrix and arbitrary exchangeable errors, whenever $1/\alpha$ is an integer and $n/p \geq 1/\alpha - 1$. The test applies the marginal rank test on $1/\alpha$ linear statistics of the outcome vectors where the coefficient vectors are determined by solving a linear system such that the joint distribution of the linear statistics is invariant to a non-standard cyclic permutation group under the null hypothesis. The power can be

further enhanced by solving a secondary non-linear travelling salesman problem, for which the genetic algorithm can find a reasonably good solution. We show that CPT has comparable power with existing tests through extensive simulation studies. When testing for a single contrast of coefficients, an exact confidence interval can be obtained by inverting the test. Furthermore, we provide a selective yet extensive literature review of the century-long efforts on this problem, highlighting the novelty of our test.

Chapter 4 discusses my joint work with Professor Peng Ding on regression adjustment for Neyman-Rubin models. Extending R. A. Fisher and D. A. Freedman's results on the analysis of covariance, Lin (2013) proposed an ordinary least squares adjusted estimator of the average treatment effect in completely randomized experiments. We further study its statistical properties under the potential outcomes model in the asymptotic regimes allowing for a diverging number of covariates. We show that when $p \gg n^{1/2}$, the estimator may have a non-negligible bias and propose a bias-corrected estimator that is asymptotically normal in the regime $p = o(n^{2/3}/(\log n)^{1/3})$. Similar to Lin (2013), our results hold for non-random potential outcomes and covariates without any model specification. Our analysis requires novel analytic tools for sampling without replacement, which complement and potentially enrich the theory in other areas such as survey sampling, matrix sketching, and transductive learning.

Contents

List of Figures

Acknowledgments

First and foremost, I would like to thank my terrific advisors at UC Berkeley, Professor Peter Bickel and Professor Michael Jordan, Professor Noureddine El Karoui, Professor William Fithian and Professor Peng Ding. Without their tremendous efforts and patience, I would not have been able to work as an academic statistician and contact with a multitude of areas.

My first formal project was supervised by Noureddine and Peter, who impressed me with their sagacity, knowledgeability and sharpness. I learned so many deep insights from the discussion between Noureddine and Peter in our regular weekly meetings for three years. During the period when I doubted if the problem could be solved, Noureddine came up with the remarkable ideas and techniques which turn out to be the key to the project. Although the final paper is 90-pages long, Noureddine checked the proof line by line and revised the paper in great details, even in the midst of his sabbatical. Had I not been advised by him, I would not have even touched a corner of the iceberg. I deeply appreciate his tremendous efforts and patience as a remarkable advisor for over three years. Later on I was so fortunate to keep working with Peter on other projects beyond pure theory driven by real-world problems. Peter is the most ingenious statistician I have ever interacted with. He has numerous ideas which appear to be abstract and vague at the beginning, but always turn out to work and lead to mind-blowing methodologies. I cannot forget the exciting moments when I managed to understand the essence of Peter's proposals, followed by a big "wow".

Being advised by Mike has been yet another stroke of great fortune. Mike is a "walking encyclopedia" with a vast knowledge across numerous disciplines, without which I could not have seen interesting results in different areas. He has always been kind and supportive to me as well as my crazy research ideas. The acronym SCSG, coined by Mike for one of our algorithm, is perfect to describe his figure in my mind – Savvy, Creative, Supportive and Gentle. Mike also provided an extraordinary research environment, marked by his remarkable weekly group meeting. As a curiosity-driven researcher, it is of great benefit for me to read materials on diverse topics, ranging from causal inference to stochastic differential equations to mechanism designs, together with his wonderful students.

My collaboration with Will started from his fabulous course on selective inference. Being one of the five enrolled student, my questions filled in almost every one of his lecture. I was thankful that he did not kick me out of the classroom for being overly challenging and was totally impressed by the clarity of the answers and the deep insights behind. The course was so interesting and inspiring that my course project was later turned into my first conference publication. In the following collaborations, Will was never falling short of creative ideas or accurate intuition. His geometry-driven thinking patterns complements my algebra-driven perspective and greatly improves my research skills.

Peng was my role model in college and I cannot express how excited I were when he chose to join our department as an assistant professor. I must attribute all my research interests on causal inference to Peng, who taught me this long-standing topic seriously and thoroughly. Our collaborations were always smooth and efficient due to his kindness and patience. Beyond his intelligence, I was greatly impressed by his wide knowledge of statistical

history as a junior faculty, which significantly impacted my vision and philosophy of being a statistician.

My thanks also go to other professors, in particular Professor Venkat Anantharam, who was extremely kind to be on both my qualifying exam and dissertation committees and provided helpful comments; Professor Cari Kaufman, who provided guidance and led me into the Bayesian world in my first semester at Berkeley; Professor Avi Feller, who is one of the core organizer of the weekly causal reading group which drastically influenced my research and motivated a line of joint works; Professor Jasjeet Sekhon, who provided crucial comments in our joint work; Professor Bin Yu, who taught the excellent 215A course that exemplified the charm of applied statistics an reshaped my principle of being a good statistian; Professor Martin Wainwright, who taught the wonderful 210B course which laid the solid theoretical foundation for my research; Professor Christopher Paciorek, who provided enormous support for softwares and computation in my research; Professor Elizaveta Levina from University of Michigan, who invited me to join the force of her project and provided instructive and insightful guidance; Professor Guido Imbens from Stanford University, who gave a thought-provoking talk at Berkeley and motivated a joint project. The thanks are also extended to Professor David Tse, Professor Stefan Wager, Professor Emmanuel Candès, Professor Cho-Jui Hsieh, Professor Elizaveth Levina, Professor Xuming He, Professor Yingying Fan, Professor Fredrik Sävje, Professor Kai Zhang, Professor Chaitra Nagaraja and Professor Linda Zhao for inviting me to give academic talks, which are encouraging as a junior researcher. In addition, I would like to thank my past and current collaborators Cheng Ju, Yuting Ye, Jianbo Chen, Alexander D'Amour, Aaditya Ramdas, Chiao-Yu Yang, Nhat Ho, Yuchen Wu, Tianxi Li, Sharmodeep Bhattacharyya, Purnamrita Sarkar, Melih Elibol, Samuel Horvath, Hongyuan Cao, Zitong Yang, Xingmei Lou and Xiaodong Li.

Next I would to express gratitude to our department and Ph.D. program, which I am very proud of. I am grateful to all staff, especially La Shana, who is always there helping me with numerous subtle issues patiently. I am also very thankful to my excellent fellow students from whom I learn a lot from our academic discussions, in particular Eli Ben-Michael, Joseph Borja, Yuansi Chen, Billy Fang, Han Feng, Ryan Giordano, Johnny Hong, Steve Howard, Kenneth Hung, Chi Jin, Sören Künzel, Hongwei Li, Lisa Li, Xiao Li, Tianyi Lin, Sujayam Saha, Jake Soloff, Sara Stoudt, Wenpin Tang, Yu Wang, Yuting Wei, Jason Wu, Siqi Wu, Zhiyi You, Da Xu, Renyuan Xu, Chelsea Zhang and Yumeng Zhang. Further I am indebted to my academic friends outside Berkeley, including but not limited to Yu Bai, Fang Cai, Xi Chen, Chao Gao, Xinzhou Guo, Zhichao Jiang, Asad Lodhia, Eugene Katsevich, Jason Lee, Song Mei, Nicole Pashley, Zhimei Ren, Feng Ruan, Weijie Su, Qingyun Sun, Pragya Sur, Jingshen Wang, Jingshu Wang, Sheng Xu, Yiqiao Zhong and Qingyuan Zhao.

Finally, I owe the most to my family. My wife Xiaoman Luo always has the magic to bring me peace and confidence when I was stressful, anxious and helpless. Her sense of humor is the major source of happiness outside my academic life. Meeting and marrying her is my greatest achievement over the past five years that is more important than any publication or academic achievement. My parents are always supportive in spite of 6500 miles between us. I could not have made any achievement without their unconditional love and support.

Chapter 1

Introduction

Inference from data lies at the heart of modern scientific research. Etymologically, the word "inference" means to "carry forward" and can be dated back to late 16th century from medieval Latin. Despite the solid philosophical and logical foundation, inference is never an easy task in practice due to uncertainty inherent in data. Statistics, pinoneered in 17th century and rapidly developed since early 20th century, is a discipline to generate frameworks and methodologies to understand and handle uncertainty in inference and decision making. Perhaps for this reason, statistical inference grows as a major approach of inference which is widely adopted in scientific areas.

Recent years have seen a remarkable burst of advances in data collection technology, which have created a dizzying array of exciting application areas for statistical inference. Nowadays phrases like "data science" and "big data" become the new fashion sweeping the social media. As a college student majored in statistics, I was deeply attracted by various fancy concepts and methodologies in modern statistics, marked by the development in 1990s such as sparse regression methods, statistical learning methods, social networks, etc.. But at the same time, my curiosity of classical statistics accrues as I delved further into the area. "What happened in statistics before 1990s?" – This is a question always haunting my minds. After all, the development over the past century laid the foundation for the success of modern statistics in the era of big data. Although I occasionally learned some classical topics from the textbooks, it is not even close to a complete story.

My journey to the old territory of statistics began upon reading Ronald A. Fisher's 1922 article "On the Mathematical Foundation of Theoretical Statistics". In this pioneering work, he summarized the purpose of statistical methods as "the reduction of data" and more specifically, he wrote:

> A quantity of data, which usually by its mere bulk is incapable of entering the mind, is to be replaced by relatively few quantities which shall adequately represent the whole, or which, in other words, shall contain as much as possible, ideally the whole, of the relevant information contained in the original data.

He further clarified the distinction between a hypothetical population and a sample, between

an estimand and an estimator, thereby emphasizing the importance to identify the "source of randomness" in statistical inference. Furthermore, he categorized statistical problems into three types:-

> (1) Problems of Specification. These arise in the choice of the mathematical form of the population.
> (2) Problems of Estimation. These involve the choice of methods of calculating from a sample statistical derivates, or as we shall call them statistics, which are designed to estimate the values of the parameters of the hypothetical population.
> (3) Problems of Distribution. These include discussions of the distribution of statistics derived from samples, or in general any functions of quantities whose distribution is known.

Over the last century, "problems of specification" led to a plethora of statistical models (e.g. linear models, randomization models, time series models, etc.) and identification strategies; "problems of estimation" motivated the decision theoretic framework and criteria (e.g. unbiasedness, minimaxity, admissibility, etc.); "problems of distribution" generated the framework of hypothesis testing and the notion of confidence intervals, as well as the solid asymptotic distributional theory.

This remarkable categorization is still valid and quite comprehensive in modern statistics, which is equipped by advanced techniques and refined methodologies but mostly aims at handling the above three tasks. It is therefore valuable for researchers to look back on history, itself being the future of earlier history, to understand how ideas, languages, techniques and methodologies evolved, as opposed to what they appeared in textbooks written from hindsight. For instance, had I been a statistician in 1970, I would be more likely than a statistician today to be familiar with Edgeworth expansion, due to the approximation theory for t-test and F-test in absence of normality (e.g. Bartlett 1935; Wallace 1958). As a result, it would be more likely for me to understand, or even to discover, the mind-blowing connection between Edgeworth expansion and higher-order accuracy of bootstrap, developed in late 1980s (e.g. Hall 1989, 1992). Similarly, had I been familiar with the early development of design-based inference (e.g. Neyman 1923; Welch 1937; Cornfield 1944) and survey sampling (e.g. Neyman 1934; Cochran 1977), it would be easier for me to understand the modern design-based causal inference under the potential outcomes framework (e.g. Freedman 2008b,a; Lin 2013; Bloniarz et al. 2016; Abadie et al. 2017). Those who are familiar with classical statistics are more likely able to find and polish the "scattered pearls" that were under-studied or forgotten over the course of history to bring back their brilliance.

On the other hand, the models and the methodologies in classical statistics may not be fully understood in spite of the long history. For instance, the linear model is over 100 years old but it still inspires new research questions in modern statistics. One remarkable example is the breakdown of classical maximum likelihood theory for linear models in moderate dimensions, where the number of predictors grows linearly with sample size. (Bean et al. 2013) showed that the optimal M-estimator in this regime is no longer the maximum likelihood estimator but is associated with a complicated loss function determined by a nonlinear system

that involves the design properties, the sample size per parameter as well as the error distribution (El Karoui et al. 2011). The astonishing finding quickly attracted further attention (e.g. El Karoui 2013, 2015; Donoho and Montanari 2015; Donoho and Montanari 2016; Sur et al. 2017; Sur and Candès 2019). Although some earlier works (e.g. Huber 1973a; Bickel and Freedman 1983a) found evidence of non-standard properties of moderate dimensional regime, the aforementioned line of work was fueled by the advances in random matrix theory and statistical physics. These works are not purely theoretical pursuit. Instead, they suggest that the standard softwares may report misleading numbers in many applications even for well-studied linear models. This is a huge warning for practitioners and will inspire further efforts in the future to robustify the built-in algorithms. This inspiring example suggests the tremendous value of investigating classical statistical problems from new perspectives and equipped with advanced techniques.

In my dissertation, I will investigate three classical statistical problems but develop novel ideas and techniques to solve them, which I refer to as "modern". Of course, this is an exaggeration since three examples are far too restrictive to show the glamour of modern statistical inference. Nonetheless, they are epitomes of the elegance and the surprise when modern statistical knowledge meets classical statistical problems. In particular, all works in the dissertation deal with "problems of distribution", in which I found the classical statistics leave numerous unsolved questions while modern techniques and methodologies have great potential to come into play. I sketch the three works in each of the following subsections respectively.

1.1 Regression M-Estimates in Moderate Dimensions

Given a linear model $y = X\beta^* + \epsilon$ with outcome vector $y \in \mathbb{R}^n$, design matrix $X \in \mathbb{R}^{n\times p}$, coefficient vector $\beta^* \in \mathbb{R}^p$ and stochastic errors $\epsilon \in \mathbb{R}^n$, an regression M-estimator is defined as

$$\hat{\beta}(\rho) = \underset{\beta\in\mathbb{R}^p}{\arg\min} \sum_{i=1}^{n} \rho(y_i - x_i^T\beta).$$

M-estimators were proposed by Peter J. Huber in 1960s (Huber 1964) and have been widely studied in literature (e.g. Relles 1968; Yohai 1972; Huber 1973a; Yohai and Maronna 1979a; Portnoy 1984, 1985; Mammen 1989, 1993). In a nutshell, when the sample size per parameter n/p tends to infinity, under some regularity conditions, $\hat{\beta}(\rho)$ is consistent in L_2 metric and is asymptotically normal in the sense that for any fixed sequence of vectors $a_n \in \mathbb{R}^p$,

$$\frac{a_n^T(\hat{\beta}(\rho) - \beta^*)}{\sqrt{a_n^T\Sigma_n a_n}} \Longrightarrow N(0,1), \quad \text{where } \Sigma = \mathrm{Cov}(\hat{\beta}(\rho)). \tag{1.1}$$

However, the story completely changes in the moderate dimensional regime, where $p/n \to \kappa \in (0,1)$. In moderate dimensions, the sample size per parameter is bounded away from infinity and thus there are insufficient samples for estimating every coefficient accurately. For

least-squares estimators, Huber (1973a) proved that (1.1) is impossible for every sequence of a_n's in moderate dimensions. For general M-estimators with particular random designs, El Karoui et al. (2011) showed the inconsistency of $\hat{\beta}(\rho)$ in L_2 metric and characterized the limiting L_2 risk as the solution of a delicate nonlinear system involving κ, the distribution of X and the distribution of errors. On the other hand, Bean et al. (2013) proved (1.1) with Gaussian design matrices for any fixed sequence of a_n's in moderate dimensions. This is not contradicted to Huber (1973a) as the latter assumes a fixed design and thus the claim (1.1) only involves the randomness from ϵ, while Bean et al. (2013)'s result also considers the randomness of design matrices which brings more regularity.

These works inspired a line of studies that extended the results to general settings (El Karoui 2013, 2015; Donoho and Montanari 2015; Donoho and Montanari 2016; Sur et al. 2017; Sur and Candès 2019). However, most of them focused on special random designs, such as Gaussian matrices or random matrices with elliptically distributed rows. Furthermore, their central research question is to determine the limiting risk of $\hat{\beta}(\rho)$. Although some attempts have been made to the "problem of distribution", the results are based on Gaussian designs (Bean et al. 2013; Donoho and Montanari 2016; Sur et al. 2017; Sur and Candès 2019), with a few exceptions on more general random designs (El Karoui 2015, 2018), and some of them are about the "bulk distribution" of all coefficients which is less interpretable to practitioners. No distributional result was established previously for general M-estimators with fixed designs in moderate dimensions.

In this chapter, we ask a classical question: what is the asymptotic distribution of a given coordinate of $\hat{\beta}(\rho)$ in moderate dimensions assuming a fixed design. This question is surprisingly hard to answer than it appears to be, mainly due to the fundamental difficulty lying in the moderate dimensional regime. Unlike the low dimensional regime , in which the estimator has asymptotically linearity and thus the Linderberg-Feller-type central limit theorem can be applied to prove the asymptotic normality, the Taylor-expansion-type argument does not carry over to in moderate dimensional regime because there is only bounded number of samples on average for each parameter. Instead, we apply the second-order Poincaré inequality (Chatterjee 2009) that can be regarded as a generalization of classical central limit theorem to nonlinear transformation of independent random variables. In addition, we replace the Taylor-expansion-type argument by a more involved leave-one-out argument that generalizes El Karoui (2013)'s techniques to fixed-designs. In summary, we prove the following result.

Theorem 1.1.1 (Informal Version)**.** *Under appropriate conditions on the design matrix X, the distribution of ϵ and the loss function ρ, as $p/n \to \kappa \in (0,1)$, while $n \to \infty$,*

$$\max_{1\le j\le p} d_{\mathrm{TV}}\left(\mathcal{L}\left(\frac{\hat{\beta}_j(\rho)-\beta_j^*}{\sqrt{\mathrm{Var}(\hat{\beta}_j(\rho))}}\right), N(0,1)\right) = o(1)$$

where $d_{\mathrm{TV}}(\cdot,\cdot)$ is the total variation distance and $\mathcal{L}(\cdot)$ denotes the law.

We also show a counterexample, namely the one-way analysis of variance problem with non-normal errors, to emphasize that our technical assumptions are not an artifact of the proof but essential to some extent, thereby revealing the non-standard property of the moderate dimensional regime.

This chapter is adapted from my joint work with Professor Noureddine El Karoui and Professor Peter J. Bickel. The paper was published on Probability Theory and Related Fields on December, 2018 (Lei et al. 2018). The idea was originated from Noureddine El Karoui and Peter Bickel as an extension of their earlier works (El Karoui et al. 2011; El Karoui 2013; Bean et al. 2013; El Karoui 2015, 2018). Noureddine El Karoui and Peter Bickel provided joint advising on this work, with joint meetings of the three of us weekly over the course of two years or so.

1.2 Exact Inference for Linear Models

Chapter 2 highlights the difficulty in deriving asymptotics even for a single coordinate with a bounded number of samples per parameter. However, the moderate dimensional regime is quite common in practice as $n/p \leq 50$ in many applications. This may suggest the frangibility of classical asymptotic theory which back up the numbers reported (e.g. p-values, confidence intervals) by standard softwares. It is thus natural to ask if there exists a robust inferential procedure in moderate dimensional regime.

In this chapter, we consider the problem of testing a linear hypothesis, under the linear models studied in Chapter 2, in the form $H_0 : R^T\beta^* = 0$, where $R \in \mathbb{R}^{p\times r}$ is a matrix with full collumn rank. In particular, if $R = (1, 0, \ldots, 0)^T$, then it is equivalent to testing for the first coordinate. Suppose we can find a valid test, then a confidence interval can be obtained for β_1^* by inverting the test, thereby yielding a valid inferential procedure, at least for a single coordinate.

Testing linear hypotheses for linear models is a century-long problem started in 1920s and various qualitatively different strategies have been proposed to tackle this problem, including normal theory based methods (e.g. Fisher 1922; Fisher 1924; Snedecor 1934), permutation-based methods (e.g. Pitman 1937b,a; Pitman 1938), rank-based methods (e.g. Friedman 1937; Theil 1950a), tests based on regression R-estimates (e.g. Hájek 1962), M-estimates (e.g. Huber 1973a), L-estimates (e.g. Bickel 1973), resampling-based methods (e.g. Freedman 1981) and other methods (e.g. Brown and Mood 1951; Daniels 1954; Hartigan 1970; Meinshausen 2015). However, as opposed to the location problems and analysis of variance problems, none of those tests are provably robust to the moderate dimensional regime under reasonably general assumptions.

In this chapter, we propose the cyclic permutation test (CPT), which is an exact non-randomized test for a given confidence level α, for *arbitrary fixed design matrices* and *arbitrary exchangeable errors*, provided that $1/\alpha$ is an integer and $n/p \geq 1/\alpha - 1$. For instance, CPT only requires $n/p \geq 19$ when $\alpha = 0.05$ and thus works in moderate dimensions. Notably, exact tests for general linear hypotheses are rare over the past century and they

are all restricted to linear models with stringent assumptions. By contrast, CPT is exact in finite samples and almost assumption-free except for the exchangeability of errors. We show that CPT has comparable power with existing tests, which may not have guarantee of validity, through extensive numerical experiments. The existence of such a non-standard, assumption-free but powerful test suggests that "problem of distribution" may be tackled by new techniques.

This chapter is adapted from my joint work with Professor Peter J. Bickel. The preprint was posted on ArXiv on July, 2019 (Lei and Bickel 2019).

1.3 Regression Adjustment for Neyman-Rubin Models

In 1923, Jerzy Neyman proposed a model for analyzing agonormic trials in his master thesis (Neyman 1923), which is later known as *randomization model* (Scheffé 1959), and quickly became one of the main pillar in analysis of experimental data (e.g. Kempthorne 1952) and survey sampling (e.g. Cochran 1977). Notably, Donald B. Rubin introduced this model into causal inference, established the framework of potential outcomes and generalized it to observational studies in his seminal work (Rubin 1974). For this reason, the randomization model is also called Neyman-Rubin model in causal inference literature.

Neyman-Rubin model is fundamentally different from linear models. The linear model with fixed designs, marked by analysis of variance, assumes that the treatment assignment is fixed and the outcome is a random variable centered at a linear function of treatment variables. By contrast, the Neyman-Rubin model assumes that the treatment assignment is random with a known distribution and the outcome is a fixed number given the treatment values. To be concrete, given a binary treatment T with observed outcomes Y^{obs}, the linear model assumes $Y_i^{\text{obs}} = \alpha + \beta T_i + \epsilon_i$ where ϵ_i is a random variable while the Neyman-Rubin model assumes $Y_i^{\text{obs}} = Y_i(1)T + Y_i(0)(1-T)$ where $Y_i(1)$ and $Y_i(0)$, called potential outcomes, are two numbers that are either fixed or independent of the treatment T_i. Clearly, the source of randomness is different based on two models. Inference based on linear models was usually classified as *model-based inference*, because it uses the functional relation between the outcome and the treatment, while inference based on Neyman-Rubin models was usually classified as *design-based inference*; see Särndal et al. (e.g. 1978) and Abadie et al. (2017). On the other hand, the inferential targets are usually different for two models. For linear models, the effect of the treatment can be easily defined as β, the coefficient of the treatment variable; for Neyman-Rubin models, the effect of the treatment is usually defined as the average of individual effects, i.e. $1/n \sum_{i=1}^{n} (Y_i(1) - Y_i(0))$. The former can be regarded as a special case of the latter if we treat $Y_i(1) = \alpha + \beta + \epsilon_i$ and $Y_i(0) = \alpha + \epsilon_i$. Inference based on Neyman-Rubin model is more general, though at the cost of the knowledge of the treatment assignment mechanism. Nonetheless, for experimental data, it comes as a free lunch as the assignment mechanism is known by design. Therefore the Neyman-Rubin model is a robust alternative to the linear model in cases where the researcher has more knowledge of the treatment assignment mechanism than that of the functional relation between observed

outcomes and the treatment.

In many applications, baseline covariates are usually collected together with the treatment assignment (e.g. demographic information of experimental subjects). A natural approach is to run a linear regression of the observed outcome on the treatment assignment and the covariates and estimate the effect of the treatment by the corresponding regression coefficient. The fundamental difference between two models does not prevent us from evaluating this procedure, which is clearly valid for a linear model, under the Neyman-Rubin model. However, Freedman (2008b) criticized this approach, showing that it may be less efficient than the naive difference-in-means estimator which completely ignores covariates. He pointed out that the failure is driven by the different sources of randomness between linear models and Neyman-Rubin models. Interestingly, Lin (2013) proposed a simple remedy by adding the interaction terms between the treatment and the covariates into the regression and showed that this estimator is never less efficient than the difference-in-means estimator in the asymptotic regime where the number of covariates p stays fixed while the sample size n tends to infinity.

Based on my experience in linear models as mentioned in the last two subsections, the asymptotics based on fixed-p regime may not be reliable. For a real problem with $n = 1000$ and $p = 50$, is the asymptotic result a plausible approximation? Bloniarz et al. (2016) took the first in a high-dimensional setting where $p \gg n$. However they considered a different estimator and assumed an approximately sparse relation between the potential outcomes and the covariates. Instead, we consider Lin (2013) in a more classical setting where no assumption is imposed on the potential outcomes except some regularity conditions involving the finite sample moments. Specifically, for completely randomized experiments, we show that Lin (2013)'s estimator is consistent when $\kappa \log p \to 0$ and asymptotically normal when $\kappa p \to 0$ under mild moment conditions, where κ is the maximum leverage score of the covariate matrix. In the favorable case where leverage scores are all close together, his estimator is consistent when $p = o(n/\log n)$ and is asymptotically normal when $p = o(n^{1/2})$. Beyond this regime, we find that the estimator may have a non-negligible bias. For this reason, we propose a bias-corrected estimator that is consistent when $\kappa \log p \to 0$ and is asymptotically normal, with the same variance in the fixed-p regime, when $\kappa^2 p \log p \to 0$. In the favorable case, the latter condition reduces to $p = o(n^{2/3}/(\log n)^{1/3})$. Our analyses require novel concentration inequalities for sampling without replacement, driven by modern probability theory.

This chapter is adapted from my joint work with Professor Peng Ding. The preprint was posted on ArXiv on June, 2018 (Lei and Ding 2018).

Chapter 2

Regression M-Estimates in Moderate Dimensions

2.1 Introduction

High-dimensional statistics has a long history (Huber 1973a; Wachter 1976, 1978) with considerable renewed interest over the last two decades. In many applications, the researcher collects data which can be represented as a matrix, called a design matrix and denoted by $X \in \mathbb{R}^{n\times p}$, as well as a response vector $y \in \mathbb{R}^n$ and aims to study the connection between X and y. The linear model is among the most popular models as a starting point of data analysis in various fields. A linear model assumes that

$$y = X\beta^* + \epsilon, \tag{2.1}$$

where $\beta^* \in \mathbb{R}^p$ is the coefficient vector which measures the marginal contribution of each predictor and ϵ is a random vector which captures the unobserved errors.

The aim of this chapter is to provide valid inferential results for features of β^*. For example, a researcher might be interested in testing whether a given predictor has a negligible effect on the response, or equivalently whether $\beta_j^* = 0$ for some j. Similarly, linear contrasts of β^* such as $\beta_1^* - \beta_2^*$ might be of interest in the case of the group comparison problem in which the first two predictors represent the same feature but are collected from two different groups.

An M-estimator, defined as

$$\hat{\beta}(\rho) = \underset{\beta\in\mathbb{R}^p}{\arg\min} \frac{1}{n}\sum_{i=1}^{n} \rho(y_i - x_i^T\beta) \tag{2.2}$$

where ρ denotes a loss function, is among the most popular estimators used in practice (Relles 1968; Huber 1973a). In particular, if $\rho(x) = \frac{1}{2}x^2$, $\hat{\beta}(\rho)$ is the famous Least Square Estimator (LSE). We intend to explore the distribution of $\hat{\beta}(\rho)$, based on which we can achieve the inferential goals mentioned above.

The most well-studied approach is the asymptotic analysis, which assumes that the scale of the problem grows to infinity and use the limiting result as an approximation. In regression problems, the scale parameter of a problem is the sample size n and the number of predictors p. The classical approach is to fix p and let n grow to infinity. It has been shown (Relles 1968; Yohai 1972; Huber 1972; Huber 1973a) that $\hat{\beta}(\rho)$ is consistent in terms of L_2 norm and asymptotically normal in this regime. The asymptotic variance can be then approximated by the bootstrap (Bickel and Freedman 1981). Later on, the studies are extended to the regime in which both n and p grow to infinity but p/n converges to 0 (Yohai and Maronna 1979b; Portnoy 1984, 1985, 1986, 1987; Mammen 1989). The consistency, in terms of the L_2 norm, the asymptotic normality and the validity of the bootstrap still hold in this regime. Based on these results, we can construct a 95% confidence interval for β_{0j} simply as $\hat{\beta}_j(\rho) \pm 1.96\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_j(\rho))}$ where $\widehat{\mathrm{Var}}(\hat{\beta}_j(\rho))$ is calculated by bootstrap. Similarly we can calculate p-values for the hypothesis testing procedure.

We ask whether the inferential results developed under the low-dimensional assumptions and the software built on top of them can be relied on for moderate and high-dimensional analysis? Concretely, if in a study $n = 50$ and $p = 40$, can the software built upon the assumption that $p/n \simeq 0$ be relied on when $p/n = .8$? Results in random matrix theory (Marčenko and Pastur 1967) already offer an answer in the negative side for many PCA-related questions in multivariate statistics. The case of regression is more subtle: For instance for least-squares, standard degrees of freedom adjustments effectively take care of many dimensionality-related problems. But this nice property does not extend to more general regression M-estimates.

Once these questions are raised, it becomes very natural to analyze the behavior and performance of statistical methods in the regime where p/n is fixed. Indeed, it will help us to keep track of the inherent statistical difficulty of the problem when assessing the variability of our estimates. In other words, we assume in this chapter that $p/n \to \kappa > 0$ while let n grows to infinity. Due to identifiability issues, it is impossible to make inference on β^* if $p > n$ without further structural or distributional assumptions. We discuss this point in details in Section 2.2.3. Thus we consider the regime where $p/n \to \kappa \in (0, 1)$. We call it the moderate p/n regime. This regime is also the natural regime in random matrix theory (Marčenko and Pastur 1967; Wachter 1978; Johnstone 2001; Bai and Silverstein 2010). It has been shown that the asymptotic results derived in this regime sometimes provide an extremely accurate approximation to finite sample distributions of estimators at least in certain cases (Johnstone 2001) where n and p are both small.

2.1.1 Qualitatively Different Behavior of Moderate p/n Regime

First, $\hat{\beta}(\rho)$ is no longer consistent in terms of L_2 norm and the risk $\mathbb{E}\|\hat{\beta}(\rho) - \beta^*\|^2$ tends to a non-vanishing quantity determined by κ, the loss function ρ and the error distribution through a complicated system of non-linear equations (El Karoui et al. 2011; El Karoui 2013, 2015; Bean et al. 2012). This L_2-inconsistency prohibits the use of standard perturbation-

analytic techniques to assess the behavior of the estimator. It also leads to qualitatively different behaviors for the residuals in moderate dimensions; in contrast to the low-dimensional case, they cannot be relied on to give accurate information about the distribution of the errors. However, this seemingly negative result does not exclude the possibility of inference since $\hat{\beta}(\rho)$ is still consistent in terms of $L_{2+\nu}$ norms for any $\nu > 0$ and in particular in L_∞ norm. Thus, we can at least hope to perform inference on each coordinate.

Second, classical optimality results do not hold in this regime. In the regime $p/n \to 0$, the maximum likelihood estimator is shown to be optimal (Huber 1964; Huber 1972; Bickel and Doksum 2015). In other words, if the error distribution is known then the M-estimator associated with the loss $\rho(\cdot) = -\log f_\epsilon(\cdot)$ is asymptotically efficient, provided the design is of appropriate type, where $f_\epsilon(\cdot)$ is the density of entries of ϵ. However, in the moderate p/n regime, it has been shown that the optimal loss is no longer the log-likehood but an other function with a complicated but explicit form (Bean et al. 2013), at least for certain designs. The suboptimality of maximum likelihood estimators suggests that classical techniques fail to provide valid intuition in the moderate p/n regime.

Third, the joint asymptotic normality of $\hat{\beta}(\rho)$, as a p-dimensional random vector, may be violated for a fixed design matrix X. This has been proved for least-squares by Huber (1973a) in his pioneering work. For general M-estimators, this negative result is a simple consequence of the results of El Karoui et al. (2011): They exhibit an ANOVA design (see below) where even marginal fluctuations are not Gaussian. By contrast, for random design, they show that $\hat{\beta}(\rho)$ is jointly asymptotically normal when the design matrix is elliptical with general covariance by using the non-asymptotic stochastic representation for $\hat{\beta}(\rho)$ as well as elementary properties of vectors uniformly distributed on the uniform sphere in $\mathbb{R}^p$; See section 2.2.3 of El Karoui et al. (2011) or the supplementary material of Bean et al. (2013) for details. This does not contradict Huber (1973a)'s negative result in that it takes the randomness from both X and ϵ into account while Huber (1973a)'s result only takes the randomness from ϵ into account. Later, El Karoui (2015) shows that each coordinate of $\hat{\beta}(\rho)$ is asymptotically normal for a broader class of random designs. This is also an elementary consequence of the analysis in El Karoui (2013). However, to the best of our knowledge, beyond the ANOVA situation mentioned above, there are no distributional results for fixed design matrices. This is the topic of this chapter.

Last but not least, bootstrap inference fails in this moderate-dimensional regime. This has been shown by Bickel and Freedman (1983b) for least-squares and residual bootstrap in their influential work. Recently, El Karoui and Purdom (2015) studied the results to general M-estimators and showed that all commonly used bootstrapping schemes, including pairs-bootstrap, residual bootstrap and jackknife, fail to provide a consistent variance estimator and hence valid inferential statements. These latter results even apply to the marginal distributions of the coordinates of $\hat{\beta}(\rho)$. Moreover, there is no simple, design independent, modification to achieve consistency (El Karoui and Purdom 2015).

2.1.2 Our Contributions

In summary, the behavior of the estimators we consider in this chapter is completely different in the moderate p/n regime from its counterpart in the low-dimensional regime. As discussed in the next section, moving one step further in the moderate p/n regime is interesting from both the practical and theoretical perspectives. Our main contribution is to establish coordinate-wise asymptotic normality of $\hat{\beta}(\rho)$ for certain *fixed design matrices* X in this regime under technical assumptions. The following theorem informally states our main result.

Theorem 2.1.1 (Informal Version of Theorem 2.3.1 in Section 2.3)**.** *Under appropriate conditions on the design matrix X, the distribution of ϵ and the loss function ρ, as $p/n \to \kappa \in (0,1)$, while $n \to \infty$,*

$$\max_{1\le j\le p} d_{\mathrm{TV}}\left(\mathcal{L}\left(\frac{\hat{\beta}_j(\rho)-\mathbb{E}\hat{\beta}_j(\rho)}{\sqrt{\mathrm{Var}(\hat{\beta}_j(\rho))}}\right), N(0,1)\right) = o(1)$$

where $d_{\mathrm{TV}}(\cdot,\cdot)$ is the total variation distance and $\mathcal{L}(\cdot)$ denotes the law.

It is worth mentioning that the above result can be extended to finite dimensional linear contrasts of $\hat{\beta}$. For instance, one might be interested in making inference on $\beta_1^* - \beta_2^*$ in the problems involving the group comparison. The above result can be extended to give the asymptotic normality of $\hat{\beta}_1 - \hat{\beta}_2$.

Besides the main result, we have several other contributions. First, we use a new approach to establish asymptotic normality. Our main technique is based on the second-order Poincaré inequality (SOPI), developed by Chatterjee (2009) to derive, among many other results, the fluctuation behavior of linear spectral statistics of random matrices. In contrast to classical approaches such as the Lindeberg-Feller central limit theorem, the second-order Poincaré inequality is capable of dealing with nonlinear and potentially implicit functions of independent random variables. Moreover, we use different expansions for $\hat{\beta}(\rho)$ and residuals based on double leave-one-out ideas introduced in El Karoui et al. (2011), in contrast to the classical perturbation-analytic expansions. See aforementioned paper and follow-ups. An informal interpretation of the results of Chatterjee (2009) is that if the Hessian of the nonlinear function of random variables under consideration is sufficiently small, this function acts almost linearly and hence a standard central limit theorem holds.

Second, to the best of our knowledge this is the first inferential result for fixed (non ANOVA-like) design in the moderate p/n regime. Fixed designs arise naturally from an experimental design or a conditional inference perspective. That is, inference is ideally carried out without assuming randomness in predictors; see Section 2.2.2 for more details. We clarify the regularity conditions for coordinate-wise asymptotic normality of $\hat{\beta}(\rho)$ explicitly, which are checkable for LSE and also checkable for general M-estimators if the error distribution is known. We also prove that these conditions are satisfied with by a broad class of designs.

The ANOVA-like design described in Section 2.3.3 exhibits a situation where the distribution of $\hat{\beta}_j(\rho)$ is not going to be asymptotically normal. As such the results of Theorem 2.3.1 below are somewhat surprising.

For complete inference, we need both the asymptotic normality and the asymptotic bias and variance. Under suitable symmetry conditions on the loss function and the error distribution, it can be shown that $\hat{\beta}(\rho)$ is unbiased (see Section 2.3.2 for details) and thus it is left to derive the asymptotic variance. As discussed at the end of Section 2.1.1, classical approaches, e.g. bootstrap, fail in this regime. For least-squares, classical results continue to hold and we discuss it in section 2.5 for the sake of completeness. However, for M-estimators, there is no closed-form result. We briefly touch upon the variance estimation in Section 2.3.4. The derivation for general situations is beyond the scope of this chapter and left to the future research.

2.1.3 Outline

The rest of the chapter is organized as follows: In Section 2.2, we clarify details which are mentioned in the current section. In Section 2.3, we state the main result (Theorem 2.3.1) formally and explain the technical assumptions. Then we show several examples of random designs which satisfy the assumptions with high probability. In Section 4, we introduce our main technical tool, second-order Poincaré inequality (Chatterjee 2009), and apply it on M-estimators as the first step to prove Theorem 2.3.1. Since the rest of the proof of Theorem 2.3.1 is complicated and lengthy, we illustrate the main ideas in Appendix A.1. The rigorous proof is left to Appendix A.2. In Section 2.5, we provide reminders about the theory of least-squares estimation for the sake of completeness, by taking advantage of its explicit form. In Section 2.6, we display the numerical results. The proof of other results are stated in Appendix A.3 and more numerical experiments are presented in Appendix A.4.

2.2 More Details on Background

2.2.1 Moderate p/n Regime: a more informative type of asymptotics?

In Section 2.1, we mentioned that the ratio p/n measures the difficulty of statistical inference. The moderate p/n regime provides an approximation of finite sample properties with the difficulties fixed at the same level as the original problem. Intuitively, this regime should capture more variation in finite sample problems and provide a more accurate approximation. We will illustrate this via simulation.

Consider a study involving 50 participants and 40 variables; we can either use the asymptotics in which p is fixed to be 40, n grows to infinity or p/n is fixed to be 0.8, and n grows to infinity to perform approximate inference. Current software rely on low-dimensional asymptotics for inferential tasks, but there is no evidence that they yield more accurate inferential

statements than the ones we would have obtained using moderate dimensional asymptotics. In fact, numerical evidence (Johnstone 2001; El Karoui et al. 2013; Bean et al. 2013) show that the reverse is true.

We exhibit a further numerical simulation showing that. Consider a case that $n = 50$, ϵ has i.i.d. entries and X is one realization of a matrix generated with i.i.d. gaussian (mean 0, variance 1) entries. For $\kappa \in \{0.1, 0.2, \ldots, 0.9\}$ and different error distributions, we use the Kolmogorov-Smirnov (KS) statistics to quantify the distance between the finite sample distribution and two types of asymptotic approximation of the distribution of $\hat{\beta}_1(\rho)$.

Specifically, we use the Huber loss function $\rho_{\text{Huber},k}$ with default parameter $k = 1.345$ (Huber 1981), i.e.

$$\rho_{\text{Huber},k}(x) = \begin{cases} \frac{1}{2}x^2 & |x| \le k \\ k(|x| - \frac{1}{2}k) & |x| > k \end{cases}$$

Specifically, we generate three design matrices $X^{(0)}$, $X^{(1)}$ and $X^{(2)}$: $X^{(0)}$ for small sample case with a sample size $n = 50$ and a dimension $p = n\kappa$; $X^{(1)}$ for low-dimensional asymptotics (p fixed) with a sample size $n = 1000$ and a dimension $p = 50\kappa$; and $X^{(2)}$ for moderate-dimensional asymptotics (p/n fixed) with a sample size $n = 1000$ and a dimension $p = n\kappa$. Each of them is generated as one realization of an i.i.d. standard gaussian design and then treated as fixed across $K = 100$ repetitions. For each design matrix, vectors ϵ of appropriate length are generated with i.i.d. entries. The entry has either a standard normal distribution, or a t_3-distribution, or a standard Cauchy distribution, i.e. t_1. Then we use ϵ as the response, or equivalently assume $\beta^* = 0$, and obtain the M-estimators $\hat{\beta}^{(0)}, \hat{\beta}^{(1)}, \hat{\beta}^{(2)}$. Repeating this procedure for $K = 100$ times results in K replications in three cases. Then we extract the first coordinate of each estimator, denoted by $\{\hat{\beta}^{(0)}_{k,1}\}_{k=1}^K, \{\hat{\beta}^{(1)}_{k,1}\}_{k=1}^K, \{\hat{\beta}^{(2)}_{k,1}\}_{k=1}^K$. Then the two-sample Kolmogorov-Smirnov statistics can be obtained by

$$\text{KS}_1 = \sqrt{\frac{n}{2}} \max_x |\hat{F}^{(0)}_n(x) - \hat{F}^{(1)}_n(x)|, \quad \text{KS}_2 = \sqrt{\frac{n}{2}} \max_x |\hat{F}^{(0)}_n(x) - \hat{F}^{(2)}_n(x)|,$$

where $\hat{F}^{(r)}_n$ is the empirical distribution of $\{\hat{\beta}^{(r)}_{k,1}\}_{k=1}^K$. We can then compare the accuracy of two asymptotic regimes by comparing KS_1 and KS_2. The smaller the value of KS_i, the better the approximation.

Figure 2.1 displays the results for these error distributions. We see that for gaussian errors and even t_3 errors, the p/n-fixed/moderate-dimensional approximation is uniformly more accurate than the widely used p-fixed/low-dimensional approximation. For Cauchy errors, the low-dimensional approximation performs better than the moderate-dimensional one when p/n is small but worsens when the ratio is large especially when p/n is close to 1. Moreover, when p/n grows, the two approximations have qualitatively different behaviors: the p-fixed approximation becomes less and less accurate while the p/n-fixed approximation does not suffer much deterioration when p/n grows. The qualitative and quantitative differences of these two approximations reveal the practical importance of exploring the p/n-fixed asymptotic regime. (See also Johnstone (2001).)

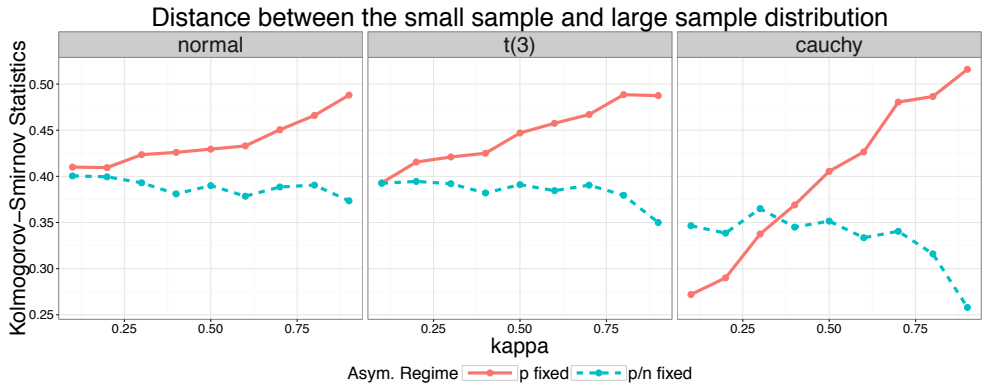


Figure 2.1: Axpproximation accuracy of p-fixed asymptotics and p/n-fixed asymptotics: each column represents an error distribution; the x-axis represents the ratio κ of the dimension and the sample size and the y-axis represents the Kolmogorov-Smirnov statistic; the red solid line corresponds to p-fixed approximation and the blue dashed line corresponds to p/n-fixed approximation.

2.2.2 Random vs fixed design?

As discussed in Section 2.1.1, assuming a fixed design or a random design could lead to qualitatively different inferential results.

In the random design setting, X is considered as being generated from a super population. For example, the rows of X can be regarded as an i.i.d. sample from a distribution known, or partially known, to the researcher. In situations where one uses techniques such as cross-validation (Stone 1974), pairs bootstrap in regression (Efron and Efron 1982) or sample splitting (Wasserman and Roeder 2009), the researcher effectively assumes exchangeability of the data $(x_i^T, y_i)_{i=1}^n$. Naturally, this is only compatible with an assumption of random design. Given the extremely widespread use of these techniques in contemporary machine learning and statistics, one could argue that the random design setting is the one under which most of modern statistics is carried out, especially for prediction problems. Furthermore, working under a random design assumption forces the researcher to take into account two sources of randomness as opposed to only one in the fixed design case. Hence working under a random design assumption should yield conservative confidence intervals for β_j^*.

In other words, in settings where the researcher collects data without control over the values of the predictors, the random design assumption is arguably the more natural one of the two.

However, it has now been understood for almost a decade that common random design assumptions in high-dimension (e.g. $x_i = \Sigma^{1/2} z_i$ where $z_{i,j}$'s are i.i.d with mean 0 and variance 1 and a few moments and Σ "well behaved") suffer from considerable geometric limitations, which have substantial impacts on the performance of the estimators considered

in this chapter (El Karoui et al. 2011). As such, confidence statements derived from that kind of analysis can be relied on only after performing a few graphical tests on the data (see El Karoui (2010)). These geometric limitations are simple consequences of the concentration of measure phenomenon (Ledoux 2001).

On the other hand, in the fixed design setting, X is considered a fixed matrix. In this case, the inference only takes the randomness of ϵ into consideration. This perspective is popular in several situations. The first one is the experimental design. The goal is to study the effect of a set of factors, which can be controlled by the experimenter, on the response. In contrast to the observational study, the experimenter can design the experimental condition ahead of time based on the inference target. For instance, a one-way ANOVA design encodes the covariates into binary variables (see Section 2.3.3 for details) and it is fixed prior to the experiment. Other examples include two-way ANOVA designs, factorial designs, Latin-square designs, etc. (Scheffe 1999).

Another situation which is concerned with fixed design is the survey sampling where the inference is carried out conditioning on the data (Cochran 1977). Generally, in order to avoid unrealistic assumptions, making inference conditioning on the design matrix X is necessary. Suppose the linear model (2.1) is true and identifiable (see Section 2.2.3 for details), then all information of β^* is contained in the conditional distribution $\mathcal{L}(y|X)$ and hence the information in the marginal distribution $\mathcal{L}(X)$ is redundant. The conditional inference framework is more robust to the data generating procedure due to the irrelevance of $\mathcal{L}(X)$.

Also, results based on fixed design assumptions may be preferable from a theoretical point of view in the sense that they could potentially be used to establish corresponding results for certain classes of random designs. Specifically, given a marginal distribution $\mathcal{L}(X)$, one only has to prove that $\mathcal{X}$ satisfies the assumptions for fixed design with high probability.

In conclusion, fixed and random design assumptions play complementary roles in moderate dimensional settings. We focus on the least understood of the two, the fixed design case, in this chapter.

2.2.3 Modeling and Identification of Parameters

The problem of identifiability is especially important in the fixed design case. Define $\beta^*(\rho)$ in the population as

$$\beta^*(\rho) = \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\rho(y_i - x_i^T \beta). \tag{2.3}$$

One may ask whether $\beta^*(\rho) = \beta^*$ regardless of ρ in the fixed design case. We provide an affirmative answer in the following proposition by assuming that ϵ_i has a symmetric distribution around 0 and ρ is even.

Proposition 2.2.1. *Suppose X has a full column rank and $\epsilon_i \overset{d}{=} -\epsilon_i$ for all i. Further assume ρ is an even convex function such that for any $i = 1, 2, \ldots$ and $\alpha \neq 0$,*

$$\frac{1}{2}\left(\mathbb{E}\rho(\epsilon_i - \alpha) + \mathbb{E}\rho(\epsilon_i + \alpha)\right) > \mathbb{E}\rho(\epsilon_i). \tag{2.4}$$

Then $\beta^*(\rho) = \beta^*$ regardless of the choice of ρ.

The proof is left to Appendix A.3. It is worth mentioning that Proposition 2.2.1 only requires the marginals of ϵ to be symmetric but does not impose any constraint on the dependence structure of ϵ. Further, if ρ is strongly convex, then for all $\alpha \neq 0$,

$$\frac{1}{2}\left(\rho(x - \alpha) + \rho(x + \alpha)\right) > \rho(x).$$

As a consequence, the condition (2.4) is satisfied provided that ϵ_i is non-zero with positive probability.

If ϵ is asymmetric, we may still be able to identify β^* if ϵ_i are i.i.d. random variables. In contrast to the last case, we should incorporate an intercept term as a shift towards the centroid of ρ. More precisely, we define $\alpha^*(\rho)$ and $\beta^*(\rho)$ as

$$(\alpha^*(\rho), \beta^*(\rho)) = \underset{\alpha\in\mathbb{R},\beta\in\mathbb{R}^p}{\arg\min} \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\rho(y_i - \alpha - x_i^T\beta).$$

Proposition 2.2.2. *Suppose $(\mathbf{1}, X)$ is of full column rank and ϵ_i are i.i.d. such that $\mathbb{E}\rho(\epsilon_1 - \alpha)$ as a function of α has a unique minimizer $\alpha(\rho)$. Then $\beta^*(\rho)$ is uniquely defined with $\beta^*(\rho) = \beta^*$ and $\alpha^*(\rho) = \alpha(\rho)$.*

The proof is left to Appendix A.3. For example, let $\rho(z) = |z|$. Then the minimizer of $\mathbb{E}\rho(\epsilon_1 - a)$ is a median of ϵ_1, and is unique if ϵ_1 has a positive density. It is worth pointing out that incorporating an intercept term is essential for identifying β^*. For instance, in the least-square case, $\beta^*(\rho)$ no longer equals to β^* if $\mathbb{E}\epsilon_i \neq 0$. Proposition 2.2.2 entails that the intercept term guarantees $\beta^*(\rho) = \beta^*$, although the intercept term itself depends on the choice of ρ unless more conditions are imposed.

If ϵ_i's are neither symmetric nor i.i.d., then β^* cannot be identified by the previous criteria because $\beta^*(\rho)$ depends on ρ. Nonetheless, from a modeling perspective, it is popular and reasonable to assume that ϵ_i's are symmetric or i.i.d. in many situations. Therefore, Proposition 2.2.1 and Proposition 2.2.2 justify the use of M-estimators in those cases and M-estimators derived from different loss functions can be compared because they are estimating the same parameter.

2.3 Main Results

2.3.1 Notation and Assumptions

Let $x_i^T \in \mathbb{R}^{1\times p}$ denote the i-th row of X and $X_j \in \mathbb{R}^{n\times 1}$ denote the j-th column of X. Throughout the chapter we will denote by $X_{ij} \in \mathbb{R}$ the (i, j)-th entry of X, by $X_{[j]} \in \mathbb{R}^{n\times(p-1)}$

the design matrix X after removing the j-th column, and by $x_{i,[j]}^T \in \mathbb{R}^{1\times(p-1)}$ the vector x_i^T after removing j-th entry. The M-estimator $\hat{\beta}(\rho)$ associated with the loss function ρ is defined as

$$\hat{\beta}(\rho) = \arg\min_{\beta\in\mathbb{R}^p} \frac{1}{n}\sum_{k=1}^n \rho(y_k - x_k^T\beta) = \arg\min_{\beta\in\mathbb{R}^p} \frac{1}{n}\sum_{k=1}^n \rho(\epsilon_k - x_k^T(\beta-\beta^*)) \tag{2.5}$$

We define $\psi = \rho'$ to be the first derivative of ρ. We will write $\hat{\beta}(\rho)$ simply $\hat{\beta}$ when no confusion can arise.

When the original design matrix X does not contain an intercept term, we can simply replace X by $(\mathbf{1}, X)$ and augment β into a $(p+1)$-dimensional vector $(\alpha, \beta^T)^T$. Although being a special case, we will discuss the question of intercept in Section 2.3.2 due to its important role in practice.

Equivariance and reduction to the null case

Notice that our target quantity $\frac{\hat{\beta}_j - \mathbb{E}\hat{\beta}_j}{\sqrt{\mathrm{Var}(\hat{\beta}_j)}}$ is invariant to the choice of β^*, provided that β^* is identifiable as discussed in Section 2.2.3, we can assume $\beta^* = 0$ without loss of generality. In this case, we assume in particular that the design matrix X has full column rank. Then $y_k = \epsilon_k$ and

$$\hat{\beta} = \arg\min_{\beta\in\mathbb{R}^p} \frac{1}{n}\sum_{k=1}^n \rho(\epsilon_k - x_k^T\beta).$$

Similarly we define the leave-j-th-predictor-out version as

$$\hat{\beta}_{[j]} = \arg\min_{\beta\in\mathbb{R}^{p-1}} \frac{1}{n}\sum_{k=1}^n \rho(\epsilon_k - x_{k,[j]}^T\beta).$$

Based on these notations we define the full residuals R_k as

$$R_k = \epsilon_k - x_k^T\hat{\beta}, \quad k = 1, 2, \ldots, n$$

and the leave-j-th-predictor-out residual as

$$r_{k,[j]} = \epsilon_k - x_{k,[j]}^T\hat{\beta}_{[j]}, \quad k = 1, 2, \ldots, n, \quad j = 1, \ldots, p.$$

Three $n \times n$ diagonal matrices are defined as

$$D = \mathrm{diag}(\psi'(R_k))_{k=1}^n, \quad \tilde{D} = \mathrm{diag}(\psi''(R_k))_{k=1}^n, \quad D_{[j]} = \mathrm{diag}(\psi'(r_{k,[j]}))_{k=1}^n. \tag{2.6}$$

We say a random variable Z is σ^2-sub-gaussian if for any $\lambda \in \mathbb{R}$,

$$\mathbb{E}e^{\lambda Z} \le e^{\frac{\lambda^2\sigma^2}{2}}.$$

In addition, we use $J_n \subset \{1, \ldots, p\}$ to represent the indices of parameters which are of interest. Intuitively, more entries in J_n would require more stringent conditions for the asymptotic normality.

Finally, we adopt Landau's notation $(O(\cdot), o(\cdot), O_p(\cdot), o_p(\cdot))$. In addition, we say $a_n = \Omega(b_n)$ if $b_n = O(a_n)$ and similarly, we say $a_n = \Omega_p(b_n)$ if $b_n = O_p(a_n)$. To simplify the logarithm factors, we use the symbol $\text{polyLog}(n)$ to denote any factor that can be upper bounded by $(\log n)^\gamma$ for some $\gamma > 0$. Similarly, we use $\frac{1}{\text{polyLog}(n)}$ to denote any factor that can be lower bounded by $\frac{1}{(\log n)^{\gamma'}}$ for some $\gamma' > 0$.

2.3.2 Technical Assumptions and main result

Before stating the assumptions, we need to define several quantities of interest. Let

$$\lambda_+ = \lambda_{\max}\left(\frac{X^T X}{n}\right), \quad \lambda_- = \lambda_{\min}\left(\frac{X^T X}{n}\right)$$

be the largest (resp. smallest) eigenvalue of the matrix $\frac{X^T X}{n}$. Let $e_i \in \mathbb{R}^n$ be the i-th canonical basis vector and

$$h_{j,0} \triangleq (\psi(r_{1,[j]}), \ldots, \psi(r_{n,[j]}))^T, \quad h_{j,1,i} \triangleq (I - D_{[j]}X_{[j]}(X_{[j]}^T D_{[j]} X_{[j]})^{-1} X_{[j]}^T) e_i.$$

Finally, let

$$\Delta_C = \max\left\{\max_{j\in J_n} \frac{|h_{j,0}^T X_j|}{\|h_{j,0}\|_2}, \max_{i\le n, j\in J_n} \frac{|h_{j,1,i}^T X_j|}{\|h_{j,1,i}\|_2}\right\},$$
$$Q_j = \text{Cov}(h_{j,0})$$

Based on the quantities defined above, we state our technical assumptions on the design matrix X followed by the main result. A detailed explanation of the assumptions follows.

A1 $\rho(0) = \psi(0) = 0$ and there exists positive numbers $K_0 = \Omega\left(\frac{1}{\text{polyLog}(n)}\right)$, $K_1, K_2 = O\left(\text{polyLog}(n)\right)$, such that for any $x \in \mathbb{R}$,

$$K_0 \le \psi'(x) \le K_1, \quad \left|\frac{d}{dx}(\sqrt{\psi'}(x))\right| = \frac{|\psi''(x)|}{\sqrt{\psi'(x)}} \le K_2;$$

A2 $\epsilon_i = u_i(W_i)$ where $(W_1, \ldots, W_n) \sim N(0, I_{n\times n})$ and u_i are smooth functions with $\|u_i'\|_\infty \le c_1$ and $\|u_i''\|_\infty \le c_2$ for some $c_1, c_2 = O(\text{polyLog}(n))$. Moreover, assume $\min_i \text{Var}(\epsilon_i) = \Omega\left(\frac{1}{\text{polyLog}(n)}\right)$.

A3 $\lambda_+ = O(\text{polyLog}(n))$ and $\lambda_- = \Omega\left(\frac{1}{\text{polyLog}(n)}\right)$;

A4 $\min_{j\in J_n} \frac{X_j^T Q_j X_j}{\text{tr}(Q_j)} = \Omega\left(\frac{1}{\text{polyLog}(n)}\right)$;

A5 $\mathbb{E}\Delta_C^8 = O\left(\text{polyLog}(n)\right)$.

Theorem 2.3.1. *Under assumptions* $\boldsymbol{A}1 - \boldsymbol{A}5$*, as* $p/n \to \kappa$ *for some* $\kappa \in (0,1)$*, while* $n \to \infty$*,*

$$\max_{j \in J_n} d_{\mathrm{TV}} \left(\mathcal{L} \left(\frac{\hat{\beta}_j - \mathbb{E}\hat{\beta}_j}{\sqrt{\mathrm{Var}(\hat{\beta}_j)}} \right), N(0,1) \right) = o(1),$$

where $d_{\mathrm{TV}}(P,Q) = \sup_A |P(A) - Q(A)|$ *is the total variation distance.*

We provide several examples where our assumptions hold in Section 2.3.3. We also provide an example where the asymptotic normality does not hold in Section 2.3.3. This shows that our assumptions are not just artifacts of the proof technique we developed, but that there are (probably many) situations where asymptotic normality will not hold, even coordinate-wise.

Discussion of Assumptions

Now we discuss assumptions **A**1 - **A**5. Assumption **A**1 implies the boundedness of the first-order and the second-order derivatives of ψ. The upper bounds are satisfied by most loss functions including the L_2 loss, the smoothed L_1 loss, the smoothed Huber loss, etc. The non-zero lower bound K_0 implies the strong convexity of ρ and is required for technical reasons. It can be removed by considering first a ridge-penalized M-estimator and taking appropriate limits as in El Karoui (2013, 2015). In addition, in this chapter we consider the smooth loss functions and the results can be extended to non-smooth case via approximation.

For unregularized M-estimators, the strong convexity is also assumed by other works El Karoui (2013) and Donoho and Montanari (2016). However, we believe that this assumption is unnecessary and can be removed at least for well-behaved design matrices. In fact, we can extend our results to strictly convex loss functions, where ψ' is always positive by imposing slightly stronger assumptions on the designs. This includes the class of optimal loss functions in the moderate p/n regime derived in (Bean et al. 2013). However, the proofs are very delicate and beyond the scope of this chapter so we plan to leave it in our future works.

Assumption **A**2 was proposed in Chatterjee (2009) for the second-order Poincaré inequality discussed in Section 2.4.1. It means that the results apply to non-Gaussian distributions, such as the uniform distribution on $[0,1]$ by taking $u_i = \Phi$, the cumulative distribution function of standard normal distribution. Through the gaussian concentration (Ledoux 2001), we see that **A**2 implies that ϵ_i are c_1^2-sub-gaussian. Thus **A**2 controls the tail behavior of ϵ_i. The bounds on the infinity norm of u_i' and u_i'' are required only for the direct application of Chatterjee's results. In fact, a look at his proof suggests that one can obtain a similar result to his Second-Order Poincaré inequality involving moment bounds on $u_i'(W_i)$ and $u_i''(W_i)$. This would be a way to weaken our assumptions to permit to have the heavy-tailed distributions expected in robustness studies. This requires substantial work and an extension of the main results of Chatterjee (2009). Because the technical part of the chapter is already long, we leave this interesting statistical question to future works.

On the other hand, since we are considering strongly convex loss-functions, it is not completely unnatural to restrict our attention to light-tailed errors. Furthermore, efficiency - and not only robustness - questions are one of the main reasons to consider these estimators in the moderate-dimensional context. The potential gains in efficiency obtained by considering regression M-estimates (Bean et al. 2013) apply in the light-tailed context, which further justify our interest in this theoretical setup.

Assumption **A**3 is completely checkable since it only depends on X. It controls the singularity of the design matrix. Under **A**1 and **A**3, it can be shown that the objective function is strongly convex with curvature (the smallest eigenvalue of the Hessian matrix) lower bounded by $\Omega\left(\frac{1}{\text{polyLog}(n)}\right)$ everywhere.

Assumption **A**4 is controlling the left tail of quadratic forms. It is fundamentally connected to aspects of the concentration of measure phenomenon (Ledoux 2001). This condition is proposed and emphasized under the random design setting by El Karoui et al. (2013). Essentially, it means that for a matrix Q_j ,which does not depend on X_j, the quadratic form $X_j^T Q_j X_j$ should have the same order as $\text{tr}(Q_j)$.

Assumption **A**5 is proposed by El Karoui (2013) under the random design settings. It is motivated by leave-one-predictor-out analysis. Note that Δ_C is the maximum of linear contrasts of X_j, whose coefficients do not depend on X_j. It is easily checked for design matrix X which is a realization of a random matrix with i.i.d sub-gaussian entries for instance.

Remark 2.3.2. *In certain applications, it is reasonable to make the following additional assumption:*

***A**6 ρ is an even function and ϵ_i's have symmetric distributions.*

*Although assumption **A**6 is not necessary to Theorem 2.3.1, it can simplify the result. Under assumption **A**6, when X is full rank, we have, if $\stackrel{d}{=}$ denotes equality in distribution,*

$$\begin{aligned}\hat{\beta} - \beta^* &= \arg\min_{\eta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \rho(\epsilon_i - x_i^T \eta) = \arg\min_{\eta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \rho(-\epsilon_i + x_i^T \eta) \\ &\stackrel{d}{=} \arg\min_{\eta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \rho(\epsilon_i + x_i^T \eta) = \beta^* - \hat{\beta}.\end{aligned}$$

This implies that $\hat{\beta}$ is an unbiased estimator, provided it has a mean, which is the case here. Unbiasedness is useful in practice, since then Theorem 2.3.1 reads

$$\max_{j \in J_n} d_{\text{TV}} \left(\mathcal{L}\left(\frac{\hat{\beta}_j - \beta_j^*}{\sqrt{\text{Var}(\hat{\beta}_j)}} \right), N(0,1) \right) = o(1) \ .$$

For inference, we only need to estimate the asymptotic variance.

An important remark concerning Theorem 2.3.1

When J_n is a subset of $\{1, \ldots, p\}$, the coefficients in J_n^c become nuisance parameters. Heuristically, in order for identifying $\beta^*_{J_n}$, one only needs the subspaces $\text{span}(X_{J_n})$ and $\text{span}(X_{J_n^c})$ to be distinguished and X_{J_n} has a full column rank. Here X_{J_n} denotes the sub-matrix of X with columns in J_n. Formally, let

$$\hat{\Sigma}_{J_n} = \frac{1}{n} X_{J_n}^T (I - X_{J_n^c}(X_{J_n^c}^T X_{J_n^c})^- X_{J_n^c}^T) X_{J_n}$$

where A^- denotes the generalized inverse of A, and

$$\tilde{\lambda}_+ = \lambda_{\max}\left(\hat{\Sigma}_{J_n}\right), \quad \tilde{\lambda}_- = \lambda_{\min}\left(\hat{\Sigma}_{J_n}\right).$$

Then $\hat{\Sigma}_{J_n}$ characterizes the behavior of X_{J_n} after removing the effect of $X_{J_n^c}$. In particular, we can modify the assumption **A**3 by

A3* $\tilde{\lambda}_+ = O(\text{polyLog}(n))$ and $\tilde{\lambda}_- = \Omega\left(\frac{1}{\text{polyLog}(n)}\right)$.

Then we are able to derive a stronger result in the case where $|J_n| < p$ than Theorem 2.3.1 as follows.

Corollary 2.3.3. *Under assumptions **A**1-2, **A**4-5 and **A**3*, as $p/n \to \kappa$ for some $\kappa \in (0, 1)$,*

$$\max_{j \in J_n} d_{\text{TV}}\left(\mathcal{L}\left(\frac{\hat{\beta}_j - \mathbb{E}\hat{\beta}_j}{\sqrt{\text{Var}(\hat{\beta}_j)}}\right), N(0,1)\right) = o(1).$$

It can be shown that $\tilde{\lambda}_+ \leq \lambda_+$ and $\tilde{\lambda}_- \geq \lambda_-$ and hence the assumption **A**3* is weaker than **A**3. It is worth pointing out that the assumption **A**3* even holds when $X_{J_n}^c$ does not have full column rank, in which case $\beta^*_{J_n}$ is still identifiable and $\hat{\beta}_{J_n}$ is still well-defined, although $\beta^*_{J_n^c}$ and $\hat{\beta}_{J_n^c}$ are not; see Appendix A.3.2 for details.

2.3.3 Examples

Throughout this subsection (except subsubsection 2.3.3), we consider the case where X is a realization of a random matrix, denoted by Z (to be distinguished from X). We will verify that the assumptions **A**3-**A**5 are satisfied with high probability under different regularity conditions on the distribution of Z. This is a standard way to justify the conditions for fixed design (Portnoy 1984, 1985) in the literature on regression M-estimates.

Random Design with Independent Entries

First we consider a random matrix Z with i.i.d. sub-gaussian entries.

Proposition 2.3.4. *Suppose Z has i.i.d. mean-zero σ^2-sub-gaussian entries with $\mathrm{Var}(Z_{ij}) = \tau^2 > 0$ for some $\sigma = O(\mathrm{polyLog}(n))$ and $\tau = \Omega\left(\frac{1}{\mathrm{polyLog}(n)}\right)$, then, when X is a realization of Z, assumptions **A**3-**A**5 for X are satisfied with high probability over Z for $J_n = \{1, \ldots, p\}$.*

In practice, the assumption of identical distribution might be invalid. In fact the assumptions **A**4, **A**5 and the first part of **A**3 ($\lambda_+ = O(\mathrm{polyLog}(n))$) are still satisfied with high probability if we only assume the independence between entries and boundedness of certain moments. To control λ_-, we rely on Litvak et al. (2005) which assumes symmetry of each entry. We obtain the following result based on it.

Proposition 2.3.5. *Suppose Z has independent σ^2-sub-gaussian entries with*

$$Z_{ij} \overset{d}{=} -Z_{ij}, \quad \mathrm{Var}(Z_{ij}) > \tau^2$$

*for some $\sigma = O(\mathrm{polyLog}(n))$ and $\tau = \Omega\left(\frac{1}{\mathrm{polyLog}(n)}\right)$, then, when X is a realization of Z, assumptions **A**3-**A**5 for X are satisfied with high probability over Z for $J_n = \{1, \ldots, p\}$.*

Under the conditions of Proposition 2.3.5, we can add an intercept term into the design matrix. Adding an intercept allows us to remove the mean-zero assumption for Z_{ij}'s. In fact, suppose Z_{ij} is symmetric with respect to μ_j, which is potentially non-zero, for all i, then according to section 2.3.2, we can replace Z_{ij} by $Z_{ij} - \mu_j$ and Proposition 2.3.6 can be then applied.

Proposition 2.3.6. *Suppose $Z = (\mathbf{1}, \tilde{Z})$ and $\tilde{Z} \in \mathbb{R}^{n\times(p-1)}$ has independent σ^2-sub-gaussian entries with*

$$\tilde{Z}_{ij} - \mu_j \overset{d}{=} \mu_j - \tilde{Z}_{ij}, \quad \mathrm{Var}(\tilde{Z}_{ij}) > \tau^2$$

*for some $\sigma = O(\mathrm{polyLog}(n))$, $\tau = \Omega\left(\frac{1}{\mathrm{polyLog}(n)}\right)$ and arbitrary μ_j. Then, when X is a realization of Z, assumptions **A**3*, **A**4 and **A**5 for X are satisfied with high probability over Z for $J_n = \{2, \ldots, p\}$.*

Dependent Gaussian Design

To show that our assumptions handle a variety of situations, we now assume that the observations, namely the rows of Z, are i.i.d. random vectors with a covariance matrix Σ. In particular we show that the Gaussian design, i.e. $z_i \overset{i.i.d.}{\sim} N(0, \Sigma)$, satisfies the assumptions with high probability.

Proposition 2.3.7. *Suppose $z_i \overset{i.i.d.}{\sim} N(0, \Sigma)$ with $\lambda_{\max}(\Sigma) = O(\mathrm{polyLog}(n))$ and $\lambda_{\min}(\Sigma) = \Omega\left(\frac{1}{\mathrm{polyLog}(n)}\right)$, then, when X is a realization of Z, assumptions **A**3-**A**5 for X are satisfied with high probability over Z for $J_n = \{1, \ldots, p\}$.*

This result extends to the matrix-normal design (e.g. Muirhead 1982, Chapter 3), i.e. $(Z_{ij})_{i\leq n, j\leq p}$ is one realization of a np-dimensional random variable Z with multivariate gaussian distribution

$$\text{vec}(Z) \triangleq (z_1^T, z_2^T, \ldots, z_n^T) \sim N(0, \Lambda \otimes \Sigma),$$

and $\otimes$ is the Kronecker product. It turns out that assumptions $\mathbf{A}3 - \mathbf{A}5$ are satisfied if both Λ and Σ are well-behaved.

Proposition 2.3.8. *Suppose Z is matrix-normal with* $\text{vec}(Z) \sim N(0, \Lambda \otimes \Sigma)$ *and*

$$\lambda_{\max}(\Lambda), \lambda_{\max}(\Sigma) = O\left(\text{polyLog}(n)\right), \quad \lambda_{\min}(\Lambda), \lambda_{\min}(\Sigma) = \Omega\left(\frac{1}{\text{polyLog}(n)}\right).$$

Then, when X is a realization of Z,assumptions $\boldsymbol{A}3$-$\boldsymbol{A}5$ for X are satisfied with high probability over Z for $J_n = \{1, \ldots, p\}$.

In order to incorporate an intercept term, we need slightly more stringent condition on Λ. Instead of assumption $\mathbf{A}3$, we prove that assumption $\mathbf{A}3^*$ - see subsubsection 2.3.2 - holds with high probability.

Proposition 2.3.9. *Suppose Z contains an intercept term, i.e. $Z = (\boldsymbol{1}, \tilde{Z})$ and $\tilde{Z}$ satisfies the conditions of Proposition 2.3.8. Further assume that*

$$\frac{\max_i |(\Lambda^{-\frac{1}{2}}\boldsymbol{1})_i|}{\min_i |(\Lambda^{-\frac{1}{2}}\boldsymbol{1})_i|} = O\left(\text{polyLog}(n)\right). \tag{2.7}$$

Then, when X is a realization of Z, assumptions $\boldsymbol{A}3^*$, $\boldsymbol{A}4$ and $\boldsymbol{A}5$ for X are satisfied with high probability over Z for $J_n = \{2, \ldots, p\}$.

When $\Lambda = I$, the condition (2.7) is satisfied. Another non-trivial example is the exchangeable case where Λ_{ij} are all equal for $i \neq j$. In this case, $\mathbf{1}$ is an eigenvector of Λ and hence it is also an eigenvector of $\Lambda^{-\frac{1}{2}}$. Thus $\Lambda^{-\frac{1}{2}}\mathbf{1}$ is a multiple of $\mathbf{1}$ and the condition (2.7) is satisfied.

Elliptical Design

Furthermore, we can move from Gaussian-like structure to generalized elliptical models where $z_i = \zeta_i \Sigma^{1/2} \mathcal{Z}_i$ where $\{\zeta_i, \mathcal{Z}_{ij} : i = 1, \ldots, n; j = 1, \ldots, p\}$ are independent random variables, $\mathcal{Z}_{ij}$ having for instance mean 0 and variance 1. The elliptical family is quite flexible in modeling data. It represents a type of data formed by a common driven factor and independent individual effects. It is widely used in multivariate statistics (Anderson 1962; Tyler 1987) and various fields, including finance (Cizek et al. 2005) and biology (Posekany et al. 2011). In the context of high-dimensional statistics, this class of model was used to refute universality claims in random matrix theory (El Karoui 2009). In robust regression, El Karoui et al. (2011) used elliptical models to show that the limit of $\|\hat{\beta}\|_2^2$ depends on the distribution of ζ_i

and hence the geometry of the predictors. As such, studies limited to Gaussian-like design were shown to be of very limited statistical interest. See also the deep classical inadmissibility results (Baranchik 1973; Jurečkovà and Klebanov 1997). However, as we will show in the next proposition, the common factors ζ_i do not distort the shape of the asymptotic distribution. A similar phenomenon happens in the random design case - see El Karoui et al. (2013) and Bean et al. (2013).

Proposition 2.3.10. *Suppose Z is generated from an elliptical model, i.e.*

$$Z_{ij} = \zeta_i \mathcal{Z}_{ij},$$

*where ζ_i are independent random variables taking values in $[a, b]$ for some $0 < a < b < \infty$ and $\mathcal{Z}_{ij}$ are independent random variables satisfying the conditions of Proposition 2.3.4 or Proposition 2.3.5. Further assume that $\{\zeta_i : i = 1, \ldots, n\}$ and $\{\mathcal{Z}_{ij} : i = 1, \ldots, n; j = 1, \ldots, p\}$ are independent. Then, when X is a realization of Z, assumptions **A**3-**A**5 for X are satisfied with high probability over Z for $J_n = \{1, \ldots, p\}$.*

Thanks to the fact that ζ_i is bounded away from 0 and ∞, the proof of Proposition 2.3.10 is straightforward, as shown in Appendix A.3. However, by a more refined argument and assuming identical distributions ζ_i, we can relax this condition.

Proposition 2.3.11. *Under the conditions of Proposition 2.3.10 (except the boundedness of ζ_i) and assume ζ_i are i.i.d. samples generated from some distribution F, independent of n, with*

$$P\left(\zeta_1 \geq t\right) \leq c_1 e^{-c_2 t^\alpha},$$

*for some fixed $c_1, c_2, \alpha > 0$ and $F^{-1}(q) > 0$ for any $q \in (0, 1)$ where F^{-1} is the quantile function of F and is continuous. Then, when X is a realization of Z, assumptions **A**3-**A**5 for X are satisfied with high probability over Z for $J_n = \{1, \ldots, p\}$.*

A counterexample

Consider a one-way ANOVA situation. In other words, let the design matrix have exactly 1 non-zero entry per row, whose value is 1. Let $\{k_i\}_{i=1}^n$ be integers in $\{1, \ldots, p\}$. And let $X_{i,j} = 1(j = k_i)$. Furthermore, let us constrain $n_j = |\{i : k_i = j\}|$ to be such that $1 \leq n_j \leq 2\lfloor p/n \rfloor$. Taking for instance $k_i = (i \mod p)$ is an easy way to produce such a matrix. The associated statistical model is just $y_i = \epsilon_i + \beta^*_{k_i}$.

It is easy to see that

$$\hat{\beta}_j = \arg\min_{\beta \in \mathbb{R}} \sum_{i:k_i=j} \rho(y_i - \beta_j) = \arg\min_{\beta \in \mathbb{R}} \sum_{i:k_i=j} \rho(\epsilon_i - (\beta_j - \beta^*_j)) \; .$$

This is of course a standard location problem. In the moderate-dimensional setting we consider, n_j remains finite as $n \to \infty$. So $\hat{\beta}_j$ is a non-linear function of finitely many random variables and will in general not be normally distributed.

For concreteness, one can take $\rho(x) = |x|$, in which case $\hat{\beta}_j$ is a median of $\{y_i\}_{\{i:k_i=j\}}$. The cdf of $\hat{\beta}_j$ is known exactly by elementary order statistics computations (see David and Nagaraja (1981)) and is not that of a Gaussian random variable in general. In fact, the ANOVA design considered here violates the assumption **A**3 since $\lambda_- = \min_j n_j/n = O(1/n)$. Further, we can show that the assumption **A**5 is also violated, at least in the least-square case; see Section 2.5.1 for details.

2.3.4 Comments and discussions

Asymptotic Normality in High Dimensions

In the p-fixed regime, the asymptotic distribution is easily defined as the limit of $\mathcal{L}(\hat{\beta})$ in terms of weak topology (Van der Vaart 1998). However, in regimes where the dimension p grows, the notion of asymptotic distribution is more delicate. a conceptual question arises from the fact that the dimension of the estimator $\hat{\beta}$ changes with n and thus there is no well-defined distribution which can serve as the limit of $\mathcal{L}(\hat{\beta})$, where $\mathcal{L}(\cdot)$ denotes the law. One remedy is proposed by Mallows (1972). Under this framework, a triangular array $\{W_{n,j}, j = 1, 2, \ldots, p_n\}$, with $\mathbb{E}W_{n,j} = 0, \mathbb{E}W_{n,j}^2 = 1$, is called jointly asymptotically normal if for any deterministic sequence $a_n \in \mathbb{R}^{p_n}$ with $\|a_n\|_2 = 1$,

$$\mathcal{L}\left(\sum_{j=1}^{p_n} a_{n,j} W_{n,j}\right) \to N(0,1).$$

When the zero mean and unit variance are not satisfied, it is easy to modify the definition by normalizing random variables.

Definition 2.3.12 (joint asymptotic normality)**.** $\{W_n : W_n \in \mathbb{R}^{p_n}\}$ *is jointly asymptotically normal if and only if for any sequence* $\{a_n : a_n \in \mathbb{R}^{p_n}\}$*,*

$$\mathcal{L}\left(\frac{a_n^T(W_n - \mathbb{E}W_n)}{\sqrt{a_n^T \operatorname{Cov}(W_n) a_n}}\right) \to N(0,1).$$

The above definition of asymptotic normality is strong and appealing but was shown not to hold for least-squares in the moderate p/n regime (Huber 1973a). In fact, Huber (1973a) shows that $\hat{\beta}^{LS}$ is jointly asymtotically normal only if

$$\max_i (X(X^TX)^{-1}X^T)_{i,i} \to 0.$$

When $p/n \to \kappa \in (0,1)$, provided X is full rank,

$$\max_i (X(X^TX)^{-1}X^T)_{i,i} \geq \frac{1}{n}\operatorname{tr}(X(X^TX)^{-1}X^T) = \frac{p}{n} \to \kappa > 0.$$

In other words, in moderate p/n regime, the asymptotic normality cannot hold for all linear contrasts, even in the case of least-squares.

In applications, however, it is usually not necessary to consider all linear contrasts but instead a small subset of them, e.g. all coordinates or low dimensional linear contrasts such as $\beta_1^* - \beta_2^*$. We can naturally modify Definition 2.3.12 and adapt to our needs by imposing constraints on a_n. A popular concept, which we use in Section 2.1 informally, is called coordinate-wise asymptotic normality and defined by restricting a_n to be the canonical basis vectors, which have only one non-zero element. An equivalent definition is stated as follows.

Definition 2.3.13 (coordinate-wise asymptotic normal)**.** $\{W_n : W_n \in \mathbb{R}^{p_n}\}$ *is coordinate-wise asymptotically normal if and only if for any sequence* $\{j_n : j_n \in \{1, \ldots, p_n\}\}$*,*

$$\mathcal{L}\left(\frac{W_{n,j_n} - \mathbb{E}W_{n,j_n}}{\sqrt{\mathrm{Var}(W_{n,j_n})}}\right) \to N(0,1).$$

A more convenient way to define the coordinate-wise asymptotic normality is to introduce a metric $d(\cdot,\cdot)$, e.g. Kolmogorov distance and total variation distance, which induces the weak convergence topology. Then W_n is coordinate-wise asymptotically normal if and only if

$$\max_j d\left(\mathcal{L}\left(\frac{W_{n,j} - \mathbb{E}W_{n,j}}{\sqrt{\mathrm{Var}(W_{n,j})}}\right), N(0,1)\right) = o(1).$$

Variance and bias estimation

To complete the inference, we need to compute the bias and variance. As discussed in Remark 2.3.2, the M-estimator is unbiased if the loss function and the error distribution are symmetric. For the variance, it is easy to get a conservative estimate via resampling methods such as Jackknife as a consequence of Efron-Stein's inequality; see El Karoui (2013) and El Karoui and Purdom (2015) for details. Moreover, by the variance decomposition formula,

$$\mathrm{Var}(\hat{\beta}_j) = \mathbb{E}\left[\mathrm{Var}(\hat{\beta}_j|X)\right] + \mathrm{Var}\left[\mathbb{E}(\hat{\beta}_j|X)\right] \geq \mathbb{E}\left[\mathrm{Var}(\hat{\beta}_j|X)\right],$$

the unconditional variance, when X is a random design matrix, is a conservative estimate. The unconditional variance can be calculated by solving a non-linear system; see El Karoui (2013) and Donoho and Montanari (2016).

However, estimating the exact variance is known to be hard. El Karoui and Purdom (2015) show that the existing resampling schemes, including jacknife, pairs-bootstrap, residual bootstrap, etc., are either too conservative or too anti-conservative when p/n is large. The challenge, as mentioned in El Karoui (2013) and El Karoui and Purdom (2015), is due to the fact that the residuals $\{R_i\}$ do not mimic the behavior of $\{\epsilon_i\}$ and that the resampling methods effectively modifies the geometry of the dataset from the point of view of the statistics of interest. We believe that variance estimation in moderate p/n regime should rely on different methodologies from the ones used in low-dimensional estimation.

2.4 Proof Sketch

Since the proof of Theorem 2.3.1 is somewhat technical, we illustrate the main idea in this section.

First notice that the M-estimator $\hat{\beta}$ is an implicit function of independent random variables $\epsilon_1, \ldots, \epsilon_n$, which is determined by

$$\frac{1}{n}\sum_{i=1}^{n} x_i \psi(\epsilon_i - x_i\hat{\beta}) = 0. \tag{2.8}$$

The Hessian matrix of the loss function in (2.5) is $\frac{1}{n}X^T DX \succeq D_0\lambda_- I_p$ under the notation introduced in section 2.3.1. The assumption **A**3 then implies that the loss function is strongly convex, in which case $\hat{\beta}$ is unique. Then $\hat{\beta}$ can be seen as a non-linear function of ϵ_i's. A powerful central limit theorem for this type of statistics is the second-order Poincaré inequality (SOPI), developed in Chatterjee (2009) and used there to re-prove central limit theorems for linear spectral statistics of large random matrices. We recall one of the main results for the convenience of the reader.

Proposition 2.4.1 (SOPI; Chatterjee (2009))**.** *Let $\mathscr{W} = (u_1(W_1), \ldots, u_n(W_n))$ where $W_i \overset{i.i.d.}{\sim} N(0,1)$ and $\|u_i'\|_\infty \le c_1, \|u_i''\|_\infty \le c_2$. Take any $g \in C^2(\mathbb{R}^n)$ and let $\nabla_i g$, ∇g and $\nabla^2 g$ denote the i-th partial derivative, gradient and Hessian of g. Let*

$$\kappa_0 = \left(\mathbb{E}\sum_{i=1}^{n}\left|\nabla_i g(\mathscr{W})\right|^4\right)^{\frac{1}{2}}, \quad \kappa_1 = (\mathbb{E}\|\nabla g(\mathscr{W})\|_2^4)^{\frac{1}{4}}, \quad \kappa_2 = (\mathbb{E}\|\nabla^2 g(\mathscr{W})\|_{\mathrm{op}}^4)^{\frac{1}{4}},$$

and $U = g(\mathscr{W})$. If U has finite fourth moment, then

$$d_{\mathrm{TV}}\left(\mathcal{L}\left(\frac{U-\mathbb{E}U}{\sqrt{\mathrm{Var}(U)}}\right), N(0,1)\right) \le \frac{2\sqrt{5}(c_1c_2\kappa_0 + c_1^3\kappa_1\kappa_2)}{\mathrm{Var}(U)}.$$

From (2.8), it is not hard to compute the gradient and Hessian of $\hat{\beta}_j$ with respect to ϵ. Recalling the definitions in Equation (2.6) on p. 17, we have

Lemma 2.4.2. *Suppose $\psi \in C^2(\mathbb{R}^n)$, then*

$$\frac{\partial\hat{\beta}_j}{\partial\epsilon^T} = e_j^T(X^TDX)^{-1}X^TD \tag{2.9}$$

$$\frac{\partial\hat{\beta}_j}{\partial\epsilon\partial\epsilon^T} = G^T\,\mathrm{diag}(e_j^T(X^TDX)^{-1}X^T\tilde{D})G \tag{2.10}$$

where e_j is the j-th cononical basis vectors in $\mathbb{R}^p$ and

$$G = I - X(X^TDX)^{-1}X^TD.$$

Recalling the definitions of K_i's in Assumption **A**1 on p. 19, we can bound κ_0, κ_1 and κ_2 as follows.

Lemma 2.4.3. *Let $\kappa_{0j}, \kappa_{1j}, \kappa_{2j}$ defined as in Proposition 2.4.1 by setting $\mathscr{W} = \epsilon$ and $g(\mathscr{W}) = \hat{\beta}_j$. Let*

$$M_j = \mathbb{E}\|e_j^T (X^T D X)^{-1} X^T D^{\frac{1}{2}}\|_\infty, \tag{2.11}$$

then

$$\kappa_{0j}^2 \le \frac{K_1^2}{(nK_0\lambda_-)^{\frac{3}{2}}} \cdot M_j, \quad \kappa_{1j}^4 \le \frac{K_1^2}{(nK_0\lambda_-)^2}, \quad \kappa_{2j}^4 \le \frac{K_2^4}{(nK_0\lambda_-)^{\frac{3}{2}}} \cdot \left(\frac{K_1}{K_0}\right)^4 \cdot M_j.$$

As a consequence of the second-order Poincaré inequality , we can bound the total variation distance between $\hat{\beta}_j$ and a normal distribution by M_j and $\mathrm{Var}(\hat{\beta}_j)$. More precisely, we prove the following Lemma.

Lemma 2.4.4. *Under assumptions **A**1-**A**3,*

$$\max_j d_{\mathrm{TV}} \left(\mathcal{L}\left(\frac{\hat{\beta}_j - \mathbb{E}\hat{\beta}_j}{\sqrt{\mathrm{Var}(\hat{\beta}_j)}} \right), N(0,1) \right) = O_p\left(\frac{\max_j (nM_j^2)^{\frac{1}{8}}}{n \cdot \min_j \mathrm{Var}(\hat{\beta}_j)} \cdot \mathrm{polyLog}(n) \right).$$

Lemma 2.4.4 is the key to prove Theorem 2.3.1. To obtain the coordinate-wise asymptotic normality, it is left to establish an upper bound for M_j and a lower bound for $\mathrm{Var}(\hat{\beta}_j)$. In fact, we can prove that

Lemma 2.4.5. *Under assumptions **A**1 - **A**5,*

$$\max_j M_j = O\left(\frac{\mathrm{polyLog}(n)}{n} \right), \quad \min_j \mathrm{Var}(\hat{\beta}_j) = \Omega\left(\frac{1}{n \cdot \mathrm{polyLog}(n)} \right).$$

Then Lemma 2.4.4 and Lemma 2.4.5 together imply that

$$\max_j d_{\mathrm{TV}} \left(\mathcal{L}\left(\frac{\hat{\beta}_j - \mathbb{E}\hat{\beta}_j}{\sqrt{\mathrm{Var}(\hat{\beta}_j)}} \right), N(0,1) \right) = O\left(\frac{\mathrm{polyLog}(n)}{n^{\frac{1}{8}}} \right) = o(1).$$

Appendix A.1, provides a roadmap of the proof of Lemma 2.4.5 under a special case where the design matrix X is one realization of a random matrix with i.i.d. sub-gaussian entries. It also serves as an outline of the rigorous proof in Appendix A.2.

2.4.1 Comment on the Second-Order Poincaré inequality

Notice that when g is a linear function such that $g(z) = \sum_{i=1}^n a_i z_i$, then the Berry-Esseen inequality (Esseen 1945a) implies that

$$d_K\left(\mathcal{L}\left(\frac{W-\mathbb{E}W}{\sqrt{\mathrm{Var}(W)}}\right), N(0,1)\right) \preceq \frac{\sum_{i=1}^n |a_i|^3}{(\sum_{i=1}^n a_i^2)^{\frac{3}{2}}},$$

where

$$d_K(F,G) = \sup_x |F(x) - G(x)|.$$

On the other hand, the second-order Poincaré inequality implies that

$$d_K\left(\mathcal{L}\left(\frac{W-\mathbb{E}W}{\sqrt{\mathrm{Var}(W)}}\right), N(0,1)\right) \leq d_{\mathrm{TV}}\left(\mathcal{L}\left(\frac{W-\mathbb{E}W}{\sqrt{\mathrm{Var}(W)}}\right), N(0,1)\right) \preceq \frac{(\sum_{i=1}^n a_i^4)^{\frac{1}{2}}}{\sum_{i=1}^n a_i^2}\ .$$

This is slightly worse than the Berry-Esseen bound and requires stronger conditions on the distributions of variates but provides bounds for TV metric instead of Kolmogorov metric. This comparison shows that second-order Poincaré inequality can be regarded as a generalization of the Berry-Esseen bound for non-linear transformations of independent random variables.

2.5 Least-Squares Estimator

The Least-Squares Estimator is a special case of an M-estimator with $\rho(x) = \frac{1}{2}x^2$. Because the estimator can then be written explicitly, the analysis of its properties is extremely simple and it has been understood for several decades (see arguments in e.g. Huber (1973a)[Lemma 2.1] and Huber (1981)[Proposition 2.2]). In this case, the hat matrix $H = X(X^TX)^{-1}X^T$ captures all the problems associated with dimensionality in the problem. In particular, proving the asymptotic normality simply requires an application of the Lindeberg-Feller theorem.

It is however somewhat helpful to compare the conditions required for asymptotic normality in this simple case and the ones we required in the more general setup of Theorem 2.3.1. We do so briefly in this section.

2.5.1 Coordinate-Wise Asymptotic Normality of LSE

Under the linear model (2.1), when X is full rank,

$$\hat{\beta}^{LS} = \beta^* + (X^TX)^{-1}X^T\epsilon,$$

thus each coordinate of $\hat{\beta}^{LS}$ is a linear contrast of ϵ with zero mean. Instead of assumption **A**2, which requires ϵ_i to be sub-gaussian, we only need to assume $\max_i \mathbb{E}|\epsilon_i|^3 < \infty$, under

which the Berry-Essen bound for non-i.i.d. data (Esseen 1945a) implies that

$$d_K\left(\mathcal{L}\left(\frac{\hat{\beta}_j-\beta_j^*}{\sqrt{\mathrm{Var}(\hat{\beta}_j)}}\right),N(0,1)\right)\preceq\frac{\|e_j(X^TX)^{-1}X^T\|_3^3}{\|e_j^T(X^TX)^{-1}X^T\|_2^3}\leq\frac{\|e_j(X^TX)^{-1}X^T\|_\infty}{\|e_j(X^TX)^{-1}X^T\|_2}.$$

This motivates us to define a matrix specific quantity $S_j(X)$ such that

$$S_j(X)=\frac{\|e_j^T(X^TX)^{-1}X^T\|_\infty}{\|e_j^T(X^TX)^{-1}X^T\|_2} \tag{2.12}$$

then the Berry-Esseen bound implies that $\max_{j\in J_n}S_j(X)$ determines the coordinate-wise asymptotic normality of $\hat{\beta}^{LS}$.

Theorem 2.5.1. *If* $\max_i\mathbb{E}|\epsilon_i|^3<\infty$*, then*

$$\max_{j\in J_n}d_K\left(\frac{\hat{\beta}_{LS,j}-\beta_{0,j}}{\sqrt{\mathrm{Var}(\hat{\beta}_{LS,j})}},N(0,1)\right)\leq A\cdot\max_i\frac{\mathbb{E}|\epsilon_i|^3}{(\mathbb{E}\epsilon_i^2)^{\frac{3}{2}}}\cdot\max_{j\in J_n}S_j(X),$$

where A is an absolute constant and $d_K(\cdot,\cdot)$ is the Kolmogorov distance, defined as

$$d_K(F,G)=\sup_x|F(x)-G(x)|.$$

It turns out that $\max_{j\in J_n}S_j(X)$ plays in the least-squares setting the role of Δ_C in assumption **A**5. Since it has been known that a condition like $S_j(X)\to 0$ is necessary for asymptotic normality of least-square estimators (Huber 1973a, Proposition 2.2), this shows in particular that our Assumption **A**5, or a variant, is also needed in the general case. See Appendix A.3.4 for details.

2.5.2 Discussion

Naturally, checking the conditions for asymptotic normality is much easier in the least-squares case than in the general case under consideration in this chapter. In particular:

1. Asymptotic normality conditions can be checked for a broader class of random design matrices. See Appendix A.3.4 for details.
2. For orthogonal design matrices, i.e $X^TX=cI$ for some $c>0$, $S_j(X)=\frac{\|X_j\|_\infty}{\|X_j\|_2}$. Hence, the condition $S_j(X)=o(1)$ is true if and only if no entry dominates the $j-th$ row of X.
3. The ANOVA-type counterexample we gave in Section 2.3.3 still provides a counter-example. The reason now is different: namely the sum of finitely many independent random variables is evidently in general non-Gaussian. In fact, in this case, $S_j(X)=\frac{1}{\sqrt{n_j}}$ is bounded away from 0.

Inferential questions are also extremely simple in this context and essentially again dimension independent for the reasons highlighted above. Theorem 2.5.1 naturally reads,

$$\frac{\hat{\beta}_j - \beta_j^*}{\sigma\sqrt{e_j^T (X^T X)^{-1} e_j}} \xrightarrow{d} N(0,1). \tag{2.13}$$

Estimating σ is still simple under minimal conditions provided $n - p \to \infty$: see Bickel and Freedman (1983b, Theorem 1.3) or standard computations concerning the normalized residual sum-of-squares (using variance computations for the latter may require up to 4 moments for ϵ_i's). Then we can replace σ in (2.13) by $\hat{\sigma}$ with

$$\hat{\sigma}^2 = \frac{1}{n-p} \sum_{k=1}^{n} R_k^2$$

where $R_k = y_k - x_k^T \hat{\beta}$ and construct confidence intervals for β_j^* based on $\hat{\sigma}$. If $n - p$ does not tend to ∞, the normalized residual sum of squares is evidently not consistent even in the case of Gaussian errors, so this requirement may not be dispensed of.

2.6 Numerical Results

As seen in the previous sections and related papers, there are five important factors that affect the distribution of $\hat{\beta}$: the design matrix X, the error distribution $\mathcal{L}(\epsilon)$, the sample size n, the ratio κ, and the loss function ρ. The aim of this section is to assess the quality of the agreement between the asymptotic theoretical results of Theorem 2.3.1 and the empirical, finite-dimensional properties of $\hat{\beta}(\rho)$. We also perform a few simulations where some of the assumptions of Theorem 2.3.1 are violated to get an intuitive sense of whether those assumptions appear necessary or whether they are simply technical artifacts associated with the method of proof we developed. As such, the numerical experiments we report on in this section can be seen as a complement to Theorem 2.3.1 rather than only a simple check of its practical relevance.

The design matrices we consider are one realization of random design matrices of the following three types:

(i.i.d. design) : $X_{ij} \overset{i.i.d.}{\sim} F$;

(elliptical design) : $X_{ij} = \zeta_i \tilde{X}_{ij}$, where $\tilde{X}_{ij} \overset{i.i.d.}{\sim} N(0,1)$ and $\zeta_i \overset{i.i.d.}{\sim} F$. In addition, $\{\zeta_i\}$ is independent of $\{\tilde{X}_{ij}\}$;

(partial Hadamard design) : a matrix formed by a random set of p columns of a $n \times n$ Hadamard matrix, i.e. a $n \times n$ matrix whose columns are orthogonal with entries restricted to ± 1.

Here we consider two candidates for F in i.i.d. design and elliptical design: standard normal distribution $N(0,1)$ and t-distribution with two degrees of freedom (denoted t_2). For the error distribution, we assume that ϵ has i.i.d. entries with one of the above two distributions, namely $N(0,1)$ and t_2. The t-distribution violates our assumption **A**2.

To evaluate the finite sample performance, we consider $n \in \{100, 200, 400, 800\}$ and $\kappa \in \{0.5, 0.8\}$. In this section we will consider a Huber loss with $k = 1.345$ (Huber 1981), i.e.

$$\rho(x) = \begin{cases} \frac{1}{2}x^2 & |x| \le k \\ kx - \frac{k^2}{2} & |x| > k \end{cases}$$

$k = 1.345$ is the default in `R` and yields 95% relative efficiency for Gaussian errors in low-dimensional problems. We also carried out the numerical work for L_1-regression, i.e. $\rho(x) = |x|$. See Appendix A.4 for details.

2.6.1 Asymptotic Normality of A Single Coordinate

First we simulate the finite sample distribution of $\hat{\beta}_1$, the first coordinate of $\hat{\beta}$. For each combination of sample size n (100, 200, 400 and 800), type of design (i.i.d, elliptical and Hadamard), entry distribution F (normal and t_2) and error distribution $\mathcal{L}(\epsilon)$ (normal and t_2), we run 50 simulations with each consisting of the following steps:

(Step 1) Generate one design matrix X;

(Step 2) Generate 300 error vectors ϵ;

(Step 3) Regress each $Y = \epsilon$ on the design matrix X and end up with 300 random samples of $\hat{\beta}_1$, denoted by $\hat{\beta}_1^{(1)}, \ldots, \hat{\beta}_1^{(300)}$;

(Step 4) Estimate the standard deviation of $\hat{\beta}_1$ by the sample standard error $\widehat{\mathrm{sd}}$;

(Step 5) Construct a confidence interval $\mathcal{I}^{(k)} = \left[\hat{\beta}_1^{(k)} - 1.96 \cdot \widehat{\mathrm{sd}}, \hat{\beta}_1^{(k)} + 1.96 \cdot \widehat{\mathrm{sd}}\right]$ for each $k = 1, \ldots, 300$;

(Step 6) Calculate the empirical 95% coverage by the proportion of confidence intervals which cover the true $\beta_1 = 0$.

Finally, we display the boxplots of the empirical 95% coverages of $\hat{\beta}_1$ for each case in Figure 2.2. It is worth mentioning that our theories cover two cases: 1) i.i.d design with normal entries and normal errors (orange bars in the first row and the first column), see Proposition 2.3.4; 2) elliptical design with normal factors ζ_i and normal errors (orange bars in the second row and the first column), see Proposition 2.3.10.

We first discuss the case $\kappa = 0.5$. In this case, there are only two samples per parameter. Nonetheless, we observe that the coverage is quite close to 0.95, even with a sample size as small as 100, in both cases that are covered by our theories. For other cases, it is interesting

to see that the coverage is valid and most stable in the partial hadamard design case and is not sensitive to the distribution of multiplicative factor in elliptical design case even when the error has a t_2 distribution. For i.i.d. designs, the coverage is still valid and stable when the entry is normal. By contrast, when the entry has a t_2 distribution, the coverage has a large variation in small samples. The average coverage is still close to 0.95 in the i.i.d. normal design case but is slightly lower than 0.95 in the i.i.d. t_2 design case. In summary, the finite sample distribution of $\hat{\beta}_1$ is more sensitive to the entry distribution than the error distribution. This indicates that the assumptions on the design matrix are not just artifacts of the proof but are quite essential.

The same conclusion can be drawn from the case where $\kappa = 0.8$ except that the variation becomes larger in most cases when the sample size is small. However, it is worth pointing out that even in this case where there is 1.25 samples per parameter, the sample distribution of $\hat{\beta}_1$ is well approximated by a normal distribution with a moderate sample size ($n \geq 400$). This is in contrast to the classical rule of thumb which suggests that 5-10 samples are needed per parameter.

2.6.2 Asymptotic Normality for Multiple Marginals

Since our theory holds for general J_n, it is worth checking the approximation for multiple coordinates in finite samples. For illustration, we consider 10 coordinates, namely $\hat{\beta}_1 \sim \hat{\beta}_{10}$, simultaneously and calculate the minimum empirical 95% coverage. To avoid the finite sample dependence between coordinates involved in the simulation, we estimate the empirical coverage independently for each coordinate. Specifically, we run 50 simulations with each consisting of the following steps:

(Step 1) Generate one design matrix X;

(Step 2) Generate 3000 error vectors ϵ;

(Step 3) Regress each $Y = \epsilon$ on the design matrix X and end up with 300 random samples of $\hat{\beta}_j$ for each $j = 1, \ldots, 10$ by using the $(300(j-1)+1)$-th to $300j$-th response vector Y;

(Step 4) Estimate the standard deviation of $\hat{\beta}_j$ by the sample standard error $\widehat{\mathrm{sd}}_j$ for $j = 1, \ldots, 10$;

(Step 5) Construct a confidence interval $\mathcal{I}_j^{(k)} = \left[\hat{\beta}_j^{(k)} - 1.96 \cdot \widehat{\mathrm{sd}}_j, \hat{\beta}_j^{(k)} + 1.96 \cdot \widehat{\mathrm{sd}}_j\right]$ for each $j = 1, \ldots, 10$ and $k = 1, \ldots, 300$;

(Step 6) Calculate the empirical 95% coverage by the proportion of confidence intervals which cover the true $\beta_j = 0$, denoted by C_j, for each $j = 1, \ldots, 10$,

(Step 7) Report the minimum coverage $\min_{1 \leq j \leq 10} C_j$.

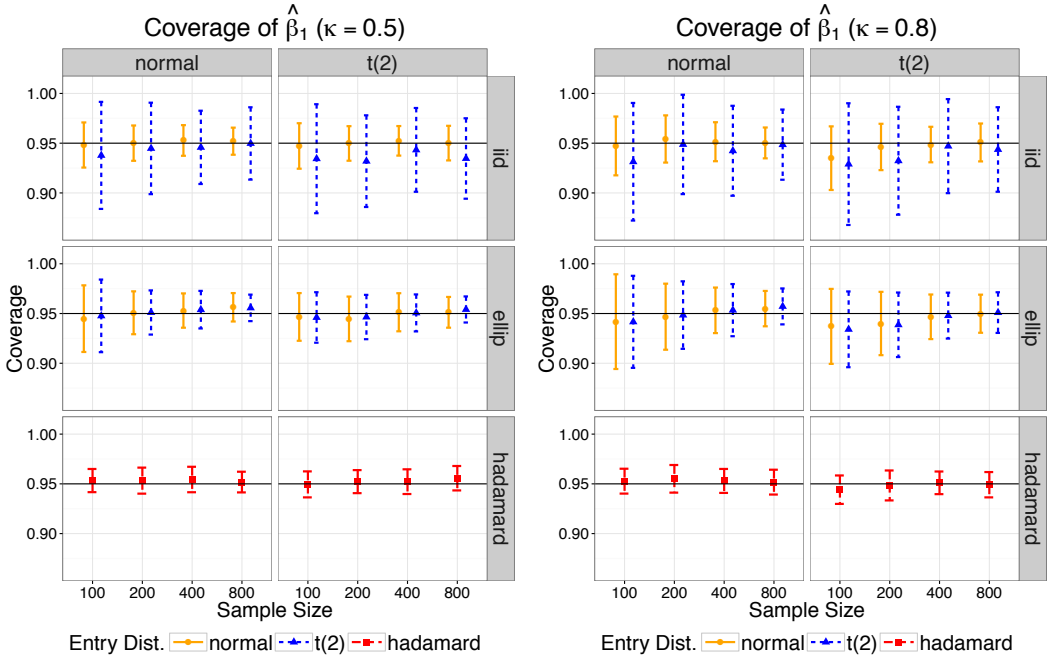


Figure 2.2: Empirical 95% coverage of $\hat{\beta}_1$ with $\kappa = 0.5$ (left) and $\kappa = 0.8$ (right) using $\text{Huber}_{1.345}$ loss. The x-axis corresponds to the sample size, ranging from 100 to 800; the y-axis corresponds to the empirical 95% coverage. Each column represents an error distribution and each row represents a type of design. The orange solid bar corresponds to the case F = Normal; the blue dotted bar corresponds to the case $F = \text{t}_2$; the red dashed bar represents the Hadamard design.

If the assumptions **A**1 - **A**5 are satisfied, $\min_{1\le j\le 10} C_j$ should also be close to 0.95 as a result of Theorem 2.3.1. Thus, $\min_{1\le j\le 10} C_j$ is a measure for the approximation accuracy for multiple marginals. Figure 2.3 displays the boxplots of this quantity under the same scenarios as the last subsection. In two cases that our theories cover, the minimum coverage is increasingly closer to the true level 0.95. Similar to the last subsection, the approximation is accurate in the partial hadamard design case and is insensitive to the distribution of multiplicative factors in the elliptical design case. However, the approximation is very inaccurate in the i.i.d. t_2 design case. Again, this shows the evidence that our technical assumptions are not artifacts of the proof.

On the other hand, the figure 2.3 suggests using a conservative variance estimator, e.g. the Jackknife estimator, or corrections on the confidence level in order to make simultaneous inference on multiple coordinates. Here we investigate the validity of Bonferroni correction by modifying the step 5 and step 6. The confidence interval after Bonferroni correction is

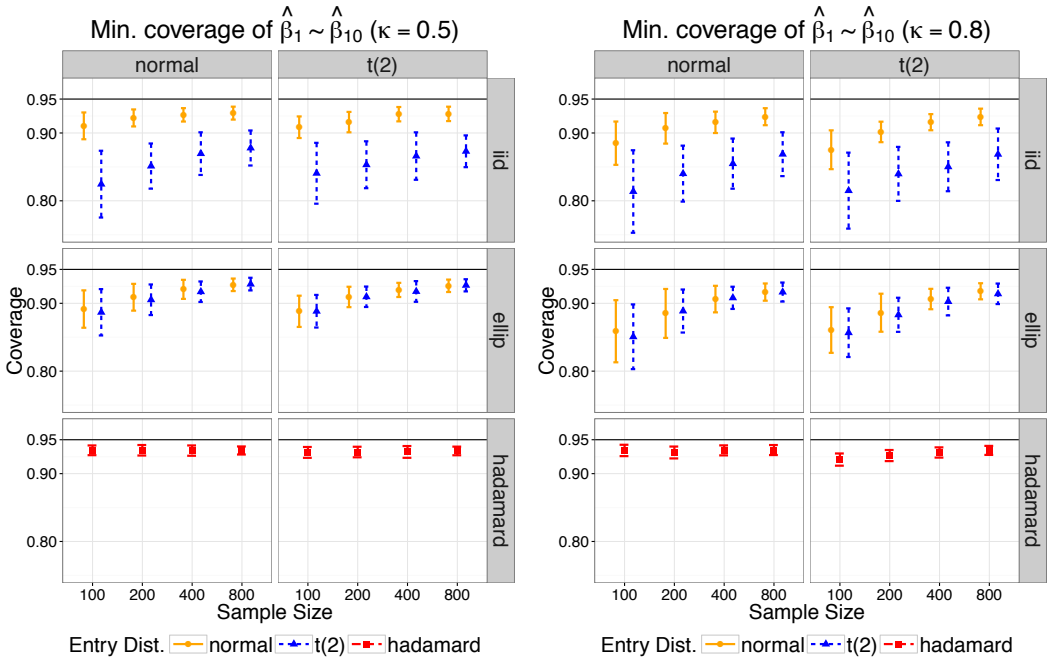


Figure 2.3: Mininum empirical 95% coverage of $\hat{\beta}_1 \sim \hat{\beta}_{10}$ with $\kappa = 0.5$ (left) and $\kappa = 0.8$ (right) using $\text{Huber}_{1.345}$ loss. The x-axis corresponds to the sample size, ranging from 100 to 800; the y-axis corresponds to the minimum empirical 95% coverage. Each column represents an error distribution and each row represents a type of design. The orange solid bar corresponds to the case $F = \text{Normal}$; the blue dotted bar corresponds to the case $F = \text{t}_2$; the red dashed bar represents the Hadamard design.

obtained by

$$\mathcal{I}_j^{(k)} = \left[\hat{\beta}_j^{(k)} - z_{1-\alpha/20} \cdot \widehat{\text{sd}}_j, \hat{\beta}_j^{(k)} + z_{1-\alpha/20} \cdot \widehat{\text{sd}}_j\right] \tag{2.14}$$

where $\alpha = 0.05$ and z_γ is the γ-th quantile of a standard normal distribution. The proportion of k such that $0 \in \mathcal{I}_j^{(k)}$ for all $j \leq 10$ should be at least 0.95 if the marginals are all close to a normal distribution. We modify the confidence intervals in step 5 by (2.14) and calculate the proportion of k such that $0 \in \mathcal{I}_j^{(k)}$ for all j in step 6. Figure 2.4 displays the boxplots of this coverage. It is clear that the Bonferroni correction gives the valid coverage except when $n = 100, \kappa = 0.8$ and the error has a t_2 distribution.

2.7 Conclusion

We have proved coordinate-wise asymptotic normality for regression M-estimates in the moderate-dimensional asymptotic regime $p/n \to \kappa \in (0, 1)$, for fixed design matrices under

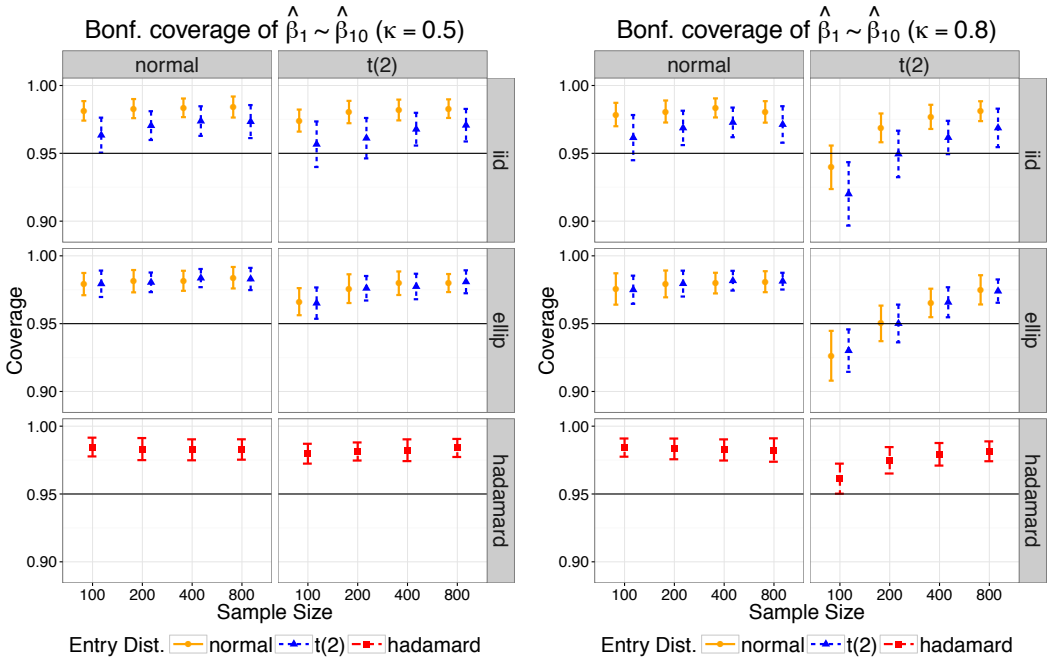


Figure 2.4: Empirical 95% coverage of $\hat{\beta}_1 \sim \hat{\beta}_{10}$ after Bonferroni correction with $\kappa = 0.5$ (left) and $\kappa = 0.8$ (right) using $\text{Huber}_{1.345}$ loss. The x-axis corresponds to the sample size, ranging from 100 to 800; the y-axis corresponds to the empirical uniform 95% coverage after Bonferroni correction. Each column represents an error distribution and each row represents a type of design. The orange solid bar corresponds to the case $F = \text{Normal}$; the blue dotted bar corresponds to the case $F = \text{t}_2$; the red dashed bar represents the Hadamard design.

appropriate technical assumptions. Our design assumptions are satisfied with high probability for a broad class of random designs. The main novel ingredient of the proof is the use of the second-order Poincaré inequality. Numerical experiments confirm and complement our theoretical results.

Acknowledgment

Peter J. Bickel and Lihua Lei were supported by the NSF DMS-1160319 and NSF DMS-1713083. Noureddine El Karoui was supported by the NSF DMS-1510172. The authors thank anonymous reviewers for helpful discussions and suggestions.

I thank Professor Noureddine El Karoui and Professor Peter J. Bickel for their excellent supervision on this work. The paper was published on Probability Theory and Related Fields on December, 2018 (Lei et al. 2018). The idea was originated from Noureddine El Karoui and Peter Bickel as an extension of their earlier works (El Karoui et al. 2011; El Karoui 2013;

Bean et al. 2013; El Karoui 2015, 2018). Noureddine El Karoui and Peter Bickel provided joint advising on this work, with joint meetings of the three of us weekly over the course of two years or so.

Chapter 3

Exact Inference for Linear Models

3.1 Introduction

In this article, we consider the following fixed-design linear model

$$y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i, \qquad i = 1, \ldots, n, \tag{3.1}$$

where ϵ_i's are stochastic errors and x_{ij}'s are treated as fixed quantities. Throughout we will use the following compact notation

$$y = \beta_0 \mathbf{1} + X\beta + \epsilon, \tag{3.2}$$

where $y = (y_i)$ denote the response vector, $X = (x_{ij}) \in \mathbb{R}^{n\times p}$ denote the design matrix, $\epsilon = (\epsilon)$ denote the error terms and $\mathbf{1} \in \mathbb{R}^n$ denote the vector with all entries equal to one. Two driving forces in early history of statistics – location problems and analysis of variance (ANOVA) problems – are both special cases of linear models.

Our focus is on testing a general linear hypothesis:

$$H_0 : R^T\beta = 0, \quad \text{where } R \in \mathbb{R}^{p\times r} \text{ is a fixed matrix with rank } r. \tag{3.3}$$

Testing linear hypotheses in linear models is ubiquitous and fundamental in numerous areas. One important example is to test whether a particular coefficient is zero, i.e. $H_0 : \beta_1 = 0$, a special case where $R = (1, 0, \ldots, 0)^T \in \mathbb{R}^{p\times 1}$. Another important example is to test the global null, i.e. $H_0 : \beta = 0$, equivalent to the linear hypothesis with $R = I_{p\times p}$. We refer to Chapter 7 of Lehmann and Romano (2006) for an extensive discussion of other examples. By inverting a test with (asymptotically) valid type-I error control, we can obtain a confidence interval/region for $R\beta$. This is of particular interest when $r = 1$, which corresponds to a single linear contrast of the regression coefficient.

This is one of the most fundamental and long-lasting problem in statistics as well as a convenient powerful prototype to generate methodology that works for more complicated

statistical problems. In the past century, several categories of methodology were proposed: normal theory based tests (Fisher 1922; Fisher 1924), permutation tests (Pitman 1937b; Pitman 1938), rank-based tests (Friedman 1937), tests based on regression R-estimates (Hájek 1962), M-estimates (Huber 1973b) and L-estimates (Bickel 1973), resampling based tests (Freedman 1981) and other tests (e.g. median-based tests (Theil 1950a; Brown and Mood 1951), symmetry-based tests (Hartigan 1970) and non-standard tests (Meinshausen 2015)). we only give the earliest reference we can track for each category to highlight the chronicle of methodology development. We will provide an extensive literature review in Section 3.4.

For a given confidence level $1-\alpha$, a test is exact if the type-I error is exactly α, in finite samples without any asymptotics. Exact tests are intellectually and practical appealling because they provide strong error control without requirement of large sample or artificial asymptotic regimes. However, perhaps surprisingly, there is no test that is exact under reasonably general assumptions to the best of our knowledge. Below is a brief summary of the conditions under which the existing tests are exact.

- Regression t-tests and F-tests are exact with **normal errors**;
- Permutation tests are exact for **global null** or **two-way layouts** (e.g. Brown and Maritz 1982);
- Rank-based tests are exact for **ANOVA problems**;
- Tests based on regression R/M/L-estimates can be exact for **global null**;
- Hartigan (1970)'s test is exact for **certain forms of balanced ANOVA problems with symmetric errors and** $r = 1$;
- Meinshausen (2015)'s test is exact for **rotationally invariant errors with known noise level**. Note that if ϵ_i's are i.i.d., rotation invariance implies the normality of ϵ_i's (Maxwell 1860);
- Other tests are exact either for global null or under unrealistically restrictive assumptions or with infeasible computation.

In this article, we develop an **exact test**, referred to as *cyclic permutation test (CPT)*, that is valid in finite samples and allows **arbitrary fixed design matrix** and **arbitrary error distributions**, provided that the error terms are **exchangeable**. Exchangeability is weaker than the frequently made i.i.d. assumption. Further, the test is **non-randomized** if $1/\alpha$ is an integer and $n/(p-r) > 1/\alpha - 1$. The former condition is true for all common choices of α, e.g. $0.1, 0.05, 0.01, 0.005$. The latter requirement is also reasonable in various applications. For instance, when $\alpha = 0.05$, the condition reads $n/(p-r) > 19$, which is true if $n/p > 19$ or $p-r$ is small. Both are typical in social science applications. We demonstrate the power of CPT through extensive simulation studies and show it is comparable to the existing ones. CPT is the first procedure that is provably exact with reasonable power under

such weak assumptions. We want to emphasize that the goal of this chapter is not to propose a procedure that is superior to existing tests, but to expand the toolbox of exact inference and hopefully to motivate novel methodology for other problems.

The rest of the article is organized as follows: Section 3.2 discusses the motivation, the implementation and the theoretical property of cyclic permutation tests. In particular, Section 3.2.6 provides a summary of the implementation of CPT. In Section 3.3, we compare CPT with five existing tests through extensive simulation studies. To save space, we only present partial results and leave others to Appendix B.1. Section 3.4 provides a selective yet extensive literature review on this topic. One main goal is to introduce various stretegies for this problem demonstrating the difficulty of developing an exact test. We put this long review at the end of this chapter to avoid distraction. Section 3.5 concludes the chapter and discusses several related issues. All programs to replicate the results in this article can be found in `https://github.com/lihualei71/CPT`.

3.2 Cyclic Permutation Test

3.2.1 Main idea

Throughout the article we denote by $[n]$ the set $\{1, \ldots, n\}$. First we show that it is sufficient to consider the sub-hypothesis:

$$H_0 : \beta_1 = \ldots = \beta_r = 0. \tag{3.4}$$

In fact, for the general linear hypothesis 3.3, let $U_R \in \mathbb{R}^{p\times r}$ be an orthonormal basis of the column span of R and $V_R \in \mathbb{R}^{p\times(p-r)}$ be an orthonormal basis of the orthogonal complement. Then $\beta = U_R U_R^T\beta + V_R V_R^T\beta$. Let $\tilde{X} = (XU_R \vdots XV_R)$ and $\tilde{\beta} = \begin{pmatrix} U_R^T\beta \\ V_R^T\beta \end{pmatrix}$. Then the linear model (3.2) can be re-formulated as

$$y = \beta_0\mathbf{1} + XU_R(U_R^T\beta) + XV_R(V_R^T\beta) + \epsilon = \beta_0\mathbf{1} + \sum_{j=1}^{r}\tilde{X}_j\tilde{\beta}_j + \sum_{j=r+1}^{p}\tilde{X}_j\tilde{\beta}_j + \epsilon. \tag{3.5}$$

On the other hand, since R has full column rank, the null hypothesis (3.3) is equivalent to $H_0 : \tilde{\beta}_1 = \ldots = \tilde{\beta}_r = 0$, which is typically referred to as a sub-hypothesis (e.g. Adichie 1978). For this reason, we will focus on (3.4) without loss of generality throughout the rest of the chapter.

Our idea is to construct a pool of linear statistics $S = (S_0, S_1, \ldots, S_m)$ such that S is distributionally invariant under the *left shifting operator* π_L under the null, in the sense that $S \overset{d}{=} \pi_L(S)$

$$S \overset{d}{=} \pi_L(S) \overset{d}{=} \pi_L^2(S) \overset{d}{=} \cdots \overset{d}{=} \pi_L^m(S), \tag{3.6}$$

where

$$\pi_L^k(S) = (S_k, S_{k+1}, \ldots, S_m, S_0, S_1, \ldots, S_{k-1}), \quad k = 1, 2, \ldots, m. \tag{3.7}$$

Let Id denote the identity mapping, then $\mathcal{G} = \{\mathrm{Id}, \pi_L, \ldots, \pi_L^m\}$ forms a group, which we refer to as the *cyclic permutation group* (CPG). We say a pool of statistics S as invariant under CPG if S satisfies (3.6). The following trivial proposition describes the the main property of CPG invariance.

Proposition 3.2.1. *Assume that $S = (S_0, S_1, \ldots, S_m)$ is invariant under CPG. Let R_0 be the rank of S_0 in descending order, defined as $R_0 = \{j \geq 0 : S_j \geq S_0\}$. Then*

$$R_0 \succeq \mathrm{Unif}([m+1]) \Longrightarrow p \triangleq \frac{R_0}{m+1} \succeq \mathrm{Unif}([0,1]) \tag{3.8}$$

where $\succeq$ denotes stochastic dominance, $\mathrm{Unif}([0,1])$ *denotes the uniform distribution on* $[0,1]$. *Furthermore, $R_0 \sim \mathrm{Unif}([m+1])$ if S has no tie with probability* 1.

Proof. Let R_j be the rank of S_j in descending order as defined in (3.8). Then the invariance of S implies the invariance of $(R_0, R_1, \ldots, R_m)$. As a result,

$$R \overset{d}{=} R_1 \overset{d}{=} \cdots \overset{d}{=} R_m.$$

Then for any k,

$$\mathbb{P}(R_0 \geq k) = \frac{1}{m+1}\sum_{j=0}^{m} \mathbb{P}(R_j \geq k) = \frac{1}{m+1}\sum_{j=0}^{m} \mathbb{E}I(R_j \geq k) = \frac{1}{m+1}\mathbb{E}\big|\{j \geq 0 : R_j \geq k\}\big|.$$

Let $S_{(1)} \geq S_{(2)} \geq \cdots \geq S_{(m+1)}$ be the order statistics of $(S_0, \ldots, S_m)$, which may involve ties. Then by definition, $R_j \geq k$ whenever $S_j \leq S_{(k-1)}$ and thus,

$$\big|\{j \geq 0 : R_j \geq k\}\big| \geq m - k + 1$$

and thus $R_0 \succeq \mathrm{Unif}([m+1])$. When there is no tie, the set $\{R_0, R_1, \ldots, R_m\}$ is always $\{1, 2, \ldots, m+1\}$ and thus

$$\mathbb{P}(R_0 \geq k) = \frac{m-k+1}{m+1}.$$

□

Based on the p-value defined in (3.8), we can derive a test that rejects the null hypothesis if $p \leq \alpha$. We refer to this simple test as *marginal rank test* (MRT). The following trivial proposition shows that MRT is valid in finite samples and can be exact under mild conditions.

Proposition 3.2.2. *Suppose $S = (S_0, S_1, \ldots, S_m)$ is invariant under CPG under H_0 and let the p-value be defined as in* (3.8). *Then $\mathbb{P}_{H_0}(p \leq \alpha) \leq \alpha$. If $1/\alpha$ is an integer and $m+1$ is divisible by $1/\alpha$, then $\mathbb{P}_{H_0}(p \leq \alpha) = \alpha$.*

In practice, the reciprocals of commonly-used confidence levels (e.g. $0.1, 0.05, 0.01, 0.005$) are integers. In these cases it is sufficient to set $m = 1/\alpha - 1$ to obtain an exact test.

The rank used in MRT only gives one-sided information and may not be suitable for two-sided tests. More concretely, S_0 may be significantly different from $S_1, \ldots, S_m$ under the alternative but the sign of the difference may depend on the true parameters. An intuitive remedy is to apply MRT on the following modified statistics

$$\tilde{S}_j = |S_j - \text{med}\left(\{S_j\}_{j=0}^m\right)|. \tag{3.9}$$

If S_0 is significantly different from $S_1, \ldots, S_m$, $\tilde{S}_0$ is significantly larger than $\tilde{S}_1, \ldots, \tilde{S}_m$. The following proposition guarantees the validity of the transformation (3.9). In particular, the transformation in (3.9) satisfies the condition.

Proposition 3.2.3. *If $S = (S_0, S_1, \ldots, S_m)$ is invariant under CPG, then $\tilde{S} = (g(S_0; S), g(S_1; S), \ldots, g(S_m; S))$ is invariant under CPG for every g such that*

$$g(x; y) = g(x; \pi_L y).$$

In this article, we consider linear statistics

$$S_j = y^T \eta_j, \quad j = 0, 1, \ldots, m,$$

and apply MRT on $\tilde{S}_0, \ldots, \tilde{S}_m$ defined in (3.9). Partition X into $(X_{[r]}\ X_{[-r]})$ and β into $(\beta_{[r]}, \beta_{[-r]})$. The linear model (3.2) implies that

$$y^T \eta_j = (\mathbf{1}^T \eta_j)\beta_0 + (X_{[r]}^T \eta_j)^T \beta_{[r]} + (X_{[-r]}^T \eta_j)^T \beta_{[-r]} + \epsilon^T \eta_j. \tag{3.10}$$

In the next three subsections we will show how to construct η_j's to guarantee the type-I error control and to enhance power. Surpringly, the only distributional assumption on ϵ is the exchangeability:

A1 ϵ has exchangeable components, i.e. for any permutation π on $[n]$

$$(\epsilon_1, \ldots, \epsilon_n) \overset{d}{=} (\epsilon_{\pi(1)}, \ldots, \epsilon_{\pi(n)}).$$

3.2.2 Construction for type-I Error Control

Under H_0, (3.10) can be simplified as

$$y^T \eta_j = \underbrace{(\mathbf{1}^T \eta_j)\beta_0 + (X_{[-r]}^T \eta_j)^T \beta_{[-r]}}_{\text{deterministic part}} + \underbrace{\epsilon^T \eta_j}_{\text{stochastic part}}. \tag{3.11}$$

To ensure the distributional invariance of $\{y^T\eta_0, \ldots, y^T\eta_m\}$ to CPG, it is sufficient to construct η_j's such that the deterministic parts are identical for all j and the noise parts are invariant under CPG. To match the deterministic part, we can simply set $X_{[-r]}^T \eta_j$ to be independent of j.

C1 there exists $\gamma_{[-r]} \in \mathbb{R}^{p-r}$ such that

$$X_{[-r]}^T \eta_j = \gamma_{[-r]} \;\; j = 0, 1, \ldots, m.$$

To ensure the invariance of the stochastic part, intuitively η_j's should be left shifted transforms of each other. To be concrete, consider the case where $n = 6$ and $m = 2$. Then given any $\eta^* = (\eta_1^*, \eta_2^*, \eta_3^*, \eta_4^*, \eta_5^*, \eta_6^*)^T$, the following construction would imply the invariance to CPG:

$$\eta_0 = (\eta_1^*, \eta_2^*, \eta_3^*, \eta_4^*, \eta_5^*, \eta_6^*)^T, \quad \eta_1 = (\eta_3^*, \eta_4^*, \eta_5^*, \eta_6^*, \eta_1^*, \eta_2^*)^T, \quad \eta_2 = (\eta_5^*, \eta_6^*, \eta_1^*, \eta_2^*, \eta_3^*, \eta_4^*)^T.$$

To see this, note that

$$(\epsilon^T\eta_0, \epsilon^T\eta_1, \epsilon^T\eta_2)^T = \begin{pmatrix} \epsilon_1 & \epsilon_2 & \epsilon_3 & \epsilon_4 & \epsilon_5 & \epsilon_6 \\ \epsilon_5 & \epsilon_6 & \epsilon_1 & \epsilon_2 & \epsilon_3 & \epsilon_4 \\ \epsilon_3 & \epsilon_4 & \epsilon_5 & \epsilon_6 & \epsilon_1 & \epsilon_2 \end{pmatrix} \eta^*,$$

and

$$(\epsilon^T\eta_1, \epsilon^T\eta_2, \epsilon^T\eta_0)^T = \begin{pmatrix} \epsilon_5 & \epsilon_6 & \epsilon_1 & \epsilon_2 & \epsilon_3 & \epsilon_4 \\ \epsilon_3 & \epsilon_4 & \epsilon_5 & \epsilon_6 & \epsilon_1 & \epsilon_2 \\ \epsilon_1 & \epsilon_2 & \epsilon_3 & \epsilon_4 & \epsilon_5 & \epsilon_6 \end{pmatrix} \eta^*.$$

By assumption **A**1,

$$\begin{pmatrix} \epsilon_1 & \epsilon_2 & \epsilon_3 & \epsilon_4 & \epsilon_5 & \epsilon_6 \\ \epsilon_5 & \epsilon_6 & \epsilon_1 & \epsilon_2 & \epsilon_3 & \epsilon_4 \\ \epsilon_3 & \epsilon_4 & \epsilon_5 & \epsilon_6 & \epsilon_1 & \epsilon_2 \end{pmatrix} \stackrel{d}{=} \begin{pmatrix} \epsilon_5 & \epsilon_6 & \epsilon_1 & \epsilon_2 & \epsilon_3 & \epsilon_4 \\ \epsilon_3 & \epsilon_4 & \epsilon_5 & \epsilon_6 & \epsilon_1 & \epsilon_2 \\ \epsilon_1 & \epsilon_2 & \epsilon_3 & \epsilon_4 & \epsilon_5 & \epsilon_6 \end{pmatrix}.$$

As a result,

$$(\epsilon^T\eta_0, \epsilon^T\eta_1, \epsilon^T\eta_2) \stackrel{d}{=} (\epsilon^T\eta_1, \epsilon^T\eta_2, \epsilon^T\eta_0).$$

Using the same argument we can show $(\epsilon^T\eta_0, \epsilon^T\eta_1, \epsilon^T\eta_2) \stackrel{d}{=} (\epsilon^T\eta_2, \epsilon^T\eta_0, \epsilon^T\eta_1)$ and thus the invariance of $(\epsilon^T\eta_0, \epsilon^T\eta_1, \epsilon^T\eta_2)$ to CPG.

In general, if n is divisible by $m + 1$ with $n = (m + 1)t$, then we can construct η_j's as a left shifted transform of a vector η^*, i.e.

$$\eta_j = \pi_L^{tj}(\eta^*) \tag{3.12}$$

where π_L is the left shifting operator defined in (3.7). More generally, if $n = (m+1)t + s$ for some integers t and $0 \le s \le m$, we can leave the last s components to be the same across η_j's while shifting the first $(m + 1)t$ entries as in (3.12).

C2 there exists $\eta_* \in \mathbb{R}^n$ such that

$$\eta_j = \left[\pi_L^{tj}((\eta_1^*, \ldots, \eta_{(m+1)t}^*)), \eta_{(m+1)t+1}^*, \ldots, \eta_n^*\right]^T,$$

where $t = \lfloor n/(m + 1) \rfloor$.

Proposition 3.2.4. *Under assumption **A**1, $(y^T\eta_0, \ldots, y^T\eta_m)$ is distributionally invariant under CPG if $(\eta_0, \ldots, \eta_m)$ satisfy **C**1 and **C**2.*

Proof. It is left to prove the invariance of $(\epsilon^T\eta_0, \ldots, \epsilon^T\eta_m)$ to CPG. Further, since the last $n-(m+1)t$ terms are the same for all j, it is left to prove the case where n is divisible by $m+1$. Let $\tilde{\Pi}$ be the permutation matrix corresponding to π_L^t. Then **C**2 implies that

$$\begin{aligned}
\pi_L(\epsilon^T\eta_0, \epsilon^T\eta_1, \ldots, \epsilon^T\eta_m) &= (\epsilon^T\eta_1, \ldots, \epsilon^T\eta_m, \epsilon^T\eta_0) \\
&= (\epsilon^T\tilde{\Pi}\eta_*, \ldots, \epsilon^T\tilde{\Pi}^m\eta_*, \epsilon^T\eta_*) \\
&= (\epsilon^T\tilde{\Pi}\eta_*, \ldots, \epsilon^T\tilde{\Pi}^m\eta_*, \epsilon^T\tilde{\Pi}^{m+1}\eta_*) \qquad (\text{Since } \tilde{\Pi}^{m+1} = \mathrm{Id}) \\
&= (\epsilon^T\eta_*, \ldots, \epsilon^T\tilde{\Pi}^{m-1}\eta_*, \epsilon^T\tilde{\Pi}^m\eta_*) \qquad (\text{Since } \tilde{\Pi}\epsilon \overset{d}{=} \epsilon) \\
&= (\epsilon^T\eta_0, \epsilon^T\eta_1, \ldots, \epsilon^T\eta_m). \qquad (3.13)
\end{aligned}$$

Repeating (3.13) for $m-1$ times, we prove the invariance of $(\epsilon^T\eta_1, \ldots, \epsilon^T\eta_m)$ under CPG. □

Now we discuss the existence of $(\eta_*, \gamma_{[-r]})$. Note that η_j is a linear transformation of η^*. Let $\Pi_j \in \mathbb{R}^{n\times n}$ be the matrix such that $\eta_j = \Pi_j\eta^*$. Then **C**1 and **C**2 imply that

$$\begin{pmatrix} -I_{p-r} & X_{[-r]}^T \\ -I_{p-r} & X_{[-r]}^T\Pi_1 \\ \vdots & \vdots \\ -I_{p-r} & X_{[-r]}^T\Pi_m \end{pmatrix} \begin{pmatrix} \gamma_{[-r]} \\ \eta_* \end{pmatrix} = 0. \qquad (3.14)$$

The above linear system has $(m+1)(p-r)$ equations and $n+p-r$ unknowns. Therefore, a non-zero solution always exists if $(m+1)(p-r) < n+p-r$.

Theorem 3.2.5. *Under assumption **A**1,*

(a) (3.14) *always has a non-zero solution if*

$$n/(p-r) > m. \qquad (3.15)$$

(b) for any solution $(\gamma_{[-r]}, \eta^*)$ of (3.14)*,*

$$(y^T\eta^*, y^T\Pi_1\eta^*, \cdots y^T\Pi_m\eta^*)$$

*is invariant under CPG under H_0, where $\Pi_j \in \mathbb{R}^{n\times n}$ is the coefficient matrix that maps η^* to η_j defined in **C**2.*

Suppose $\alpha = 0.05$ for illustration and set $m = 1/\alpha - 1 = 19$. Then the condition (3.15) reads

$$n > 19(p-r).$$

Even when $r = 1$, this is a mild condition in many applications. On the other hand, when r is large but $p - r$ is small, then (3.15) can still be satisfied even if $p > n$. This is in sharp contrast to regression F-tests and permutation F-tests that require fitting the full model and thus $p \leq n$. Furthermore, it is worth emphasizing that Theorem 3.2.5 allows arbitrary design matrices. This is fundamentally different from the asymptotically valid tests which always impose regularity conditions on X.

3.2.3 Construction for high power when $r = 1$

To guarantee reasonable power, we need $y^T\eta_0$ to be significantly different from the other statistics under the alternative. In this subsection we focus on the case where $r = 1$ to highlight the key idea.

When $\beta_1 \neq 0$, (3.10) implies that

$$y^T\eta_j = (X_1^T\eta_j)\beta_1 + W_j$$

where $W_j = \epsilon^T\eta_j + (\mathbf{1}^T\eta_*)\beta_0 + (X_{[-1]}^T\eta_*)^T\beta_{[-1]}$ and $(W_1, \ldots, W_m)$ is invariant under CPG by Theorem 3.2.5. To enhance power, it is desirable that $X_1^T\eta_0$ lies far from $\{X_1^T\eta_1, \ldots, X_1^T\eta_m\}$. In particular, we impose the following condition on η_j's:

C3$^{(1)}$ there exists $\gamma_1, \delta \in R$, such that

$$X_1^T\eta_j = \gamma_1 \ \ (j = 1, 2, \ldots, m), \quad X_1^T\eta_0 = \gamma_1 + \delta.$$

Putting **C**1, **C**2 and **C**3$^{(1)}$ together, we obtain the following linear system,

$$\left(-e_{1,p(m+1)} \vdots A(X)^T \right) \begin{pmatrix} \delta \\ \gamma \\ \eta \end{pmatrix} = 0, \tag{3.16}$$

where $e_{1,p(m+1)}$ is the first canonical basis in $\mathbb{R}^{p(m+1)}$ and

$$A(X) = \begin{pmatrix} -I_p & -I_p & \cdots & -I_p \\ X & \Pi_1^T X & \cdots & \Pi_m^T X \end{pmatrix} \in \mathbb{R}^{(n+p)\times(m+1)p}. \tag{3.17}$$

This linear system has $(m + 1)p$ equations and $n + p + 1$ variables. Thus it always has a non-zero solution if

$$n + p + 1 > p(m + 1) \Longleftrightarrow n \geq pm.$$

When $\alpha = 0.05$ and $m = 19$, this condition is still reasonable in many settings.

The normalized gap $\delta/\|\eta\|$ can be regarded as a proxy of power. Write γ for $\begin{pmatrix} \gamma_1 \\ \gamma_{[-1]} \end{pmatrix}$ and e_1 for the first canonical basis vector of $\mathbb{R}^p$. Putting conditions **C**1-**C**3 together, it is natural to consider the following optimization problem:

$$\max_{\delta\in\mathbb{R},\gamma\in\mathbb{R}^p,\eta\in\mathbb{R}^n,\|\eta\|_2=1} \delta, \quad \text{s.t.} \left(-e_{1,p(m+1)} \vdots A(X)^T \right) \begin{pmatrix} \delta \\ \gamma \\ \eta \end{pmatrix} = 0. \tag{3.18}$$

This linear programming problem can be solved by fitting a linear regression and it permits a closed-form solution. Let $O^*(X)$ denote the optimal value of the objective function, i.e. maximum achievable value of δ in this case.

Theorem 3.2.6. *Assume that $n \geq pm$. Let*

$$B(X) = \begin{pmatrix} (I - \Pi_m)^T X & (\Pi_1 - \Pi_m)^T X & \cdots & (\Pi_{m-1} - \Pi_m)^T X \end{pmatrix} \in \mathbb{R}^{n \times mp}. \tag{3.19}$$

Partition $B(X)$ into $[B(X)_1\ B(X)_{[-1]}]$ where $B(X)_1$ is the first column of $B(X)$. Further let

$$\tilde{\eta} = (I - H_{[-1]})B(X)_1, \quad \text{where } H_{[-1]} = B(X)_{[-1]}(B(X)_{[-1]}^T B(X)_{[-1]})^+ B(X)_{[-1]}^T$$

where $+$ denotes the Moore-Penrose generalized inverse. Then $O^*(X) = \|\tilde{\eta}\|_2$ and one global maximizer of (3.18) is given by

$$\eta^*(X) = \tilde{\eta}/\|\tilde{\eta}\|_2, \quad \delta^*(X) = \|\tilde{\eta}\|_2.$$

Remark 3.2.7. *When $B(X)_{[-1]}$ has full column rank, $\tilde{\eta}$ is the residual vector by regressing $B(X)_1$ on $B(X)_{[-1]}$ and $\|\tilde{\eta}\|_2^2$ is the residual sum of squares. Both can be easily computed using standard softwares. If $B(X)_{[-1]}$ does not have full column rank, then $\tilde{\eta}$ is the minimum norm ordinary least squares by regressing $B(X)_1$ on $B(X)_{[-1]}$, which is the limit of ridge estimator when the penalty tends to zero and is the limiting solution of standard gradient descent initialized at zero (e.g. Hastie et al. 2019).*

Proof. First, (3.16) can be equivalently formulated as

$$B(X)^T \eta = \delta e_{1,pm}.$$

This can be further rewritten as

$$\delta = B(X)_1^T \eta, \quad B(X)_{[-1]}^T \eta = 0. \tag{3.20}$$

For any η satisfying the second constraint,

$$H_{[-1]}\eta = 0,$$

and thus

$$B(X)_1^T \eta = B(X)_1^T (I - H_{[-1]})\eta = \tilde{\eta}^T \eta.$$

As a result,

$$\max_{B(X)_{[-1]}^T \eta = 0, \|\eta\|_2 = 1} B(X)_1^T \eta \leq \max_{\|\eta\|_2 = 1} \tilde{\eta}^T \eta = \|\tilde{\eta}\|_2.$$

In other words, we have shown that $\delta^*(X) \leq \|\tilde{\eta}\|_2$. On the other hand, the vector $\tilde{\eta}/\|\tilde{\eta}\|_2$ satisfies the constraint (3.20) and

$$B(X)_1^T \tilde{\eta}/\|\tilde{\eta}\|_2 = \|\tilde{\eta}\|_2.$$

This shows that $\delta^*(X) \geq \|\tilde{\eta}\|_2$. In this case, it is obvious that $O^*(X) = \delta^*(X)$. Therefore, $O^*(X) = \|\tilde{\eta}\|_2$ and one maximizer is $\eta^*(X) = \tilde{\eta}/\|\tilde{\eta}\|_2$. □

3.2.4 Construction for high power when $r > 1$

Similar to $\mathbf{C}3^{(1)}$, we impose the following restriction on η.

C3 there exists $\gamma_{[r]}, \delta \in \mathbb{R}^r$, such that

$$X_{[r]}^T \eta_j = \gamma_{[r]} \ \ (j = 1, 2, \ldots, m), \quad X_{[r]}^T \eta_0 = \gamma_{[r]} + \delta.$$

Combining with (3.14), we obtain an analogue of (3.16) as follows.

$$\left(-e_{1,p(m+1)}, \ldots, -e_{r,p(m+1)} \vdots A(X)^T \right) \begin{pmatrix} \delta \\ \gamma \\ \eta \end{pmatrix} = 0, \tag{3.21}$$

where $A(X)$ is defined in (3.17) and $\gamma = \begin{pmatrix} \gamma_{[r]} \\ \gamma_{[-r]} \end{pmatrix}$. This linear system involves $p(m+1)$ equations and $n + p + r$ variables. Therefore it always has a non-zero solution if

$$n + p + r > p(m+1) \Longleftrightarrow n \geq pm - r + 1.$$

Unlike the univariate case, there are infinite ways to characterized the signal strength since δ is multivariate. A sensible class of criteria is to maximize a quadratic form

$$\max_{\delta \in \mathbb{R}^r, \gamma \in \mathbb{R}^p, \eta \in \mathbb{R}^n, \|\eta\|_2 = 1} \delta^T M \delta \quad \text{s.t.} \left(-e_{1,p(m+1)}, \ldots, -e_{r,p(m+1)} \vdots A(X)^T \right) \begin{pmatrix} \delta \\ \gamma \\ \eta \end{pmatrix} = 0. \tag{3.22}$$

The following theorem gives the optimal solution given any weighting matrix M. Let $O^*(X)$ denote the optimal value of the objective function.

Theorem 3.2.8. *Assume that $n \geq pm - r + 1$. Let $B(X)$ be defined in (3.19). Partition $B(X)$ into $(B(X)_{[r]}\ B(X)_{[-r]})$ where $B(X)_{[r]}$ is the matrix formed by the first r columns of $B(X)$. Let*

$$M_r(X) = (I - H_{[-r]}) B(X)_{[r]} M B(X)_{[r]}^T (I - H_{[-r]}),$$

where

$$H_{[-r]} = B(X)_{[-r]} (B(X)_{[-r]}^T B(X)_{[-r]})^+ B(X)_{[-r]}^T$$

Further let $\lambda_{\max}(M_r(X))$ denote the maximum eigenvalue, u be any eigenvector corresponding to it and $\tilde{\eta} = (I - H_{[-r]})u$. Then $O^*(X) = \lambda_{\max}(M_r(X))$ and

$$\eta^*(X) = \tilde{\eta}/\|\tilde{\eta}\|_2, \quad \delta^*(X) = B(X)_{[r]}^T \eta^*(X)$$

is an optimal solution of (3.22).

Proof. Similar to the proof of Theorem 3.2.6, we first rewrite (3.21) as

$$B(X)_{[r]}^T \eta = \delta, \quad B(X)_{[-r]}^T \eta = 0.$$

As a result, η lies in the row null space of $B(X)_{[-r]}$ and thus there exists $\zeta \in \mathbb{R}^n$ such that

$$H_{[-r]}\eta = 0.$$

Then

$$\delta^T M \delta = \eta^T (I - H_{[-r]}) B(X)_{[r]} M B(X)_{[r]}^T (I - H_{[-r]})\eta = \eta^T M_r(X) \eta.$$

Since $\|\eta\|_2 \leq 1$,

$$\delta^T M \delta \leq \lambda_{\max}(M_r(X)).$$

On the other hand, for any eigenvector u of $M_r(X)$ corresponding to its largest eigenvalue, let $\tilde{\eta} = (I - H_{[-r]})u$ and $\eta = \tilde{\eta}/\|\tilde{\eta}\|_2$, then

$$\eta^T M_r(X)\eta = \lambda_{\max}(M_r(X)), \quad B(X)_{[-r]}\eta = 0, \quad \|\eta\|_2 = 1.$$

Thus, $\eta^*(X) = \tilde{\eta}/\|\tilde{\eta}\|_2$ is an optimal solution. As a result, $\delta^*(X) = B(X)_{[r]}^T \eta^*(X)$ and $O^*(X) = \lambda_{\max}(M_r(X))$. □

Although Theorem 3.2.8 gives the solution of (3.22) for arbitrary weight matrix M, it is not clear which M is the best choice. Since

$$\eta_j^T y = \delta^T \beta_{[r]} I(j = 0) + \tilde{W}_j$$

where $\tilde{W}_j = \gamma^T \beta + \eta_j^T \epsilon$ is invariant under CPG. Thus, $\delta^T \beta_{[r]}$ characterizes the signal strength. In principle, the "optimal" weight matrix should be depend on prior knowledge of $\beta_{[r]}$. For instance, for a Bayesian hypothesis testing problem with a prior distribution Q on $\beta_{[r]}$ under the alternative, the optimal weight matrix is $M = \mathbb{E}_Q \left[\beta_{[r]} \beta_{[r]}^T\right]$.

3.2.5 Pre-ordering rows of design Matrix

Given any X, we can easily calculate the proxy of signal strength $O^*(X)$ by Theorem 3.2.6 and Theorem 3.2.8. However, the optimal value is not invariant to row permutation of X, that is, for any permutation matrix $\Pi \in \mathbb{R}^{n \times n}$,

$$O^*(X) \neq O^*(\Pi X)$$

in general. Roughly speaking, this is because $\delta^*(X)$ involves left shifting operator, which depends on the arrangement of the rows of X. Figure 3.1 illustrates the variability of $O^*(\Pi X)$ as a function of Π for a fixed matrix with 8 rows and 3 columns, generated with i.i.d. Gaussian entries.

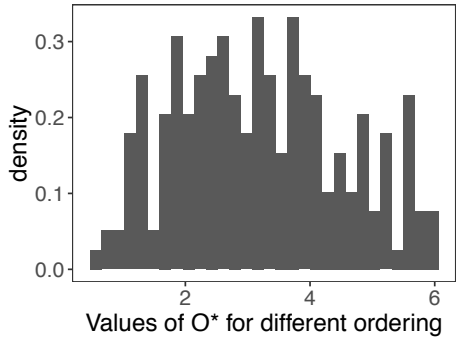


Figure 3.1: Histograms of $O^*(\Pi X)$ for a realization of a random matrix with i.i.d. Gaussian entries.

Notably, even in such regular cases the variability is non-negligible. This motivates the following secondary combinatorial optimization problem:

$$\max_{\Pi} O^*(\Pi X). \tag{3.23}$$

This is a non-linear travelling salesman problem. Note that we aim at finding a solution with reasonably large objective value insteading finding the global maximum of (3.23), which is NP-hard. For this reason, we solve (3.23) by Genetic Algorithm (GA), which is efficient albeit without worst-case convergence guarantee. In a nutshell, GA maintains a *population* of permutations, generate new permutations by two operations: *crossover* and *mutation*, and *evolves* the population via a mechanism called *selection*, based on the objective value. We refer the readers to Michalewicz (2013) for more details.

We compare GA with a simple competing algorithm that randomly selects ordering and keeps the one yielding the largest objective value. We refer to this method as Stochastic Search (SS). Although this competitor is arguably too weak and more efficient algorithms may exist (e.g. continuous relaxation of permutation matrices into double-stochastic matrices), our goal here is simply to illustrate the effectiveness of GA instead of to claim the superiority of GA. We compare the performance of GA and SS on three matrices with $n = 1000$ and $p = 20$, generated from random one-way ANOVA matrices with exactly one entry in each row at a unifromly random position, random matrices with i.i.d. standard normal entries and random matrices with i.i.d. standard Cauchy entries. The results are plotted in Figure 3.2 where the y-axis measures $O^*(\Pi X)$, scaled by the maximum achieved by GA and SS for illustration, and the x-axis measures the number of random samples each algorithm accesses. The population size is set to be 10 for GA in all scenarios.

3.2.6 Implementation of CPT

Based on previous subsections, we summarize the implementation of CPT below:

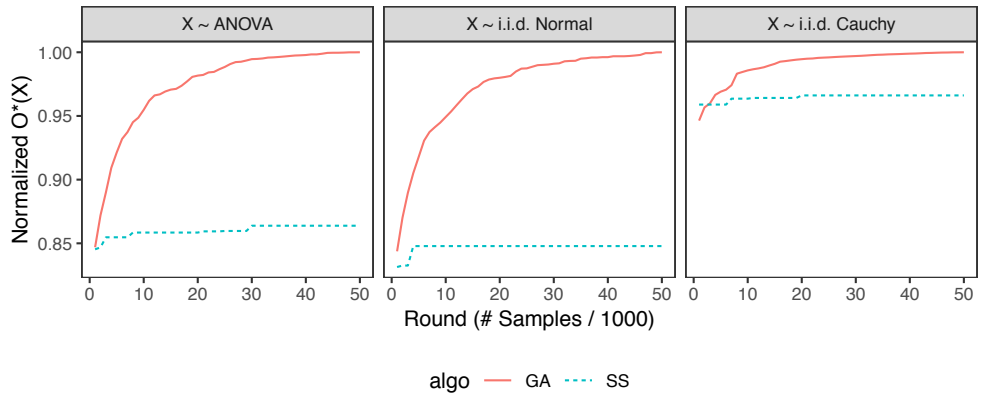


Figure 3.2: Histograms of $O^*(\Pi X)$ for three matrices as realizations of random one-way ANOVA matrices with exactly one entry in each row at a unifromly random position, random matrices with i.i.d. standard normal entries and random matrices with i.i.d. standard Cauchy entries, respectively.

Step 1 Compute a desirable pre-ordering Π_0 for the combinatorial optimization problem

$$\max_{\Pi} O^*(\Pi X),$$

where $O^*(\cdot)$ is defined in Theorem (3.2.6) when $r = 1$ and is defined in Theorem (3.2.8) when $r > 1$;

Step 2 Replace y and X by $\Pi_0 y$ and $\Pi_0 X$;

Step 3 Compute η^* via the formula in Theorem (3.2.6) or Theorem 3.2.8;

Step 4 Compute $S_j = \eta_j^T y$ for $j = 0, 1, \ldots, m$ where

$$\eta_j = \left[\pi_L^{tj}((\eta_1^*, \ldots, \eta_{(m+1)t}^*)), \eta_{(m+1)t+1}^*, \ldots, \eta_n^*\right]^T, \quad t = \lfloor n/(m+1) \rfloor;$$

Step 5 Compute $\tilde{S}_j = |S_j - \text{med}\left(\{S_j\}_{j=0}^m\right)|$;

Step 6 Compute the p-value $p = R_0/(m+1)$ where R_0 is the rank of $\tilde{S}_0$ in the set $\{\tilde{S}_0, \ldots, \tilde{S}_m\}$ in descending order;

Step 7 Reject the null hypothesis if $p \leq \alpha$.

The inputs of CPT include the design matrix X, the outcome vector y, the confidence level α, the number of statistics $m + 1$ and a sub-routine to solve Step 1. As the default, we set $m = \lceil 1/\alpha \rceil - 1$ and use Genetic Algorithm, implemented in `R` package `gaoptim`, to solve Step 1.

3.3 Experiments

3.3.1 Setup

To examine the power of our procedure, we conduct extensive numerical experiments. In all experiments below, we fix the sample size $n = 1000$ and consider three values $25, 33, 40$ for dimension p such that the sample per parameter $n/p \approx 40, 30, 25$. Given a dimension p, we consider three types of design matrices: realizations of random one-way ANOVA matrices with exactly one entry in each row at a unifromly random position, realizations of random matrices with i.i.d. standard normal entries and realizations of random matrices with i.i.d. standard Cauchy entries. For each type of design matrix, we generate 50 independent copies. Given each X, we generate 3000 copies of ϵ with i.i.d. entries from the standard normal distribution and the standard Cauchy distribution.

We consider two variants of CPTs – CPT with random ordering and CPT with GA pre-ordering, as well as five competing tests: (1) t/F tests; (2) permutation t/F tests which approximates the null distribution of the t/F statistic by the permutation distribution with $X_{[r]}$ reshuffled; (3) Freedman-Lane test (Freedman and Lane 1983; Anderson and Robinson 2001) which approximates the null distribution of the t/F statistic by the permutation distribution with regression residuals reshuffled; (4) asymptotic z-test from least absolute deviation (LAD) regression; (5) Group Bound method (Meinshausen 2015). For both permutation tests, we calculate the test based on 1000 random permutation. To further demonstrate the importance of pre-ordering step of CPT, we consider a weaker GA pre-ordering with 1000 random samples and a stronger GA pre-ordering with 10000 random samples. Three variants of CPTs are abbreviated as CPTw for CPT with weak pre-ordering, CPTs for strong pre-ordering and CPTr for CPT with random ordering. All tests will be performed with level $\alpha = 0.05$ and the number of statistics in CPT is set to be $m + 1 = 20$.

3.3.2 Testing for a single coordinate

In the first experiment, we consider testing a single coordinate, i.e. $r = 1$. Given a design matrix X and an error distribution F, we start by computing a benchmark signal-to-noise ratio β_1^* such that the t/F tests have approximately 20% power, through Monte-Carlo simulation, when y is generated from

$$y = X_1\beta_1^* + \epsilon, \quad \text{where } \epsilon_i \overset{i.i.d.}{\sim} F.$$

Then all tests are performed on X and the following 18000 outcome vectors $y_s^{(b)}$, respectively:

$$y_s^{(b)} \triangleq X_1(s\beta_1^*) + \epsilon^{(b)}, \quad \text{where } s = 0, 1, \ldots, 5,\ b = 1, \ldots, 3000.$$

For each s, the proportion of rejection among 3000 ϵ's is computed. When $s = 0$, this proportion serves as an approximation of the type-I error and should be closed to or below α for a valid test; when $s > 0$, it serves as an approximation of power and should be large for a powerful test.

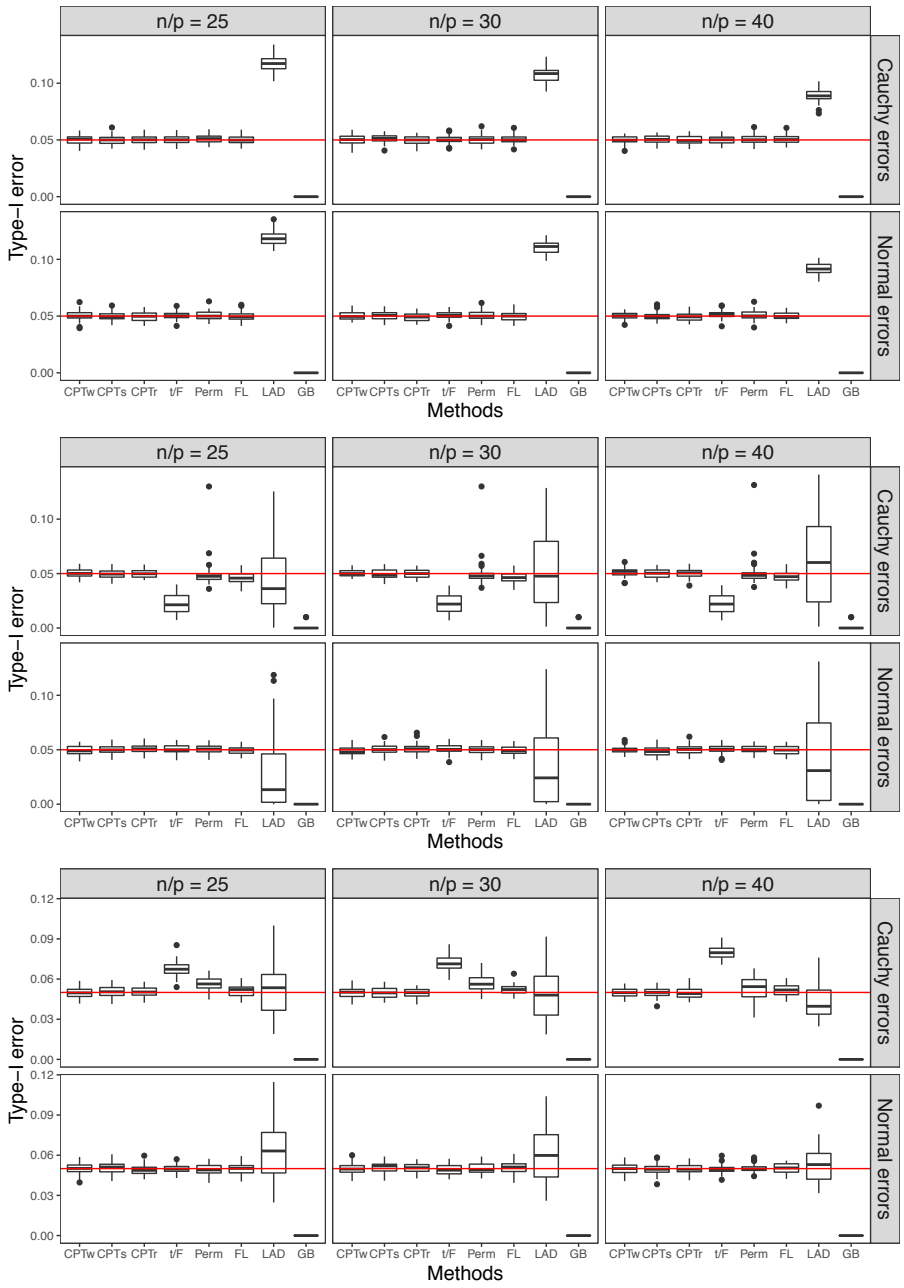


Figure 3.3: Monte-Carlo type-I error for testing a single coordinate with three types of X's: (top) realizations of random matrices with i.i.d. standard normal entries; (middle) realizations of random matrices with i.i.d. standard Cauchy entries; (bottom) realizations of random one-way ANOVA design matrices.

Figure 3.3 displays the type-I error of all tests for three types of design matrices. The box-plots display the variation among 50 independent copies of design matrices. In all cases, three variants of CPTs are valid as guaranteed by theory and Group Bound method is overly conservative. Permutation tests and Freedman-Lane tests also appear to be valid in our simulation settings even though there is no theoretical guarantee for heavy-tailed errors. When errors are Gaussian, t-test is valid as guaranteed by theory but can be conservative or anti-conseravative (i.e. invalid) with heavy-tailed errors depending on the design matrix. Interestingly, for one-way ANOVA, t-test becomes less valid as the sample size per parameter increases. On the other hand, LAD-based tests are anti-conservative when X is a realization of Gaussian matrices with both Gaussian and Cauchy errors, although the validity can be proved asymptotically under regularity conditions that are satisfied by realizations of Gaussian matrices with high probability (e.g. Pollard 1991).

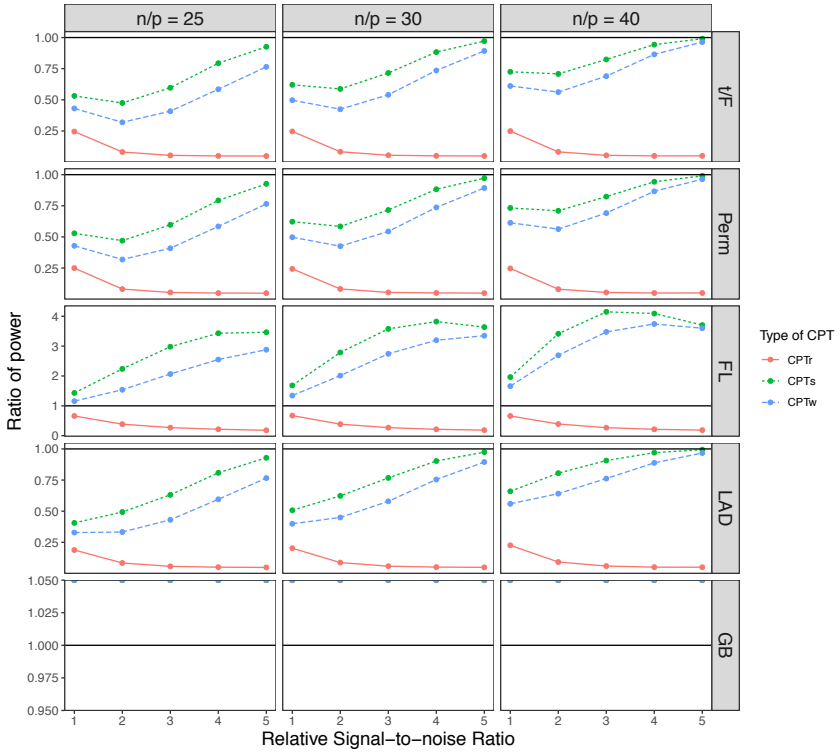


Figure 3.4: Median power ratio between each variant of CPT and each competing test for testing a single coordinate with realizations of Gaussian matrices and Gaussian errors. The black solid line marks the equal power. The missing values in the last row correspond to infinite ratios.

To save space, we only show results for the case where the design matrices are realizations of Gaussian matrices and errors are Gaussian in Figure 3.4 and the case where the design

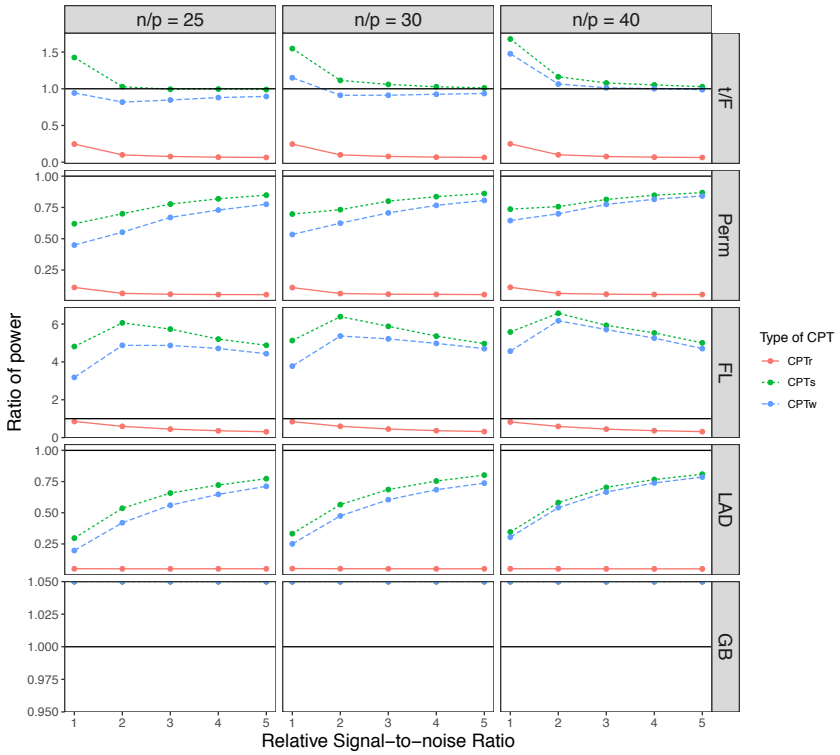


Figure 3.5: Median power ratio between each variant of CPT and each competing test for testing a single coordinate with realizations of Cauchy matrices and Cauchy errors. The black solid line marks the equal power. The missing values in the last row correspond to infinite ratios.

matrices are realizations of Cauchy matrices and errors are Cauchy in Figure 3.5, respectively. The results for other cases will be presented in Appendix B.1. All figures plot the median power ratio, from 50 independent copies of X's, between each variant of CPT (CPTw, CPTs and CPTr) and each competing test. First we see that the Group Bound method has zero power in all scenarios and thus the power ratios are infinite and missing in the plots. Second, the pre-ordering step is significant in raising the power of CPT. Third, the relative power of CPT becomes larger as n/p increases. In the first case, it is not surprising that t-tests is the most powerful ones because it is provably the uniformly most powerful unbiased (UMPU) test for linear models with Gaussian errors. The efficiency loss of CPTs against t-tests, permutation t-tests and LAD-based tests is moderate in general and is low when the sample size per parameter and the signal-to-noise ratio is large. In the second case, CPTs is more powerful than t-tests, although it is still less powerful than permutation t-tests and LAD-based tests. In both cases, CPTs is more powerful than Freedman-Lane tests even when $n/p = 25$ and the signal-to-noise ratio is small.

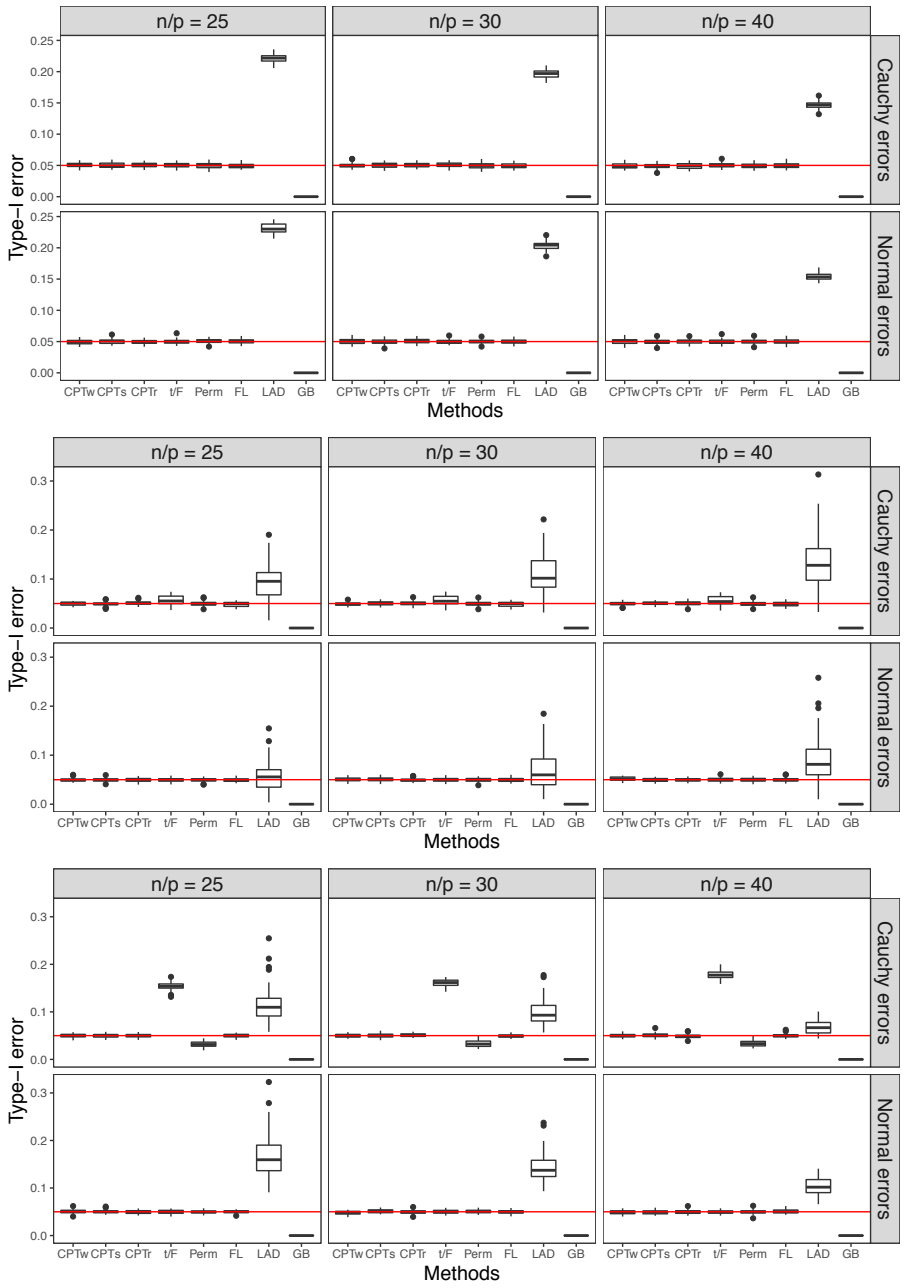


Figure 3.6: Monte-Carlo type-I error for testing five coordinates with three types of X's: (top) realizations of random matrices with i.i.d. standard normal entries; (middle) realizations of random matrices with i.i.d. standard Cauchy entries; (bottom) realizations of random one-way ANOVA design matrices.

3.3.3 Testing for multiple coordinates

Next we consider testing the first five coordinates with a Bayesian alternative hypothesis

$$\beta_{[5]} \sim N(\mathbf{1}_5, \Sigma), \quad \Sigma = \text{diag}(0.2, 0.4, 0.6, 0.8, 1).$$

All other settings are exactly the same as Section 3.3.2, except that t-tests and permutation t-tests are replaced by F-tests and permutation F-tests. For CPT, we choose the weight matrix $M = \mathbb{E}[\beta_{[5]}\beta_{[5]}^T]$. Figure 3.6 displays the Monte-Carlo type-I error of all tests. The results are qualitatively the same as the experiment in Section 3.3.2 except that F-tests and LAD-based tests become more invalid. To save space, all power comparisons are presented in Appendix B.1.

3.4 1908-2018: A Selective Review of The Century-Long Effort

Linear model is one of the most fundamental object in the history of statistics and has been developed for over a century. Nowadays it is still among most widely-used models for data analysts to demystify complex data as well as most powerful tools for statisticians to understand complicated methods and expand the toolbox for advanced tasks. It is impossible to exhaust the literature for this century-long problem. We thus provide a selective yet extensive review to highlight milestones in the past century. In particular, we will focus on the linear hypothesis testing and the estimation, which can yield the former, for vanilla linear models. We will focus on the linear models with general covariates and briefly discuss the simplified forms including location problems and ANOVA problems when necessary. However, we will exclude the topics including high dimensional sparse linear models, selective inference for linear models, linear models with dependent errors, high breakdown regression methods, linear time series, and generalized linear models. We should emphasize that these topics are at least equally important as those discussed in this section and they are exclude simply to avoid digression. Furthermore, as mentioned earlier, one purpose of this review is to highlight various stretegies for this problem and the difficulty of developing an exact test.

3.4.1 Normal theory based tests

Motivated by the seminal work by Student (1908b) and Student (1908a) which propose the one-sample and two-sample t-tests, Ronald A. Fisher derived the well-known t-distribution (Fisher 1915) and applied it to testing for a single regression coefficient in homoskedastic Gaussian linear models (Fisher 1922). In his 1922 paper, he also derived an equivalent form of F test for testing the global null under the same setting. Later he derived the F-distribution (Fisher 1924), which he characterized through "z", the half logarithm of F-statistics, and proposed the F-test for ANOVA, a special case of linear hypothesis in homoskedastic Gaussian

linear models. Both tests were elaborated in his insightful book (Fisher 1925) and the term "F-test" was coined by George W. Snedecor (Snedecor 1934).

This paramount line of work established the first generation of rigorous statistical test for linear models. They are *exact tests* of linear hypotheses in linear models with independent and identically distributed normal errors and almost arbitrary fixed-design matrices. Despite the exactness of the tests without any assumption on the design matrices, the normality assumption can rarely be justified in practice.Early investigation of the test validity with non-normal errors can be dated back to Egon S. Pearson (Pearson 1929; Pearson and Adyanthāya 1929; Pearson 1931). Unlike the large-sample theory based framework that is standard nowadays, the early work take an approximation perspective to imporve the validity for extremely small sample. It was furthered in the next a few decades (e.g. Eden and Yates 1933; Bartlett 1935; Geary 1947; Gayen 1949, 1950; David and Johnson 1951a; David and Johnson 1951b; Box 1953; Box and Watson 1962; Pearson and Please 1975) and it was mostly agreed that the regression t-test is extremely robust to non-normal errors with moderately large sample (e.g. > 30) while the regression F-test is more sensitive to the deviation from normality. It is worth emphasizing that these works were either based on mathematically unrigorous approximation or based on rigorous Edgeworth expansion theory that could be justified rigorously (e.g. Esseen 1945b; Wallace 1958; Bhattacharya and Ghosh 1978) in the asymptotic regime that the sample size tends to infinity while the dimension of the parameters stays relatively low (e.g. a small constant).

Later on, due to the popularization of rigorous large-sample theory in 1950s (e.g. Chernoff 1956) pioneered by Doob (1935), Wilks (1938), Mann and Wald (1943), and Wald (1949), investigators started to look at the regression test validity in certain asymptotic regimes. This can be dated back to Friedhelm Eicker (Eicker 1963, 1967), to the best of our knowledge, and developed by Peter J. Huber in his well-known and influential paper (Huber 1973b), which shows that the least square estimate is jointly asymptotically normal if and only if the maximum leverage score tends to zero. This clean and powerful result laid the foundation to asymptotic analysis for t-tests and F-tests (e.g. Arnold 1980). Notably these early works do not assume the dimension p stays fixed, as the simplified arguments in standard textbooks. Before 1990, the large-sample theory for least squares estimators were well established in the regime where the sample size per parameter n/p grows to infinity, under regularity conditions on the design matrices and on the errors, usually with independent and identically distributed elements and finite moments. It shows that both t-tests and F-tests are asymptotically valid and can be approximated by z-tests and χ^2-tests, respectively. For t-tests, the robustness to non-normality was even established without the typical regularity conditions (e.g. Zellner (1976) and Jensen (1979) for spherically invariant errors, Efron (1969), Cressie (1980), Benjamini (1983), and Pinelis (1994) for orthant symmetric errors) or beyond the aforementioned regime (e.g. Lei et al. 2018). By contrast, though similar results exist for F-tests (e.g. Zellner 1976), more non-robustness results were established. For instance, a line of work (e.g. Boos and Brownie 1995; Akritas and Arnold 2000; Calhoun 2011; Anatolyev 2012) showed that F-tests are asymptotically invalid, unless the errors are normal, in the moderate dimensional regime where n/p stays bounded as n approaches infinity, although

correction is available under much stronger assumptions on the design matrix or the coefficient vectors. Even with normal errors, Zhong and Chen (2011) showed that the power of F-tests diminishes as n/p approaches 1. In a nutshell, there has been tremendous effort in the past century investigating the robustness of regression t-tests and F-tests and it was agreed that t-tests are insensitive to non-normality, high dimension and irregularity of design matrices to certain extent while F-tests are less robust in general.

3.4.2 Permutation tests

Despite the tremendous attention on regression t-tests and F-tests, other methodology emerged in parallel as well. The earliest alternative is the permutation test, which justifies the significance of the test through the so-called "permutation distribution". However, the early model to justify permutation tests is the "randomization model" in contrast to the "population model" that we considered in (3.2). The "randomization model" was introduced by Jerzy S. Neyman in his master thesis (Neyman 1923),which is also known as Neyman-Rubin model (Rubin 1974), or design-based inference (Särndal et al. (1978), in contrast to model-based inference), or "conditional-on-errors" model (Kennedy (1995), in contrast to "conditional-on-treatment" model), and the term was coined by Ronald A. Fisher in 1926 (Fisher 1926). The theoretical foundation of permutation test was laid by Edwin J. G. Pitman in his three seminal papers (Pitman 1937a,b; Pitman 1938), where the last two were studied for regression problems, albeit under the "randomization model". The early work view permutation tests as better devices in terms of the logical coherence and robustness to non-normality (e.g. Geary 1927; Eden and Yates 1933; Fisher 1935a). They found that the permutation distribution for "randomization models" mostly agree with the normality-based distribution for "population models", until 1937 when Li B. Welch disproved the agreement for Latin-squares designs (Welch 1937). In the next half century, most of the work on permutation tests were established for "randomization models" without being justified under "population models", except for rank-based tests which will be discussed later. We will skip the discussion for this period and refer to Berry et al. (2013) for a thorough literature review on this line of work, because our work focuses on the "population model" like (3.2).

The general theory of permutation tests in "population models" can be dated back to the notable works by Hoeffding (1952) and Box and Andersen (1955) and further developed by Romano (1989, 1990) and Chung and Romano (2013). In regression context, early investigations were done for special cases in ANOVA (Mehra and Sen 1969; Brown and Maritz 1982; Welch 1990). For testing a single regression coefficient, Oja (1987) and Collins (1987) proposed the permutation test on a linear statistic and the F-statistic by permuting the covariate while leaving the others the same. Whereas the procedure can be easily validated for univariate regression, the validity when $p > 1$ was only justified in "randomization models". Manly (1991) proposed permuting the response vector y, which is valid for testing the global null $\beta = 0$ but not for general case. Freedman and Lane (1983), Ter Braak (1992), and Kennedy and Cade (1996) proposed three different permutation tests on regression residuals. The theory of the aforementioned tests were established in a later review paper by Anderson

and Robinson (2001). The main take-away message being that the permutation test should be performed on asymptotically pivotal statistics. For instance, to test a single coefficient, the permutation t-test asymptotically valid. This was further confirmed and extended by DiCiccio and Romano (2017) for heteroskedastic linear models with random designs.

3.4.3 Rank-based tests

Perhaps a bit surprisingly, rank-based methods for linear regression can be dated back to 1936, when Hotelling and Pabst (1936) established the hypothesis testing theory for rank correlation, nowadays known as Spearman's correlation which was originated from Galton (1894) and developed by Spearman (1904) and Pearson (1907). This work can be regarded as the application of rank-based methods for univariate linear models. Appealed by the normality free nature of rank-based tests, Milton Friedman extended the idea to one-way ANOVA (Friedman 1937). It can be identified as the first application of rank-based method for multivariate linear models and was further developed by Kendall and Smith (1939) and Friedman (1940). Friedman's test transforms continuous or ordinal outcomes into ranks and were widely studied in ANOVA problems, started by the famous Kruskal-Wallis test for one-way ANOVA (Kruskal and Wallis 1952) and developed by Hodges and Lehmann (1962), Puri and Sen (1966), Sen (1968b), Conover and Iman (1976), Conover and Iman (1981), Akritas (1990), Akritas and Arnold (1994), Brunner and Denker (1994), and Akritas et al. (1997) for two-way ANOVA problems and factorial designs. Since 1990s, due to the advance of high dimensional asymptotic theory, further progress was made on refining the procedures in presence of large number of factors or treatments (Brownie and Boos 1994; Boos and Brownie 1995; Wang and Akritas 2004; Bathke and Lankowski 2005; Bathke and Harrar 2008).

However the aforementioned works are restricted to ANOVA problems (with a few exceptions, e.g. (Sen 1968a, 1969)) and fundamentally different from the modern rank tests based on regression R-estimates, which are based on ranks of regression residuals. The first R-estimate based test can be dated back to Hájek (1962), which established asymptotically most powerful rank test for univariate regression given the error distribution. Adichie (1967a) extended the idea to testing the intercept and the regression coefficient simultaneously. It was further extended to global testing for multivariate regression (Koul 1969). Rank-based tests for testing sub-hypotheses was first proposed by Koul (1970) and Puri and Sen (1973) for bivariate regression. The general theory of testing sub-hypotheses were independently developed by Srivastava (1972), McKean and Hettmansperger (1976) and Adichie (1978). The underlying theory is based on the pinoneering work by Jana Jurečková (Jureckova 1969), as a significant generalization of Hodges and Lehmann (1963) for location problems and Adichie (1967b) for univariate regression. Her work was further extended by Jureckova (1971) and Eeden (1972). However, these approaches are computationally extensive due to the discreteness of ranks. A one-step estimator was proposed by Kraft and Van Eeden (1972), which is asymptotically equivalent to the maximum likelihood estimators if the error distribution is known. Another one-step rank-based estimator, motivated by Bickel (1975) for

M-estimators, was proposed by McKean and Hettmansperger (1978). On the other hand, Jaeckel (1972) proposed a rank-based objective function, later known as Jaeckel's dispersion function, that is convex in β whose minimizer is asymptotically equivalent to Jurečková's score-based estimators. Hettmansperger and McKean (1978) found an equivalent but mathematically more tractable formulation of Jaeckel's dispersion function as the sum of pairwise difference of regression residuals. A weighted generalization of the dispersion function was introduced by Sievers (1983), which unified Jaeckel's dispersion function and Kendall's tau based dispersion function (Sen 1968a; Sievers 1978) . Three nice survey papers were written by Adichie (1984), Aubuchon and Hettmansperger (1984) and Draper (1988). In 1990s, due to the development of quantile regression (Koenker and Bassett 1978), Gutenbrunner and Jurecková (1992) found an important coincidence between the dual problem of quantile regression and the so-called "rank-score process", which generalizes the notion introduced by Hájek and Šidák (1967) to linear models. Gutenbrunner et al. (1993) then developed rank-score test for linear hypotheses; see also Koenker (1997) for a review. Over the past decade, there were much fewer works on rank-based tests for linear models (e.g. Feng et al. 2013).

3.4.4 Tests based on regression M-estimates

Regression M-estimates were introduced by Peter J. Huber in 1964 for location problems (Huber 1964). The idea was soon extended to linear models by Relles (1968), who proved the asymptotic theory for Huber's loss with p-fixed and n tending to infinity. The theory was extended to general convex loss functions by Yohai (1972). Despite the appealing statistical properties, the computation remained challenging in 1970s. Bickel (1975) proposed one-step M-estimates that are computational tractable with the same asymptotic property as full M-estimates. In addition, he proved the uniform asymptotic linearity of M-estimates, which was a fundamental theoretical result that laid the foundation for later works. Based on Bickel (1975)'s technique, Jureckova (1977) established the relation between regression M-estimates and R-estimates. The asymptotic normality of M-estimates directly yield an asymptotically valid Wald-type test for general linear hypotheses. Schrader and Hettmansperger (1980) developed an analogue of likelihood-ratio test based on M-estimators for sub-hypotheses. It was further extended to general linear hypotheses by Silvapulle (1992). However, both Wald-type tests and likelihood-ratio-type tests involves estimating nuisance parameters. To overcome the extra efforts, Sen (1982) proposed M-test as an analogue of studentized score test M-tests, which is capable to test general linear hypotheses with merely estimates of regression coefficients under the null hypothesis. It is known that Rao's score test may not be efficient in presence of nuisance parameter. Singer and Sen (1985) discussed an efficient test, which is essentially the analogue of Neyman's $C(\alpha)$ test based on projected scores (Neyman 1959), although it brings back nuisance parameters. M-tests were later investigated and generalized in a general framework based on influence functions (e.g. Boos 1992; Markatou and Ronchetti 1997).

Similar to t/F tests but unlike regression R-estimates, the robustness against high dimensionality was investigated extensively for M-estimators in general linear models. In Huber's 1972 Wald lectures (Huber 1972), he conjectured that the asymptotic normality of M-estimates proved by Relles (1968) can be extended to the asymptotic regime where p grows with n. The conjecture was proved one year later in the regime $\kappa p^2 = o(1)$, where κ is the maximum leverage score, which implies $p = o(n^{1/3})$ (Huber 1973b). This was improved to $\kappa p^{3/2} = o(1)$ by Yohai and Maronna (1979a), which implies that $p = o(n^{2/5})$, to $p = o(n^{2/3}/\log n)$ by Portnoy (1985) under further regularity conditions on the design matrix, and to $\kappa n^{1/3}(\log n)^{2/3} = o(1)$, which implies that $p = o(n^{2/3}/(\log n)^{2/3})$. All aforementioned results work for smooth loss functions. For non-smooth loss functions, Welsh (1989) obtained the first asymptotic result in the regime $p = o(n^{1/3}/(\log n)^{2/3})$. It was improved to $p = o(n^{1/2})$ by Bai and Wu (1994). For a single coordinate, Bai and Wu (1994) showed the asymptotic normality in the regime $p = o(n^{2/3})$. These works prove that the classical asymptotic theory holds if $p \ll n^{2/3}$. However, in the moderate dimensions where p grows linear with n, the M-estimates are no longer consistent in L_2 metric and the risk $\|\hat{\beta} - \beta\|_2^2$ tends to a non-vanishing quantity determined by p/n, the loss function and the error distribution through a complicated system of non-linear equations for random designs (El Karoui et al. 2011; Bean et al. 2012; El Karoui 2013; Donoho and Montanari 2016; El Karoui 2018). This surprising phenomenon marks the failure of classical asymptotic theory for M-estimators. For least-squares estimators, Lei et al. (2018) showed that the classical t-test with appropriate studentization is still asymptotically valid under regularity conditions on the design matrix. Cattaneo et al. (2018) proposed a refined test for heteroscedastic linear models. However it is unclear how to test general linear hypotheses with general M-estimators in this regime, even for a single coordinate. Lei et al. (2018) provides the only fixed-design result for the asymptotic property of a single coordinate for general M-estimates. For the purpose of hypothesis testing, the null variance should be estimated but there is no consistent variance estimator, except for special random designs (e.g. Bean et al. 2012).

3.4.5 Tests based on regression L-estimates

L-estimators constitute an important class of robust statistics based on linear combination of order statistics. Frederick Mosteller proposed the first L-estimator for Gaussian samples (Mosteller 1946). This was further developed by Hastings et al. (1947), Lloyd (1952), Evans and Evans (1955), Jung (1956), Tukey (1960), Bickel (1965), and Gastwirth (1966). In particular, John W. Tukey advocated the trimmed mean and Winsorized mean, which he attributed to Charles P. Winsor based on their personal communication in 1941, in his far-reaching paper (Tukey 1962). One year later, the well-known Hodges and Lehmann estimator was developed (Hodges and Lehmann 1963), which established the first connection between R-estimates and L-estimates. For location problems, Bickel and Lehmann (1975) found the superiority of L-estimates over M-estimates and R-estimates.

Despite the simplicity and nice theoretical property of L-statistics, they are not easy to be generalized to linear models. The first attempt was made by Bickel (1973), which

proposed a one-step L-estimate for general linear models. However, this estimator is not equivariant to affine transformation of design matrices. Motivated by this paper, Welsh (1987) proposed a class of one-step L-estimators that are equivariant to reparametrization of the design matrix. Welsh (1991) further extended the idea to construct an adaptive L-estimator. Another line of thoughts were motivated by the pinoneering work of Koenker and Bassett (1978), which introduced the notion of regression quantiles as a natural analogue of sample quantiles for linear models. Although quantile regression yields an M-estimator, it had been the driving force for the development of regression L-estimators since 1980s. In this paper, they proposed another class of L-estimators by discrete weighted average of regression quantiles and derived its asymptotic distribution. This idea was furthered by Koenker and Portnoy (1987) to L-estimators with continuous weights, by Portnoy and Koenker (1989) to adaptive L-estimators, and by Koenker and Zhao (1994) to heteroscedastic linear models. The other notable strategy of contructing L-statistics is based on weighted least squares with "outliers" removed. Ruppert and Carroll (1980) developed two equivariant one-step estimators as analogues of trimmed mean. Both estimators can be written in the form of weighted least squares where units with extreme residuals are removed and one is based on regression quantiles. As with Ruppert and Carroll (1980), Jureckova (1983) proposed an analogue of winsorized mean. The Bahadur representation of trimmed mean estimator was derived by Jurečková (1984). A nice review article of regression L-estimators was written by Alimoradi and Saleh (1998). The asymptotic results of L-estimators induce an asymptotically valid Wald-type test with a consistent estimate of asymptotic variance. Unlike M-estimators, we are not aware of other types of tests based on L-estimates.

3.4.6 Resampling based tests

Resampling, marked by Jackknife (Quenouille 1949, 1956; Tukey 1958) and bootstrap (Efron 1979), is a generic technique to assess the uncertainty of an estimator. Although both involving resampling, resampling-based tests are fundamentally different from permutation tests, as the former is approximating the sampling distribution under the truth while the latter is approximating the sampling distribution under the null hypothesis, although they are asymptotically equivalent in many cases (e.g. Romano 1989). Miller (1974) proposed the first Jackknife-based estimate for general liner models. He showed that the estimator is asymptotically normal and the Jackknife variance estimator is consistent, thereby the Wald-type test is asymptotically valid. Hinkley (1977) pointed out that Miller's estimator is less efficient than the least-squares estimator and proposed a weighted Jackknife estimates to overcome the inefficiency. Wu (1986) proposed a general class of delete-k jackknife estimators for estimating the covariance matrix of the least-squares estimator. This was extended by Shao and Wu (1987), Shao (1988, 1989), Peddada and Patwardhan (1992), and Liu and Singh (1992).

On the other hand, David A. Freedman first studied bootstrapping procedures for linear models (Freedman 1981). He proposed and studied two types of bootstrap: residual bootstrap, where the regression residuals are resampled and added back to the fitted values,

and the pair bootstrap, where the outcome and the covariates are resampled together. In the fixed-p regime, he showed the consistency of the residual bootstrap for homoscedastic linear models and consistency of pair bootstrap for general "correlation models" including heteroscedastic linear models. Navidi (1989), Hall (1989) and Qumsiyeh (1994) established the higher order accuracy of pair bootstrap for linear models and the results were then presented under a broader framework in the influential monograph of Peter Hall (Hall 1992). Wu (1986) found that the residual bootstrap fails in heteroscedastic linear models because its sampling process is essentially homoscedastic. To overcome this, he introduced another type of bootstrapping method based on random re-scaling of regression residuals that match the first and second moments. Liu (1988) introduced a further requirement to match the third moment and improved the rate of convergence. Later Mammen (1993) coined this procedure "wild bootstrap" and proved the consistency for linear least-squares estimator under random-design homoscedastic and heteroscedastic linear models. Hu and Zidek (1995) proposed an alternative bootstrap procedure for heteroscedastic linear models that resample the score function instead of the residuals. A wild bootstrap analogue of score-based bootstrap was proposed by Kline and Santos (2012). In particular, they developed the bootstrap Wald tests and score tests for general linear hypotheses.

The bootstrap techniques were also widely studied for regression M-estimates. The residual bootstrap was extended to M-estimators with smooth loss functions by Shorack (1982). Unlike least-squares estimators, it requires a debiasing step to obtain distributional consistency. Lahiri (1992) proposed a weighted residual bootstrap that does not require debiasing. He additionally showed the higher order accuracy of the weighted bootstrap and Shorack's bootstrap for studentized M-estimators. However, this weighted bootstrap ia hard to be implemented in general. On the other hand, motivates by Bayesian bootstrap (Rubin 1981), Rao and Zhao (1992) proposed a bootstrapping procedure by randomly reweighting the objective function. This idea was extended by Chatterjee (1999) in a broader framework called "generalized bootstrap". It was later re-invented by Jin et al. (2001) and referred to as "perturbation bootstrap". The higher order accuracy of perturbation bootstrap was established by Das and Lahiri (2019). It was pointed out by (Das and Lahiri 2019) that the perturbation bootstrap coincides with wild bootstrap in for least-squares estimators. Hu and Kalbfleisch (2000) proposed another estimating function based bootstrap, as essentially an resampling version of Sen (1982)'s M-tests. Wild bootstrap was introduced for quantile regression by Feng et al. (2011).

The robustness of bootstrap methods against high dimension was widely studied in literature. Bickel and Freedman (1983a) proved the distributional consistency of residual bootstrap least-squares estimators in the regime $p = o(n)$ in terms of the linear contrasts and in the regime $p = o(n^{1/2})$ in terms of the whole vector, for fixed-design linear models with vanishing maximum leverage score. They further the failure of bootstrap in moderate dimensions where $p/n \to c \in (0, 1)$ and the usual variance re-scaling does not help because the bootstrap distribution is no longer asymptotically normal. For M-estimators, Shorack (1982) showed that the debiased residual bootstrap is distributionally consistent in the regime $p = o(n^{1/3})$ in terms of the linear contrasts. The results were extended by Mammen (1989) to the regime

$p = o(n^{2/3}/(\log n)^{2/3})$ in terms of the linear contrasts and to the regime $p = o(n^{1/2})$ in terms of the whole vector. For random designs with i.i.d. design points, Mammen (1993) proved the distributional consistency of both pair bootstrap and wild bootstrap, in terms of linear contrasts, in the regime $p = o(n^a)$ for arbitrary $a < 1$. He also proved the consistency for heteroscedastic linear models in the regime $p = o(n^{3/4})$ for pair bootstrap and the regime $p = o(n^{1/2})$ for Wild bootstrap. The was further extended by Chatterjee (1999) to generalized bootstrap, including perturbation bootstrap (Rao and Zhao 1992), m-out-of-n bootstrap (Bickel and Sakov 2008) and delete-d jackknife (Wu 1990). On the other hand, extending Bickel and Freedman (1983a)'s negative result, El Karoui and Purdom (2018) showed the failure of various bootstrap procedures for M-estimators in moderate dimensions, including pair bootstrap, residual bootstrap, wild bootstrap and jackknife.

3.4.7 Other tests

A generic strategy for hypothesis testing is through pitoval statistics. Specifically, if there exists a statistics S of which the distribution is fully known, then the rejection set $S \in \mathcal{R}^c$ for any region $\mathcal{R}$ with $P(S \in \mathcal{R}) \geq 1 - \alpha$ gives a finite-sample valid test. For linear models, it is extremely hard to a pitoval statistics under general linear hypotheses, except for Gaussian linear models for which the t/F statistics are pivotal. However, if the goal is to test all coefficients plus the intercept, i.e. $H_0 : \beta_0 = \gamma_0, \beta = \gamma$, then one can recover the stochastic errors as $\epsilon_i = y_i - \gamma_0 - x_i^T \gamma$ under the null and construct pivotal statistics based on ϵ. Taking one step further, given a pitoval statistic, one can invert the above test to obtain a finite-sample valid confidence region $\mathcal{C}$ for (β_0, β), by collecting all (γ_0, γ^*)'s to which the corresponding null hypothesis fails to be rejected. This induces a confidence region for $R\beta$ as $\mathcal{C}' = \{R\beta : (\beta_0, \beta) \in \mathcal{C}\}$. Finally, using the duality between confidence interval and hypothesis testing again, the test which rejects the null hypothesis is finite-sample valid for the linear hypothesis $H_0 : R\beta = 0$. If $r \ll p$, this seemingly "omnibus test" is in general powerless and inferior to the tests discussed in previous subsections. Nonetheless, it stimulates several non-standard but interesting tests in history that are worth discussion.

The most popular strategy to construct pivotal statistics is based on quantiles of ϵ_i's, especially the median. Assuming ϵ_i's have zero median, Fisher (1925) first introduced the sign test for location problems, which was investigated and formalized later by Cochran (1937). Thirteen years later, Henri Theil proposed an estimator for univariate linear models (Theil 1950a,b,c), later known as Theil-Sen estimator (Sen 1968a). Brown and Mood (1951) proposed the median test for general linear models by reducing the problem into a contingency table and applying the χ^2-tests. The theoretical property of Brown-Mood test was studied by Kildea (1981) and Johnstone and Velleman (1985). Daniels (1954) proposed a geometry-based test for univariate linear models, which can be regarded as a generalization of Brown-Mood test. It was later connected to the notion of regression depth (Rousseeuw and Hubert 1999) and applied in deepest regression methods (Van Aelst et al. 2002). The idea of inverting the sign test was exploited in Quade (1979) and an analogue incorporating Kendall's tau between the residuals and the covariates was proposed by Lancaster and

Quade (1985). The idea also attracted some attention in engineering literature (e.g. Campi and Weyer 2005; Campi et al. 2009) and in econometrics literature (e.g. Chernozhukov et al. 2009). It should be noted that the approach is computationally infeasible even for moderately large dimensions. Assuming further the symmetry of ϵ_i's, Hartigan (1970) proposed a non-standard test based on an interesting notion of typical values. It was designed for location problems but can be applied to certain ANOVA problems. Furthermore, Siegel (1982) proposed the repeated median estimator and Rousseeuw (1984) proposed the least median squares estimators to achieve high breakdown point.

The pivotal statistics can also be constructed in other ways. Parzen et al. (1994) proposed a bootstrap procedure based on inverting a pivotal estimating function at a random point. This procedure mimics the Fisher's fiducial inference but can be justified under common framework. Recently Meinshausen (2015) proposed the Group Bound test for subhypotheses, which even works for high-dimensional settings where $p \gg n$. However, the validity is only guaranteed for rotationally invariant errors with known noise level. This requirement is extremely strong as shown by Maxwell (1860): if ϵ_i's are further assumed to be i.i.d., then rotation invariance implies the normality of ϵ_i's.

3.5 Conclusion and Discussion

In this article, we propose Cyclic Permutation Test (CPT) for testing general linear hypotheses for linear models. This test is exact for arbitrary fixed design matrix and arbitrary exchangeable errors, whenever $1/\alpha$ is an integer and $n/(p-r) \geq 1/\alpha - 1$. Extensive simulation studies demonstrates the reasonable performance of CPT.

CPT is non-standard compared to various methodologies developed in the past century. CPT essentially constructs a pivotal statistic in finite samples based on group invariance. This is rare in the territory of distribution-free inference with complex nuisance parameters. Our goal is to expand the toolbox for exact and distribution-free inference and hopefully generate new ideas for more complicated problems. In the following subsections we discuss several extensions and future directions.

3.5.1 Confidence interval/region by inverting CPT

It is straightforward to deduce a confidence band for $\beta_{[r]}$ can be obtained by inverting CPT. Specifically, the inverted confidence band is given by

$$\mathcal{I} \triangleq \left\{ \beta_{[r]} : p(y - X\beta; X) > \alpha \right\},$$

where $p(y; X)$ is the p-value produced by CPT with a design matrix X and an outcome vector y. Under the construction **C**3,

$$\eta_j^T (y - X\beta) = \eta_j^T y - \gamma^T \beta - \delta^T \beta_{[r]} I(j = 0).$$

Thus,

$$\text{med}\left(\{\eta_j^T(y - X\beta)\}_{j=0}^m\right) = \text{med}\left(\{\eta_j^T y - \delta^T\beta_{[r]} I(j=0)\}_{j=0}^m\right) - \gamma^T\beta.$$

Then $\mathcal{I}$ can be simplified as

$$\mathcal{I} = \left\{\beta_{[r]} : \delta^T\beta_{[r]} \in [x_{\min}, x_{\max}]\right\} \tag{3.24}$$

where $x_{\min}$ and $x_{\max}$ are the infimum and the superimum of x such that

$$\frac{1}{m+1}\left(1 + \sum_{j=1}^m I\Big(|\eta_0^T y - x - m(x)| \geq |\eta_j^T y - m(x)|\Big)\right) > \alpha,$$

and

$$m(x) = \text{med}\left(\{\eta_j^T y - xI(j=0)\}_{j=0}^m\right).$$

When $r = 1$, the confidence interval (3.24) gives a useful confidence interval simply as

$$\mathcal{I} = [x_{\min}/\delta, x_{\max}/\delta].$$

However when $r > 1$, the confidence region (3.24) may not be useful because it is unbounded. More precisely, $\beta_{[r]} \in \mathcal{I}$ implies that $\beta_{[r]} + \xi \in \mathcal{I}$ for any ξ orthogonal to δ. We leave the construction of more efficient confidence regions to future research.

3.5.2 Connection to knockoff based inference

Our test is implicitly connected to the novel idea of knockoffs, proposed by Barber, Candès, et al. (2015) to control false discovery rate (FDR) for variable selection in linear models. Specifically, hey assumed a Gaussian linear model and aimed at detecting a subset of variables that control FDR in finite samples. Unlike the single hypothesis testing considered in this chapter, multiple inference requires to deal with the dependence between test statistics for each hypothesis carefully. They proposed an interesting idea of constructing a pseudo design matrix $\tilde{X}$ such that the joint distribution of $(X_1^T y, \ldots, X_p^T y, \tilde{X}_1^T y, \ldots, \tilde{X}_p^T y)$ is invariant to the pairwise swapping of $X_j^T y$ and $\tilde{X}_j^T y$ all for j with $\beta_j = 0$. Then the test statistic for testing $H_{0j} : \beta_j = 0$ is constructed by comparing $X_j^T y$ and $\tilde{X}_j^T y$ in an appropriate way, thereby obtaining a valid binary p-value p_j that is uniformly distributed on $\{1/2, 1\}$ under H_{0j}. The Knockoffs-induced p-values marginally resemble the construction of statistics in CPT with $m = 2, \eta_0 = X_j, \eta_1 = \tilde{X}_j$. On the other hand, the validity of knockoffs essentially rests on the distributional invariance of ϵ to the rotation group while the validity of CPT relies on the distributional invariance of ϵ to the cyclic permutation group. This coincidence illustrates the charm and the magical power of group invariance in statistical inference.

3.5.3 More efficient algorithm for pre-ordering

Although GA can solve the combinatorial optimization problem efficiently for problems with moderate size, it is not scalable enough to handle big data. Since the exact minimizer is not required, we can resort to other heuristic algorithms. One heuristic strategy is proposed by Fogel et al. (2013) by relaxing permutation matrice into doubly stochastic matrices, with $\Pi\mathbf{1} = \Pi^T\mathbf{1} = 0$ and $\Pi_{ij} \geq 0$, and optimize the objective using continuous optimization algorithms. Taking the case of $r = 1$ for example, by Theorem 3.2.6, (3.23) is equivalent to

$$\min_{\Pi} B(\Pi X)_1^T (I - B(\Pi X)_{[-1]} (B(\Pi X)_{[-1]}^T B(\Pi X)_{[-1]})^+ B(\Pi X)_{[-1]}^T) B(\Pi X)_1.$$

By Sherman-Morrison-Woodbury formula, the reciprocal of the above objective is the first diagonal element of $H(\Pi)$. Therefore, (3.23) is equivalent to

$$\max_{\Pi} e_1^T [B(\Pi X)^T B(\Pi X)]^{-1} e_1.$$

Denote by $h(\Pi)$ the above objective function and $H(\Pi)$ by $[B(\Pi X)^T B(\Pi X)]^{-1}$, then the derivative of h with respect to Π can be easily calculated as

$$\begin{aligned}
\frac{\partial h(\Pi)}{\partial \Pi_{ij}} &= -e_1^T H(\Pi) \left(\frac{\partial}{\partial \Pi_{ij}} [B(\Pi X)^T B(\Pi X)] \right) H(\Pi) e_1 \\
&= - e_1^T H(\Pi) \left(B(\Pi X)^T \frac{\partial}{\partial \Pi_{ij}} B(\Pi X) + \left(\frac{\partial}{\partial \Pi_{ij}} B(\Pi X) \right)^T B(\Pi X) \right) H(\Pi) e_1 \\
&= - e_1^T H(\Pi) \Big(B(\Pi X)^T B(e_i X_j^T) + B(e_i X_j^T)^T B(\Pi X) \Big) H(\Pi) e_1,
\end{aligned}$$

where the last line uses the definition of $B(X)$ in (3.19). The easy gradient computation may suggest an efficient gradient based algorithm. We leave this as a future direction.

Chapter 4

Regression Adjustment for Neyman-Rubin Models

4.1 Introduction

4.1.1 Potential outcomes and Neyman's randomization model

We use potential outcomes to define causal effects (Neyman 1923/1990). Let $Y_i(1)$ and $Y_i(0)$ be the potential outcomes if unit $i \in \{1, \ldots, n\}$ receives the treatment and control, respectively. Neyman (1923/1990) treated all the potential outcomes as fixed quantities, and defined the average treatment effect (ATE) as $\tau \equiv n^{-1} \sum_{i=1}^{n} \tau_i$, where $\tau_i = Y_i(1) - Y_i(0)$ is the individual treatment effect for unit i. In a completely randomized experiment, the experimenter randomly assigns n_1 units to the treatment group and n_0 units to the control group, with $n = n_1 + n_0$. Let T_i denote the assignment of the i-th unit where $T_i = 1$ corresponds to the treatment and $T_i = 0$ corresponds to the control. For unit i, only $Y_i^{\text{obs}} = Y_i(T_i)$ is observed while the other potential outcome $Y_i(1 - T_i)$ is missing. Although $(Y_i(1), Y_i(0))_{i=1}^n$ are fixed, the Y_i^{obs}'s are random due to the randomization of the T_i's.

Scheffé (1959, Chapter 9) called the above formulation the *randomization model*, under which all potential outcomes are fixed and the randomness comes solely from the treatment indicators. This finite-population perspective has a long history for analyzing randomized experiments (e.g. Neyman 1923/1990, 1935; Kempthorne 1952; Imbens and Rubin 2015; Mukerjee et al. 2018; Fogarty 2018; Middleton 2018). In contrast, the super-population perspective (e.g. Tsiatis et al. 2008; Berk et al. 2013; Pitkin et al. 2017) assumes that the potential outcomes and other individual characteristics are independent and identically distributed (i.i.d.) draws from some distribution. Two perspectives are both popular in the literature, but they are different in the source of randomness: the finite-population perspective quantifies the uncertainty of the sampling procedure in a single "universe" of units; in contrast, the super-population perspective also considers the uncertainty across multiple, possibly infinite, "universes" of units.

We use the conventional notation $O(\cdot), o(\cdot)$, $O_{\mathbb{P}}(\cdot)$ and $o_{\mathbb{P}}(\cdot)$. Let $\mathbf{1}$ denote the vector with

all entries 1, I denote an identity matrix, and $\mathrm{V} = \mathrm{I} - (\mathbf{1}^\mathsf{T}\mathbf{1})^{-1}\mathbf{1}\mathbf{1}^\mathsf{T}$ denote the projection matrix orthogonal to $\mathbf{1}$, with appropriate dimensions depending on the context. Let $\|\cdot\|_q$ be the vector q-norm, i.e. $\|\alpha\|_q = (\sum_{i=1}^n |\alpha_i|^q)^{1/q}$ and $\|\alpha\|_\infty = \max_{1\le i\le n} |\alpha_i|$. Let $\|\cdot\|_{\mathrm{op}}$ denote operator norm and $\|\cdot\|_F$ denote the Frobenius norm of matrices. Let $N(0,1)$ denote the standard normal distribution, and $t(\nu)$ denote standard t distribution with degrees of freedom ν with $t(1)$ being the standard Cauchy distribution. Let $\xrightarrow{\mathrm{d}}$ and $\xrightarrow{\mathbb{P}}$ denote convergences in distribution and in probability.

4.1.2 Regression-adjusted average treatment effect estimates

Let $\mathcal{T}_t = \{i : T_i = t\}$ be the indices and $n_t = |\mathcal{T}_t|$ be the fixed sample size for treatment arm $t \in \{0,1\}$. We consider a completely randomized experiment in which $\mathcal{T}_1$ is a random size-n_1 subset of $\{1,\dots,n\}$ uniformly over all $\binom{n}{n_1}$ subsets. The simple difference-in-means estimator

$$\hat{\tau}_{\mathrm{unadj}} = \frac{1}{n_1}\sum_{i\in\mathcal{T}_1} Y_i^{\mathrm{obs}} - \frac{1}{n_0}\sum_{i\in\mathcal{T}_0} Y_i^{\mathrm{obs}} = \frac{1}{n_1}\sum_{i\in\mathcal{T}_1} Y_i(1) - \frac{1}{n_0}\sum_{i\in\mathcal{T}_0} Y_i(0)$$

is unbiased with variance $S_1^2/n_1 + S_0^2/n_0 - S_\tau^2/n$ (Neyman 1923/1990), where S_1^2, S_0^2 and S_τ^2 are the finite-population variances of the $Y_i(1)$'s, $Y_i(0)$'s and τ_i's, respectively.

The experimenter usually collects pre-treatment covariates. If the covariates are predictive of the potential outcomes, it is intuitive to incorporate them in the analysis to improve the estimation efficiency. Suppose unit i has a p-dimensional vector of pre-treatment covariates $x_i \in \mathbb{R}^p$. Early works on the analysis of covariance assumed constant treatment effects (Fisher 1935b; Kempthorne 1952; Hinkelmann and Kempthorne 2007), under which a commonly-used treatment effect estimate is the coefficient of the treatment indicator of the ordinary least squares (OLS) fit of the Y_i^{obs}'s on T_i's and x_i's. Freedman (2008b) criticized this standard approach, showing that (a) it can be even less efficient than $\hat{\tau}_{\mathrm{unadj}}$ in the presence of treatment effect heterogeneity, and (b) the estimated standard error based on the OLS can be inconsistent for the true standard error under the randomization model.

Lin (2013) proposes a simple solution. Without loss of generality, we center the covariates at $n^{-1}\sum_{i=1}^n x_i = 0$ because otherwise we can replace x_i by $x_i - n^{-1}\sum_{i=1}^n x_i$. His estimator for the ATE is the coefficient of the treatment indicator in the OLS fit of the Y_i^{obs}'s on T_i's, x_i's and the *interaction terms* $T_i x_i$'s. He further shows that the Eicker–Huber–White standard error (e.g. MacKinnon 2013) is consistent for the true standard error. Lin (2013)'s results hold under the finite-population randomization model, without assuming that the linear model is correct.

We use an alternative formulation of regression adjustment and consider the following family of covariate-adjusted ATE estimator:

$$\hat{\tau}(\beta_1,\beta_0) = \frac{1}{n_1}\sum_{i\in\mathcal{T}_1}(Y_i^{\mathrm{obs}} - x_i^\mathsf{T}\beta_1) - \frac{1}{n_0}\sum_{i\in\mathcal{T}_0}(Y_i^{\mathrm{obs}} - x_i^\mathsf{T}\beta_0). \tag{4.1}$$

Because $\mathbb{E}\left(n_t^{-1}\sum_{i\in\mathcal{T}_t} x_i^\mathsf{T}\beta_t\right) = 0$, the estimator in (4.1) is unbiased for any fixed coefficient vectors $\beta_t \in \mathbb{R}^p$ $(t = 0, 1)$. It is the difference-in-means estimator with potential outcomes replaced by $(Y_i(1) - x_i^\mathsf{T}\beta_1, Y_i(0) - x_i^\mathsf{T}\beta_0)_{i=1}^n$.

Let $Y(t) = (Y_1(t), \ldots, Y_n(t))^\mathsf{T} \in \mathbb{R}^n$ denote the vector of potential outcomes under treatment t $(t = 0, 1)$, $X = (x_1, \ldots, x_n)^\mathsf{T}$ denote the matrix of covariates. Without loss of generality, we assume

$$\mathbf{1}^\mathsf{T}X = 0 \text{ and } \operatorname{rank}(X) = p, \tag{4.2}$$

i.e., the covariate matrix has centered columns and full column rank. Otherwise, we transform X to $\mathrm{V}X$ and remove the redundant columns to ensure the full column rank condition. Let β_t be the population OLS coefficient of regressing $Y(t)$ on X with an intercept:

$$\begin{aligned}(\mu_t, \beta_t) &= \underset{\mu\in\mathbb{R},\beta\in\mathbb{R}^p}{\arg\min} \|Y(t) - \mu\mathbf{1} - X\beta\|_2^2 &(4.3)\\ &= \left(\frac{1}{n}\sum_{i=1}^n Y_i(t), (X^\mathsf{T}X)^{-1}X^\mathsf{T}Y(t)\right), &(4.4)\end{aligned}$$

where (4.4) holds because X is orthogonal to $\mathbf{1}$. Li and Ding (2017, Example 9) show that the OLS coefficients (β_1, β_0) in (4.3) minimize the variance of the estimator defined in (4.1).

The classical analysis of covariance chooses $\beta_1 = \beta_0 = \hat{\beta}$, the coefficient of the covariates in the OLS fit of the Y_i^{obs}'s on T_i's and x_i's with an intercept. This strategy implicitly assumes away treatment effect heterogeneity, and can lead to inferior properties when $\beta_1 \neq \beta_0$ (Freedman 2008b). Lin (2013) chooses $\beta_1 = \hat{\beta}_1$ and $\beta_0 = \hat{\beta}_0$, the coefficients of the covariates in the OLS fit of Y_i^{obs}'s on x_i's with an intercept, in the treatment and control groups, respectively. Numerically, this is identical to the estimator obtained from the regression with interactions discussed before.

4.1.3 Our contributions

In practice, it is common to have many covariates. Therefore, it is important to approximate the sampling distribution with p growing with the sample size n at certain rate. Under the finite-population randomization model, Bloniarz et al. (2016) discussed a high dimensional regime with possibly larger p than n but assumed that the potential outcomes could be well approximated by a sparse linear combination of the covariates, under the ultra sparse regime (termed, for example, by Cai and Guo (2017)) where the number of non-zero coefficients is many fewer than $n^{1/2}/\log p$. Under a super-population framework, Wager et al. (2016) discussed covariate adjustment using the OLS and some other machine learning techniques.

We study Lin (2013)'s estimator under the finite-population perspective in the regime where $p < n$ but p grows with n at certain rate. We focus on this estimator because (a) it is widely used in practice because of its simplicity, and (b) it does not require any tuning parameter unlike other high dimensional or machine learning methods. As in the classic linear regression, asymptotic properties depend crucially on the maximum leverage score $\kappa = \max_{1\le i\le n} H_{ii}$, where the i-the leverage score H_{ii} is i-th diagonal entry of the hat matrix

$H = X(X^\mathsf{T}X)^{-1}X^\mathsf{T}$ (Huber 1973a). Under the regime $\kappa \log p \to 0$, we prove the consistency of Lin (2013)'s estimator under mild moment conditions on the population OLS residuals. In the favorable case where all leverage scores are close to their average p/n, the consistency holds if $p = o(n/\log n)$.

In addition, we prove that Lin (2013)'s estimator is asymptotically normal under $\kappa p \to 0$ and extra mild conditions, with the same variance formula as the fixed-p regime. Furthermore, we proposed a debiased estimator, which is asymptotically normal under an even weaker assumption $\kappa^2 p \log p \to 0$, with the same variance as before. In the favorable case where all leverage scores are close to their average p/n, Lin (2013)'s estimator is asymptotically normal when $p = o(n^{1/2})$, but the debiased estimator is asymptotically normal when $p = o(n^{2/3}/(\log n)^{1/3})$. Lin (2013)'s estimator may also be asymptotically normal in the latter regime, but it requires an extra condition (See Theorem 4.3.6). In our simulation, the debiased estimator indeed yields better finite-sample inferences.

For statistical inference, we propose several asymptotically conservative variance estimators, which yield valid asymptotic Wald-type confidence intervals for the ATE. We prove the results under the same regime $\kappa \log p \to 0$ with the same conditions as required for the asymptotic normality.

Importantly, our theory does not require any modeling assumptions on the *fixed* potential outcomes and the covariates. It is nonparametric.

We prove novel vector and matrix concentration inequalities for sampling without replacement. These tools are particularly useful for finite population causal inference, and can also complement and potentially enrich the theory in other areas such as survey sampling (e.g., Cochran 2007), matrix sketching (e.g., Woodruff 2014) and transductive learning (e.g., El-Yaniv and Pechyony 2009).

4.2 Regression Adjustment

4.2.1 Point Estimators

We reformulate Lin (2013)'s estimator. The ATE is the difference between the two intercepts of the population OLS coefficients in (4.4):

$$\tau = \frac{1}{n}\sum_{i=1}^{n} Y_i(1) - \frac{1}{n}\sum_{i=1}^{n} Y_i(0) = \mu_1 - \mu_0.$$

Therefore, we focus on estimating μ_1 and μ_0. Let $X_t \in \mathbb{R}^{n_t \times p}$ denote the sub-matrix formed by the rows of X, and $Y_t^{\text{obs}} \in \mathbb{R}^{n_t}$ the subvector of $Y^{\text{obs}} = (Y_1^{\text{obs}}, \ldots, Y_n^{\text{obs}})^\mathsf{T}$, with indices in $\mathcal{T}_t$ $(t = 0, 1)$. The regression-adjusted estimator follows two steps. First, for $t \in \{0, 1\}$, we regress Y_t^{obs} on X_t with an intercept, and obtain the fitted intercept $\hat{\mu}_t \in \mathbb{R}$ and coefficient of the covariate $\hat{\beta}_t \in \mathbb{R}^p$. Second, we estimate τ by

$$\hat{\tau}_{\text{adj}} = \hat{\mu}_1 - \hat{\mu}_0. \tag{4.5}$$

In general, $\hat{\tau}_{\text{adj}}$ is biased in finite samples. Correcting the bias gives stronger theoretical guarantees as our later asymptotic analysis suggests. Here we propose a bias-corrected estimator. Define the potential residuals based on the population OLS as

$$e(t) = Y(t) - \mu_t - X\beta_t, \quad (t = 0, 1). \tag{4.6}$$

The property of the OLS guarantees that $e(t)$ is orthogonal to $\mathbf{1}$ and X:

$$\mathbf{1}^{\mathsf{T}} e(t) = 0, \quad X^{\mathsf{T}} e(t) = 0, \quad (t = 0, 1). \tag{4.7}$$

Let $\hat{e} \in \mathbb{R}^n$ be the vector residuals from the sample OLS:

$$\hat{e}_i = \begin{cases} Y_i^{\text{obs}} - \hat{\mu}_1 - x_i^{\mathsf{T}}\hat{\beta}_1, & (i \in \mathcal{T}_1), \\ Y_i^{\text{obs}} - \hat{\mu}_0 - x_i^{\mathsf{T}}\hat{\beta}_0, & (i \in \mathcal{T}_0). \end{cases} \tag{4.8}$$

For any vector $\alpha \in \mathbb{R}^n$, let α_t denote the subvector of α with indices in $\mathcal{T}_t$ (e.g. $Y_t(1), e_t(1), \hat{e}_t$, etc.)

Let $H = X(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}$ be the hat matrix of X, and $H_t = X_t(X_t^{\mathsf{T}}X_t)^{-1}X_t^{\mathsf{T}}$ be the hat matrix of X_t. Let H_{ii} be the i-th diagonal element of H, also termed as the *leverage score*. Define

$$\Delta_t = \frac{1}{n}\sum_{i=1}^{n} e_i(t)H_{ii}, \quad \hat{\Delta}_t = \frac{1}{n_t}\sum_{i\in\mathcal{T}_t} \hat{e}_i H_{ii}. \tag{4.9}$$

We introduce the following debiased estimator:

$$\hat{\tau}_{\text{adj}}^{\text{de}} = \hat{\tau}_{\text{adj}} - \left(\frac{n_1}{n_0}\hat{\Delta}_0 - \frac{n_0}{n_1}\hat{\Delta}_1\right). \tag{4.10}$$

The bias correction terms in (4.10) come from higher order asymptotic expansions. When $p = 1$, (4.10) reduces the bias formula in Lin (2013, Section 6 point (iv)). Thus (4.10) is an extension to the multivariate case.

4.2.2 Variance estimators

For fixed p, Lin (2013) proved that $n^{1/2}(\hat{\tau}_{\text{adj}} - \tau)$ is asymptotically normal with variance

$$\sigma_n^2 = \frac{1}{n_1}\sum_{i=1}^{n} e_i^2(1) + \frac{1}{n_0}\sum_{i=1}^{n} e_i^2(0) - \frac{1}{n}\sum_{i=1}^{n}(e_i(1) - e_i(0))^2 \tag{4.11}$$

$$= \sum_{i=1}^{n}\left(\sqrt{\frac{n_0}{n_1 n}}e_i(1) + \sqrt{\frac{n_1}{n_0 n}}e_i(0)\right)^2. \tag{4.12}$$

The second form (4.12) follows from some simple algebra and shows that σ_n^2 is always non-negative. The first form (4.11) motivates conservative variance estimators. The third term in

(4.11) has no consistent estimator without further assumptions on $e(1)$ and $e(0)$. Ignoring it and estimating the first two terms in (4.11) by their sample analogues, we have the following variance estimator:

$$\hat{\sigma}^2 = \frac{n}{n_1(n_1-1)} \sum_{i \in \mathcal{T}_1} \hat{e}_i^2 + \frac{n}{n_0(n_0-1)} \sum_{i \in \mathcal{T}_0} \hat{e}_i^2. \tag{4.13}$$

Although (4.13) appears to be conservative due to the neglect of the third term in (4.12), we find in numerical experiments that it typically underestimates σ_n^2 in the cases beyond our theoretic limit with many covariates or many influential observations. The classic linear regression literature suggests rescaling the residual as

$$\tilde{e}_i = \begin{cases} \hat{e}_i & \text{(HC0)} \\ \sqrt{\frac{n-1}{n-p}}\hat{e}_i & \text{(HC1 correction)} \\ \frac{\hat{e}_i}{\sqrt{1-H_{t,ii}}} & \text{(HC2 correction)} \\ \frac{\hat{e}_i}{1-H_{t,ii}} & \text{(HC3 correction)} \end{cases}, \quad (i \in \mathcal{T}_t) \tag{4.14}$$

where $H_{t,ii}$ is the diagonal element of H_t corresponding to unit i. HC0 corresponds to the estimator (4.13) without corrections. Previous literature has shown that the above corrections, especially HC3, are effective in improving the finite sample performance of variance estimator in linear regression under independent super-population sampling (e.g. MacKinnon 2013; Cattaneo et al. 2018). More interestingly, it is also beneficial to borrow these HCj's to the context of a completely randomized experiment. This motivates the following variance estimators

$$\hat{\sigma}^2_{\text{HC}j} = \frac{n}{n_1(n_1-1)} \sum_{i \in \mathcal{T}_1} \tilde{e}_{i,j}^2 + \frac{n}{n_0(n_0-1)} \sum_{i \in \mathcal{T}_0} \tilde{e}_{i,j}^2 \tag{4.15}$$

where $\tilde{e}_{i,j}$ is the residual in (4.14) with j corresponding to HCj for $j = 0, 1, 2, 3$.

Based on normal approximations, we can construct Wald-type confidence intervals for the ATE based on point estimators $\hat{\tau}_{\text{adj}}$ and $\hat{\tau}^{\text{de}}_{\text{adj}}$ with estimated standard errors $\hat{\sigma}_{\text{HC}j}$.

4.3 Main Results

4.3.1 Regularity conditions

We embed the finite population quantities $\{(x_i, Y_i(1), Y_i(0))\}_{i=1}^n$ into a sequence, and impose regularity conditions on this sequence. The first condition is on the sample sizes.

Assumption 1. *$n/n_1 = O(1)$ and $n/n_0 = O(1)$.*

Assumption 1 holds automatically if treatment and control groups have fixed proportions (e.g., $n_1/n = n_0/n = 1/2$ for balanced experiments). It is not essential and can be removed at the cost of complicating the statements.

The second condition is on $\kappa = \max_{1\le i\le n} H_{ii}$, the maximum leverage score, which also plays a crucial role in classic linear models (e.g. Huber 1973a; Mammen 1989; Donoho and Huo 2001).

Assumption 2. $\kappa \log p = o(1)$.

The maximum leverage score satisfies

$$p/n = \operatorname{tr}(H)/n \le \kappa \le \|H\|_{\mathrm{op}} = 1 \Longrightarrow \kappa \in [p/n, 1]. \tag{4.16}$$

Assumption 2 permits influential observations as long as $\kappa = o(1/\log p)$. In the favorable case where $\kappa = O(p/n)$, it reduces to $p \log p/n \to 0$, which permits p to grow almost linearly with n. Moreover, it implies

$$\frac{p}{n} \le \kappa = o\left(\frac{1}{\log p}\right) = o(1) \Longrightarrow p = o(n). \tag{4.17}$$

Assumptions 1 and 2 are useful for establishing consistency. The following two extra conditions are useful for variance estimation and asymptotic normality. The third condition is on the correlation between the potential residuals from the population OLS in (4.6).

Assumption 3. *There exist a constant $\eta > 0$ independent of n such that*

$$\rho_e \triangleq \frac{e(1)^{\mathsf{T}} e(0)}{\|e(1)\|_2 \|e(0)\|_2} > -1 + \eta.$$

Assumption 3 is mild because it is unlikely to have the perfect negative sample correlation between the treatment and control residual potential outcomes in practice.

The fourth condition is on the following two measures of the potential residuals based on the population OLS in (4.6).

$$\mathcal{E}_2 = n^{-1} \max\left\{\|e(0)\|_2^2, \|e(1)\|_2^2\right\}, \quad \mathcal{E}_\infty = \max\left\{\|e(0)\|_\infty, \|e(1)\|_\infty\right\}.$$

Assumption 4. $\mathcal{E}_\infty^2/(n\mathcal{E}_2) = o(1)$.

Assumption 4 is a Lindeberg–Feller-type condition requiring that no single residual dominates others. A similar form appeared in Hájek (1960)'s finite population central limit theorem. Previous works require more stringent assumptions on the fourth moment (Lin 2013; Bloniarz et al. 2016).

4.3.2 Discussion of regularity conditions

Although the above assumptions are about fixed quantities in the finite population, it is helpful to consider the case where the quantities are realizations of random variables. This approach connects the assumptions to more comprehensible conditions on the data generating process. See Portnoy (1984, 1985) and Lei et al. (2016) for examples in other contexts.

We emphasize that we do not need the assumptions in this subsection for our main theory but use them to aid interpretation. The readers who believe our assumptions to be mild can skip this subsection at first read.

For Assumption 2, we consider the case where $(x_i)_{i=1}^n$ are realizations of i.i.d. random vectors. Anatolyev and Yaskov (2017) show that under mild conditions each leverage score concentrates around p/n. Here we further consider the magnitude of the maximum leverage score κ.

Proposition 4.3.1. *Let Z_i be i.i.d. random vectors in $\mathbb{R}^p$ with arbitrary mean. Assume that Z_i has independent entries with $\max_{1\le j\le p}\mathbb{E}|Z_{ij}-\mathbb{E}Z_{ij}|^\delta \le M = O(1)$ for some $\delta > 2$. Define $Z = (Z_1^\intercal, \dots, Z_n^\intercal)^\intercal \in \mathbb{R}^{n\times p}$ and $X = \mathrm{V}Z$ so that X has centered columns. If $p = O(n^\gamma)$ for some $\gamma < 1$, then over the randomness of Z,*

$$\max_{1\le i\le n}\left|H_{ii} - \frac{p}{n}\right| = O_{\mathbb{P}}\left(\frac{p^{2/\min\{\delta,4\}}}{n^{(\delta-2)/\delta}} + \frac{p^{3/2}}{n^{3/2}}\right), \quad \kappa = O_{\mathbb{P}}\left(\frac{p}{n} + \frac{p^{2/\min\{\delta,4\}}}{n^{(\delta-2)/\delta}}\right).$$

When $\delta > 4$, Proposition 4.3.1 implies that $\kappa = O_{\mathbb{P}}(p/n + n^{-(\delta-4)/2\delta}(p/n)^{1/2})$. In this case, Assumption 2 holds with high probability if $p = O(n^\gamma)$ for any $\gamma < 1$. In particular, the fixed-p regime corresponds to $\gamma = 0$.

The hat matrix of X is invariant to any nonsingular linear transformation of the columns. Consequently, X and XA have the same leverage scores for any invertible $A \in \mathbb{R}^{n\times n}$. Thus we can extend Proposition 4.3.1 to random matrices with correlated columns in the form of $\mathrm{V}ZA$. In particular, when $Z_i \overset{\text{i.i.d.}}{\sim} N(\mu, \mathrm{I})$ and $A = \Sigma^{1/2}$, $Z_i^\intercal A \overset{\text{i.i.d.}}{\sim} N(\Sigma^{1/2}\mu, \Sigma)$. The previous argument implies that Proposition 4.3.1 holds for $X = \mathrm{V}ZA$. We will revisit Proposition 4.3.1 when imposing further conditions on the H_{ii}'s and κ.

For Assumption 4, we consider the case where the $Y_i(t)$'s are realizations of i.i.d. random variables, and make a connection with the usual moment conditions. This helps to understand the growth rates of $\mathcal{E}_2$ and $\mathcal{E}_\infty$.

Proposition 4.3.2. *Let $Y(t) \in \mathbb{R}^n$ be a non-constant random vector with i.i.d. entries, and X be any fixed matrix with centered columns. If for some $\delta > 0$, $\mathbb{E}|Y_i(t) - \mathbb{E}Y_i(t)|^\delta < \infty$ for $t = 0, 1$, then*

$$\mathcal{E}_2 = \begin{cases} O_{\mathbb{P}}(1) & (\delta \ge 2) \\ o_{\mathbb{P}}(n^{2/\delta-1}) & (\delta < 2) \end{cases}, \quad \mathcal{E}_\infty = O_{\mathbb{P}}(n^{1/\delta}).$$

Furthermore, $\mathcal{E}_2^{-1} = O_{\mathbb{P}}(1)$ if $Y_i(1)$ or $Y_i(0)$ is not a constant.

When $\delta > 2$, Proposition 4.3.2 implies $\mathcal{E}_\infty^2/(n\mathcal{E}_2) = O_{\mathbb{P}}(n^{2/\delta-1}) = o_{\mathbb{P}}(1)$, and thus Assumption 4 holds with high probability. We will revisit Proposition 4.3.2 for the consistency of $\hat{\tau}_{\text{adj}}$ and $\hat{\tau}_{\text{adj}}^{\text{de}}$.

4.3.3 Asymptotic Expansions

We derive an asymptotic expansion of $\hat{\tau}_{\text{adj}}$.

Theorem 4.3.3. *Under Assumptions 1 and 2,*

$$\hat{\tau}_{\mathrm{adj}} - \tau = \left(\frac{\boldsymbol{1}^{\mathsf{T}} e_1(1)}{n_1} - \frac{\boldsymbol{1}^{\mathsf{T}} e_0(0)}{n_0} \right) + \left(\frac{n_1}{n_0}\Delta_0 - \frac{n_0}{n_1}\Delta_1 \right) + O_{\mathbb{P}}\left(\sqrt{\frac{\mathcal{E}_2 \kappa^2 p \log p}{n}} + \sqrt{\frac{\mathcal{E}_2 \kappa}{n}} \right). \tag{4.18}$$

The first term in (4.18) is the difference-in-means estimator of the residual potential outcomes based on the population OLS. The second term is non-standard and behaves as a "bias," which motivates the debiased estimator $\hat{\tau}^{\mathrm{de}}_{\mathrm{adj}}$ by subtracting its empirical analogue from $\hat{\tau}_{\mathrm{adj}}$.

We need to analyze Δ_t and $\hat{\Delta}_t - \Delta_t$ to simplify Theorem 4.3.3 and to derive an asymptotic expansion of $\hat{\tau}^{\mathrm{de}}_{\mathrm{adj}}$. Define

$$\Delta = \max\{|\Delta_1|, |\Delta_0|\}. \tag{4.19}$$

The Cauchy–Schwarz inequality implies

$$|\Delta| = \max_{t=0,1} |\Delta_t| \leq \sqrt{\frac{1}{n}\sum_{i=1}^{n} H_{ii}} \times \sqrt{\max_{t=0,1} \frac{1}{n}\sum_{i=1}^{n} e_i^2(t) H_{ii}} \leq \sqrt{\frac{\mathcal{E}_2 \kappa p}{n}}. \tag{4.20}$$

This helps us to obtain the following expansions.

Corollary 4.3.4. *Under Assumptions 1 and 2,*

$$\hat{\tau}_{\mathrm{adj}} - \tau = \frac{\boldsymbol{1}^{\mathsf{T}} e_1(1)}{n_1} - \frac{\boldsymbol{1}^{\mathsf{T}} e_0(0)}{n_0} + O_{\mathbb{P}}\left(\Delta + \sqrt{\frac{\mathcal{E}_2 \kappa^2 p \log p}{n}} + \sqrt{\frac{\mathcal{E}_2 \kappa}{n}} \right) \tag{4.21}$$

$$= \frac{\boldsymbol{1}^{\mathsf{T}} e_1(1)}{n_1} - \frac{\boldsymbol{1}^{\mathsf{T}} e_0(0)}{n_0} + O_{\mathbb{P}}\left(\sqrt{\frac{\mathcal{E}_2 \kappa p}{n}} \right), \tag{4.22}$$

$$\hat{\tau}^{\mathrm{de}}_{\mathrm{adj}} - \tau = \frac{\boldsymbol{1}^{\mathsf{T}} e_1(1)}{n_1} - \frac{\boldsymbol{1}^{\mathsf{T}} e_0(0)}{n_0} + O_{\mathbb{P}}\left(\sqrt{\frac{\mathcal{E}_2 \kappa^2 p \log p}{n}} + \sqrt{\frac{\mathcal{E}_2 \kappa}{n}} \right). \tag{4.23}$$

Expansion (4.21) follows from (4.18) and Assumption 1, and (4.22) holds because the upper bound in (4.20) dominates the third term of (4.18). Expansion (4.23) shows that our de-biasing strategy works because $|\hat{\Delta}_t - \Delta_t|$ is of higher order compared to the third term of (4.23). These asymptotic expansions in Corollary 4.3.4 are crucial for our later analysis.

4.3.4 Consistency

Because the first term in (4.18) is the difference-in-means of the potential residuals, Neyman (1923/1990) implies that it has mean 0 and variance σ_n^2/n. We then use Chebyshev's

inequality to obtain

$$\frac{\mathbf{1}^\mathsf{T} e_1(1)}{n_1} - \frac{\mathbf{1}^\mathsf{T} e_0(0)}{n_0} = O_{\mathbb{P}}\left(\sqrt{\frac{\sigma_n^2}{n}}\right) = O_{\mathbb{P}}\left(\sqrt{\frac{\mathcal{E}_2}{n}}\right). \tag{4.24}$$

Coupled with (4.24) and $\kappa \leq 1$, Corollary 4.3.4 implies that

$$\begin{aligned}
\hat{\tau}_{\text{adj}} - \tau &= O_{\mathbb{P}}\left(\sqrt{\frac{\mathcal{E}_2(\kappa p + 1)}{n}}\right), \\
\hat{\tau}^{\text{de}}_{\text{adj}} - \tau &= O_{\mathbb{P}}\left(\sqrt{\frac{\mathcal{E}_2(\kappa^2 p \log p + 1)}{n}}\right).
\end{aligned}$$

These expansions immediately imply the following consistency result. We essentially require the right-hand sides of the above two identities go to 0.

Theorem 4.3.5. *Under Assumptions 1 and 2, $\hat{\tau}_{\text{adj}}$ is consistent if $\mathcal{E}_2 = o\left(n/(\kappa p + 1)\right)$, and $\hat{\tau}^{\text{de}}_{\text{adj}}$ is consistent if $\mathcal{E}_2 = o\left(n/(\kappa^2 p \log p + 1)\right)$.*

In the classical fixed-p regime, Theorem 4.3.5 implies that both $\hat{\tau}_{\text{adj}}$ and $\hat{\tau}^{\text{de}}_{\text{adj}}$ are consistent when $\mathcal{E}_2 = o(n)$ because $\kappa \leq 1$. From Proposition 4.3.2, the condition $\mathcal{E}_2 = o(n)$ corresponds to the existence of finite first moment under a super-population i.i.d sampling. In the favorable case where $\kappa = O(p/n)$, the same condition $\mathcal{E}_2 = o(n)$ is sufficient for the consistency of $\hat{\tau}_{\text{adj}}$ if $p = O(n^{1/2})$ and for the consistency of $\hat{\tau}^{\text{de}}_{\text{adj}}$ if $p = O(n^{2/3}/(\log n)^{1/3})$. Thus, both estimators are robust to the heavy-tailedness of the potential residuals.

Moreover, when the residuals are not extremely heavy-tailed such that $\mathcal{E}_2 = o(n/p)$, Theorem 4.3.5 implies that both estimators are always consistent, without any further assumption on κ (except Assumption 2). The consistency can hold without a uniformly bounded second moment of the potential residuals.

4.3.5 Asymptotic normality

The first term of (4.18) is the difference-in-means estimator with potential residuals. We can use the classical finite population central limit theorem to show that it is asymptotically normal with mean 0 and variance σ_n^2/n. Therefore, the asymptotic normalities of $\hat{\tau}_{\text{adj}}$ and $\hat{\tau}^{\text{de}}_{\text{adj}}$ hold if the the remainders of (4.21) and (4.23) are asymptotically vanishing after multiplied by $n^{1/2}/\sigma_n$. We first consider $\hat{\tau}_{\text{adj}}$.

Theorem 4.3.6. *Under Assumptions 1–4, $n^{1/2}(\hat{\tau}_{\text{adj}} - \tau)/\sigma_n \xrightarrow{\text{d}} N(0,1)$ if $\kappa^2 p \log p = o(1)$ and $n\Delta^2 = o(\mathcal{E}_2)$.*

Replacing Δ in Theorem (4.3.6) by the upper bound $|\Delta| \leq \sqrt{\mathcal{E}_2 p \kappa / n}$ in (4.20), we obtain the following looser but cleaner result.

Corollary 4.3.7. *Under Assumptions 1–4,* $n^{1/2}(\hat{\tau}_{\text{adj}} - \tau)/\sigma_n \xrightarrow{\text{d}} N(0,1)$ *if* $\kappa p = o(1)$.

In the favorable case where $\kappa = O(p/n)$, the condition $\kappa p = o(1)$ reduces to $p^2/n \to 0$, i.e., $p = o(n^{1/2})$. In this case, Corollary 4.3.7 extends Lin (2013)'s result to $p = o(n^{1/2})$.

The above result can be sharpened if the leverage scores are well-behaved. In fact, because $e(t)$ has mean zero, we can rewrite Δ_t as

$$\Delta_t = n^{-1} \sum_{i=1}^{n} e_i(t) \left(H_{ii} - p/n\right).$$

The Cauchy–Schwarz inequality implies

$$\Delta = \max_{t=0,1} |\Delta_t| \le \max_{1 \le i \le n} \left| H_{ii} - \frac{p}{n} \right| \times \max_{t=0,1} \frac{1}{n} \sum_{i=1}^{n} |e_i(t)| \le \max_{1 \le i \le n} \left| H_{ii} - \frac{p}{n} \right| \sqrt{\mathcal{E}_2}.$$

Therefore, the condition $\Delta = o(\sqrt{\mathcal{E}_2/n})$ in Theorem 4.3.6 holds whenever

$$\max_{1 \le i \le n} \left| H_{ii} - \frac{p}{n} \right| = o\left(n^{-1/2}\right). \tag{4.25}$$

That is, under (4.25), the asymptotic normality of $\hat{\tau}_{\text{adj}}$ holds when the other condition in Theorem 4.3.6 holds, i.e., $\kappa^2 p \log p \to 0$. In the favorable case where $\kappa = O(p/n)$, the condition reduces to $p^3 \log p / n^2 \to 0$, which further implies $p = o(n^{2/3}/(\log n)^{1/3})$. This relaxes the constraint on the dimension to $n^{2/3}$ up to a log-factor. Under $p = o(n^{2/3}/(\log n)^{1/3})$, we can use Proposition 4.3.1 to verify that (4.25) holds with high probability if entries of X are independent and have finite 12-th moments.

Although we relaxes the constraint on the dimension, it is not ideal to impose an extra condition on the leverage scores. In contrast, the debiased estimator is asymptotically normal without any further condition.

Theorem 4.3.8. *Under Assumptions 1–4,* $n^{1/2}(\hat{\tau}^{\text{de}}_{\text{adj}} - \tau)/\sigma_n \xrightarrow{\text{d}} N(0,1)$ *if* $\kappa^2 p \log p = o(1)$.

Therefore, the debiased estimator has better theoretical guarantees. In the asymptotic regime $p = o(n^{2/3}/(\log n)^{1/3})$, we can use Proposition 4.3.1 to verify that the condition $\kappa^2 p \log p = o(1)$ holds if entries of X are independent and have finite $(6+\epsilon)$-th moments.

4.3.6 Variance estimation

The variance estimators $\hat{\sigma}^2_{\text{HC}j}$'s are all asymptotically equivalent because the correction terms in (4.14) are negligible under our asymptotic regime. We can prove that the $\hat{\sigma}^2_{\text{HC}j}$'s for all j are asymptotically conservative estimators of σ_n^2.

Theorem 4.3.9. *Under Assumptions 1–4,* $\hat{\sigma}^2_{\text{HC}j}/\sigma_n^2 \ge 1 + o_{\mathbb{P}}(1)$ *for all* $j \in \{0,1,2,3\}$.

Therefore, the Wald-type confidence intervals for the ATE are all asymptotically conservative.

4.3.7 Related works

Theoretical analyses under the finite-population randomization model are challenging due to the lack of probability tools. The closest work to ours is Bloniarz et al. (2016), which allows p to grow with n and potentially exceed n. However, they assume that the potential outcomes have sparse linear representations based on the covariates, and require $s = o(n^{1/2}/\log p)$ where s is a measure of sparsity. Under additional regularities conditions, they show that $\hat{\tau}(\hat{\beta}_1^{\text{lasso}}, \hat{\beta}_0^{\text{lasso}})$ is consistent and asymptotically normal with $(\hat{\beta}_1^{\text{lasso}}, \hat{\beta}_0^{\text{lasso}})$ being the LASSO coefficients of the covariates. Although the LASSO-adjusted estimator can handle ultra-high dimensional case where $p \gg n$, it has three limitations. First, the requirement $s < < n^{1/2}/\log p$ is stringent. For instance, the PAC-man dataset considered by Bloniarz et al. (2016) has $n = 1013$ and $p = 1172$, so the condition reads $s \ll 4.5$, which implicitly imposes a strong sparse modelling assumption.

Second, the penalty level of the LASSO depends on unobserved quantities. Although they use the cross-validation to select the penalty level, the theoretical properties of this procedure is still unclear. Third, their "restrictive eigenvalue condition" imposes certain non-singularity on the submatrices of the covariate matrix. However, (submatrices of) the covariate matrix can be ill-conditioned especially when interaction terms are included in practice. In addition, this condition is computationally challenging to check.

Admittedly, our results cannot deal with the case of $p > n$. Nevertheless, we argue that $p < n$ is an important regime in many applications.

4.4 Numerical Experiments

We perform extensive numerical experiments to confirm and complement our theory. We examine the performance of the estimators $\hat{\tau}_{\text{adj}}$ and $\hat{\tau}_{\text{adj}}^{\text{de}}$ as well as the variance estimators $\hat{\sigma}_{\text{HC}j}^2$ for $j = 0, 1, 2, 3$. We post the programs to replicate all the experimental results at `https://github.com/lihualei71/RegAdjNeymanRubin/`.

4.4.1 Data Generating Process

We examine the moderate sample performance of the estimators. We set $n = 2000, n_1 = n\pi_1$ for $\pi_1 \in \{0.2, 0.5\}$ and generate a matrix $\mathcal{X} \in \mathbb{R}^{n\times n}$ with i.i.d. entries from $t(2)$. We keep the matrix fixed. For each exponent $\gamma \in \{0, 0.05, \ldots, 0.75\}$, we let $p = \lceil n^\gamma \rceil$ and take the first p columns of $\mathcal{X}$ as the covariate matrix. In Supplementary Material III, we also simulate X with $N(0, 1)$ and $t(1)$ entries and take X from two real datasets. We select $t(2)$ distribution for presentation because it is neither too idealized as $N(0, 1)$ (where $\kappa \sim p/n$), nor too irregular as $t(1)$. It is helpful to illustrate and complement our theory.

With X, we construct the potential outcomes as

$$Y(1) = X\beta_1^* + \sigma_1^*\epsilon(1), \quad Y(0) = X\beta_0^* + \sigma_0^*\epsilon(0), \tag{4.26}$$

with $\beta_1^* = \beta_0^* = 0 \in \mathbb{R}^p$, $\sigma_1^* = \sigma_0^* = 1$, and $\epsilon(1), \epsilon(0) \in \mathbb{R}^n$. Note that for given $\epsilon(1), \epsilon(0)$ and X, both the ATE estimate ($\hat{\tau}_{\text{adj}}$ or $\hat{\tau}_{\text{adj}}^{\text{de}}$) and the variance estimate are invariant to the choices of β_1^* and β_0^*. Similarly, we generate $(\epsilon(1), \epsilon(0))$ as realizations of random vectors with i.i.d. entries from $N(0,1)$, or $t(2)$, or $t(1)$.

Given $X \in \mathbb{R}^{n \times p}$ and potential outcomes $Y(1), Y(0) \in \mathbb{R}^p$, we generate 5000 binary vectors $T \in \mathbb{R}^n$ with n_1 units assigned to treatment. For each assignment vector, we observe half of the potential outcomes.

4.4.2 Repeated Sampling Evaluations

Based on the observe data, we obtain two estimates $\hat{\tau}_{\text{adj}}$ and $\hat{\tau}_{\text{adj}}^{\text{de}}$, as well as five variance estimates $\hat{\sigma}_{\text{HC}j}^2$ $(j = 0, 1, 2, 3)$ and σ_n^2. Technically, σ_n^2 is not an estimate because it is the theoretical asymptotic variance. Below $\hat{\tau}$ can be either $\hat{\tau}_{\text{adj}}$ or $\hat{\tau}_{\text{adj}}^{\text{de}}$, and $\hat{\sigma}^2$ can be any of the five estimates.

Let $\hat{\tau}_1, \ldots, \hat{\tau}_{5000}$ denote the estimates in 5000 replicates, and τ denote the true ATE. The empirical relative absolute bias is $|5000^{-1} \sum_{k=1}^{5000} \hat{\tau}_k - \tau| / \sigma_n$.

Similarly, let $\hat{\sigma}_1^2, \ldots, \hat{\sigma}_{5000}^2$ denote the variance estimates obtained in 5000 replicates, and $\hat{\sigma}_*^2$ denote the empirical variance of $(\hat{\tau}_1, \ldots, \hat{\tau}_{5000})$. We compute the standard deviation inflation ratio $\text{SDR}(\hat{\sigma}) = 5000^{-1} \sum_{k=1}^{5000} \hat{\sigma}_k / \hat{\sigma}_*$. Note that $\hat{\sigma}_*^2$ is an unbiased estimate of true sampling variance of $\hat{\tau}$, which can be different from the theoretical asymptotic variance σ_n^2.

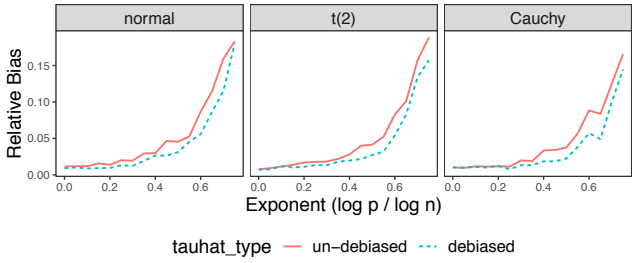
For each estimate and variance estimate, we compute the t-statistic $n^{1/2}(\hat{\tau} - \tau)/\hat{\sigma}$ and the z-score $n^{1/2}(\hat{\tau} - \tau)/\hat{\sigma}_*$. For each t-statistic and the z-score, we estimate the empirical 95% coverage by the proportion within $[-1.96, 1.96]$, the 95% quantile range of $N(0,1)$.

In summary, we compute three measures defined above: relative bias, standard deviation inflation ratios, and 95% coverage. We repeat 50 times using different random seeds and record the medians of each measure. Fig. 4.1 summarizes the results. We emphasize that for each experiment, both X and potential outcomes are fixed and the randomness only comes from treatment assignments.

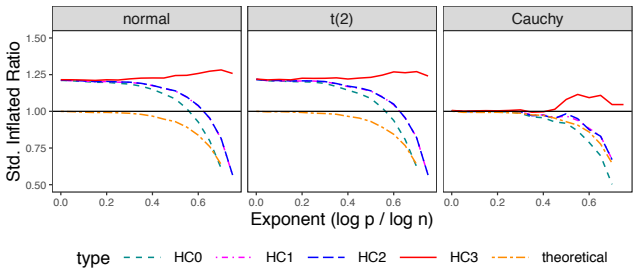
4.4.3 Results

From Figures 4.4.3 and 4.4.3, $\hat{\tau}_{\text{adj}}^{\text{de}}$ does reduce the bias regardless of the distribution of potential outcomes, especially for moderately large p. It is noteworthy that the relative bias is too small ($\leq 15\%$) to affect coverage.

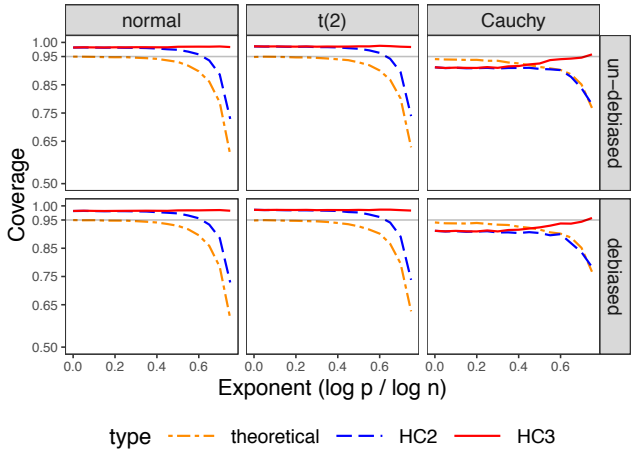
For standard deviation inflation ratios, we find that the true sampling variances of $\hat{\tau}_{\text{adj}}$ and $\hat{\tau}_{\text{adj}}^{\text{de}}$ are almost identical and thus we set the sampling variance of $\hat{\tau}_{\text{adj}}$ as the baseline variance $\hat{\sigma}_*^2$. Figures 4.4.3 and 4.4.3 shows an interesting phenomenon that the theoretical asymptotic variance σ_n^2 tends to underestimate the true sampling variance for large p. Corollary 4.3.4 partially suggests this. The theoretical asymptotic variance is simply the variance of the first term while the finite sample variance also involves the second term and, more importantly, the error term, which can be large in the presence of high dimensional or influential observations. All variance estimators overestimate σ_n^2 because they all ignore the third term of σ_n^2. However,



(a) Relative bias of $\hat{\tau}_{\text{adj}}^{\text{de}}$ and $\hat{\tau}_{\text{adj}}$.

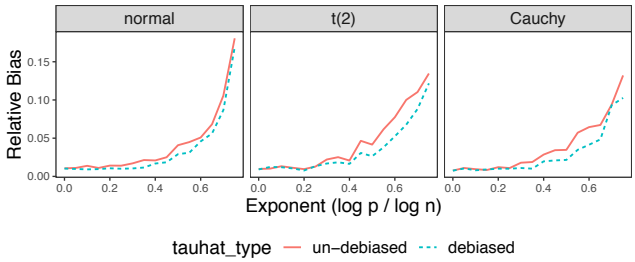


(b) Ratio of standard deviation between five standard deviation estimates, $\sigma_n, \hat{\sigma}_{\text{HC0}}, \hat{\sigma}_{\text{HC1}}, \hat{\sigma}_{\text{HC2}}, \hat{\sigma}_{\text{HC3}}$, and the true standard deviation of $\hat{\tau}_{\text{adj}}$.

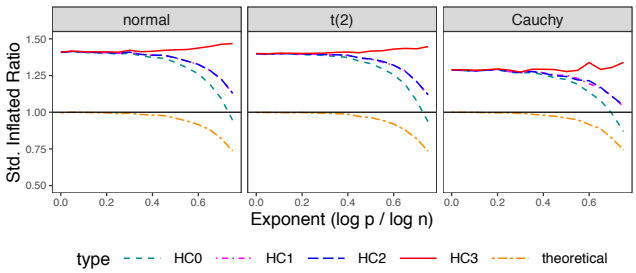


(c) Empirical 95% coverage of t-statistics derived from two estimators and four variance estimators ("theoretical" for σ_n^2, "HC2" for $\hat{\sigma}_{\text{HC2}}^2$ and "HC3" for $\hat{\sigma}_{\text{HC3}}^2$)

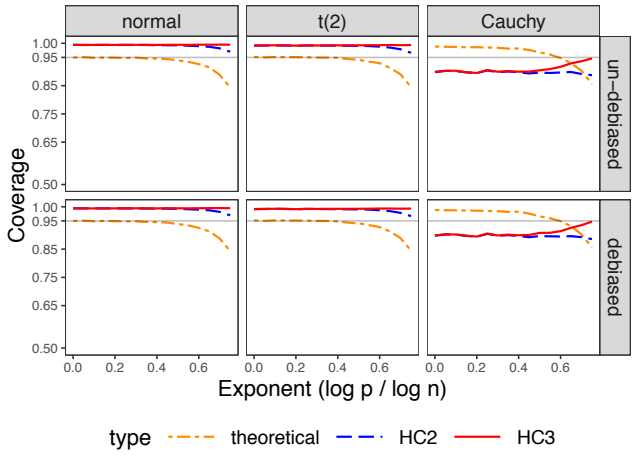
Figure 4.1: Simulation with $\pi_1 = 0.2$. X is a realization of a random matrix with i.i.d. $t(2)$ entries, and $e(t)$ is a realization of a random vector with i.i.d. entries from a distribution corresponding to each column.



(a) Relative bias of $\hat{\tau}_{\text{adj}}^{\text{de}}$ and $\hat{\tau}_{\text{adj}}$.



(b) Ratio of standard deviation between five standard deviation estimates, $\sigma_n, \hat{\sigma}_{\text{HC0}}, \hat{\sigma}_{\text{HC1}}, \hat{\sigma}_{\text{HC2}}, \hat{\sigma}_{\text{HC3}}$, and the true standard deviation of $\hat{\tau}_{\text{adj}}$.



(c) empirical 95% coverage of t-statistics derived from two estimators and four variance estimators ("theoretical" for σ_n^2, "HC2" for $\hat{\sigma}_{\text{HC2}}^2$ and "HC3" for $\hat{\sigma}_{\text{HC3}}^2$)

Figure 4.2: Simulation. X is a realization of a random matrix with i.i.d. $t(2)$ entries, and $e(t)$ is a realization of a random vector with i.i.d. entries from a distribution corresponding to each column.

all estimators, except the HC3 estimator, tend to underestimate the true sampling variance for large p. In contrast, the HC3 estimator does not suffer from anti-conservatism.

Figures 4.4.3 and 4.4.3 shows that HC0 and HC1 variance estimates lie between the theoretical asymptotic variance and the HC2 variance estimate. For better visualization, we only plot the 95% coverage of t-statistics computed from $\sigma_n^2, \hat{\sigma}_{\text{HC2}}^2$ and $\hat{\sigma}_{\text{HC3}}^2$ in Figures 4.4.3 and 4.4.3. We draw the following conclusions from Figures 4.4.3 and 4.4.3. First, as we pointed out previously, the coverage of two ATE estimates are almost identical because the relative bias is small in these scenarios. Second, as Figures 4.4.3 and 4.4.3 suggest, the t-statistic with HC3 variance estimate has the best coverage, and it protects the coverage against the increasing dimension. In contrast, the theoretical asymptotic variance and HCj ($j = 0, 1, 2$) variance estimates yield significantly lower coverage for large p. Therefore, we advocate using $\hat{\sigma}_{\text{HC3}}^2$ for variance estimation.

4.4.4 Effectiveness of debiasing

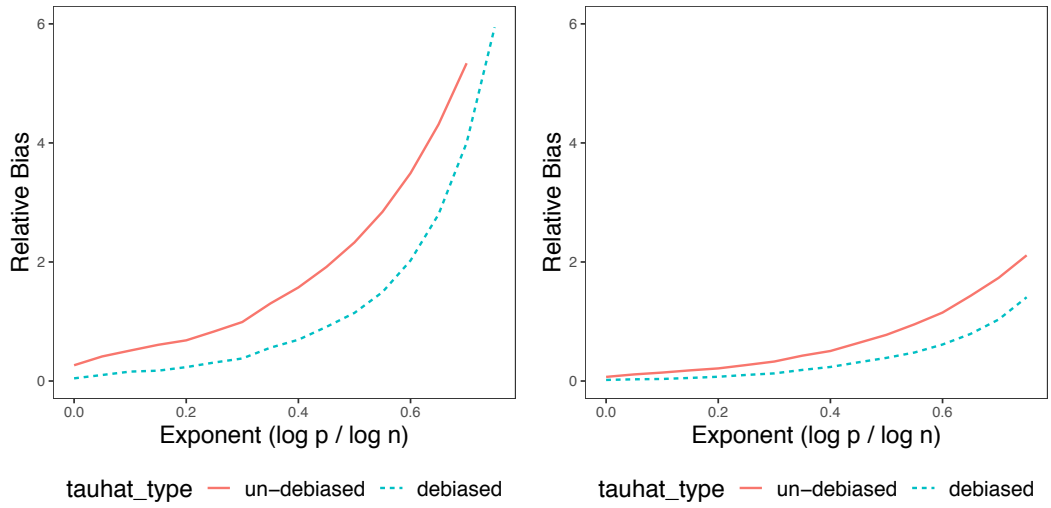
In the aforementioned settings, the debiased estimator yields almost identical inference as the undebiased estimator. This is not surprising because in the above scenarios the potential outcomes are generated from linear models and thus Lin (2013)'s estimator has bias close to zero. However, in practice, the potential outcomes might not have prefect linear relationships with the covariates. To illustrate the potential benefits of debiasing, we consider the "most-biased" situation which maximizes the "bias term", measured as the second term in the expansion (4.18). Specifically, we consider the case where $\epsilon(0) = \epsilon$ and $\epsilon(1) = 2\epsilon$ for some vector ϵ that satisfies (4.7) with sample variance 1. To maximize the bias term, we take ϵ as the solution of

$$\begin{aligned} &\max_{\epsilon \in \mathbb{R}^n} \left| \frac{n_1}{n_0}\Delta_0 - \frac{n_0}{n_1}\Delta_1 \right| = \left(\frac{2n_0}{n_1} - \frac{n_1}{n_0} \right) \left| \sum_{i=1}^{n} H_{ii}\epsilon_i \right|, \\ &\text{s.t. } \|\epsilon\|_2^2/n = 1, X^{\mathsf{T}}\epsilon = \mathbf{1}^{\mathsf{T}}\epsilon = 0. \end{aligned} \tag{4.27}$$

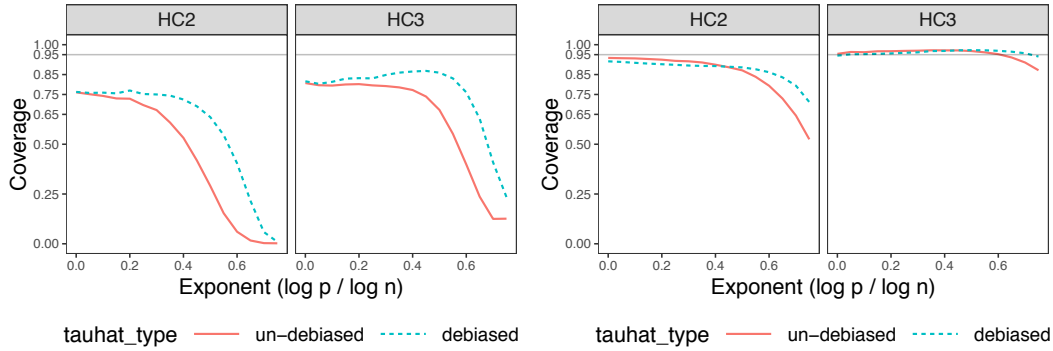
We give more details of constructing ϵ in Section C.6 of Supplementary Material III. From (4.27), the bias is amplified when the group sizes are unbalanced. Note that this setting essentially assume a non-linear relationship between the potential outcomes and the covariates.

We perform simulation detailed in Section 4.4.2 based on potential outcomes in (4.27) and report relative bias and coverage to demonstrate the effectiveness of debiasing. To save space, we only report the coverage for $\hat{\sigma}_{\text{HC2}}^2$ and $\hat{\sigma}_{\text{HC3}}^2$. Fig. 4.3 summarizes the results.

Unlike the previous settings, the relative bias in this setting is large enough to affect the coverage. From Fig. 4.4.4, as expected, we see that the relative bias is larger when the group sizes are more unbalanced. The debiased estimator reduces a fair proportion of bias in both cases and improves coverage especially when the dimension is high. We provide experimental results in more settings in Supplementary Material III, which confirm the effectiveness of debiasing.



(a) Relative bias of $\hat{\tau}^{\text{de}}_{\text{adj}}$ and $\hat{\tau}_{\text{adj}}$.



(b) Empirical 95% coverage of t-statistics derived from two estimators and two variance estimators ("HC2" for $\hat{\sigma}^2_{\text{HC2}}$ and "HC3" for $\hat{\sigma}^2_{\text{HC3}}$)

Figure 4.3: Simulation. X is a realization of a random matrix with i.i.d. $t(2)$ entries and $e(t)$ is defined in (4.27): (Left) $\pi_1 = 0.2$; (Right) $\pi_1 = 0.5$.

4.4.5 Trimming the Design Matrix

Our theory suggests that κ of the design matrix affects the statistical properties of $\hat{\tau}_{\text{adj}}$ and $\hat{\tau}^{\text{de}}_{\text{adj}}$. When there are many influential observations in the data, it is beneficial to reduce κ before regression adjustment. Because our theory holds even for mis-specified linear models, any preprocessing of X does not affect the consistency and asymptotic normality if the preprocessing does not depend on T or Y^{obs}. This is a feature of our theory. In contrast, trimming is not applicable to the theory under a super-population perspective assuming a correctly specified regression model.

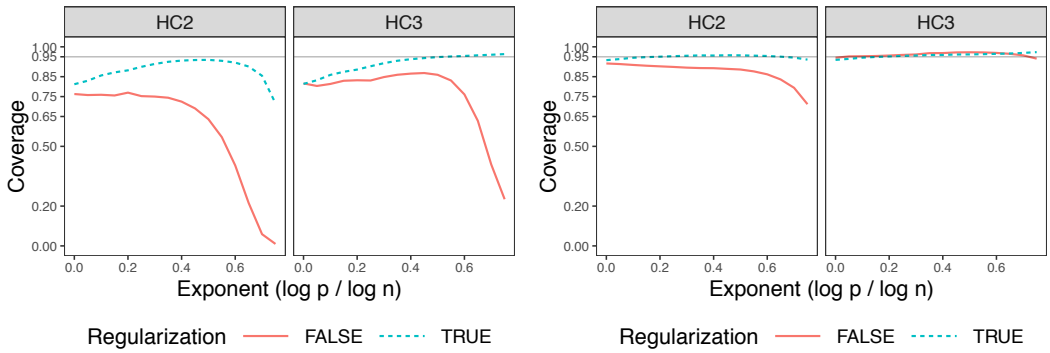


Figure 4.4: Simulation. Empirical 95% coverage of t-statistics derived from the debiased estimator with and without trimming the covariate matrix: (Left) $\pi_1 = 0.2$; (Right) $\pi_1 = 0.5$. X is a realization of a random matrix with i.i.d. $t(2)$ entries and $e(t)$ is defined in (4.27).

In Section 4.4, the entries of X are realizations of heavy-tailed random variables, and κ increases even with an infrequent extreme covariate value. For the 50 design matrices used in Section 4.4 with $p = \lceil n^{2/3} \rceil$ and $n = 2000$, the average of κ is 0.9558 with standard error 0.0384. Now we consider a simple form of trimming which thresholds each column at its 2.5% and 97.5% quantiles. Then the average of κ reduces dramatically to 0.0704 with standard error 0.0212. Fig. 4.4 shows the coverage of the t-statistics derived from $\hat{\tau}^{\text{de}}_{\text{adj}}$ with and without the trimming. It is clear that the coverage gets drastically improved after trimming.

Since the main goal of this chapter is not on statistical methodology, we only propose the above heuristic approach to illustrate the idea of trimming, motivated by our asymptotic theory. The general methodology is an interesting future research topic.

4.5 Conclusions and Practical Suggestions

Fisher (1935b) advocated using the analysis of covariance under treatment-unit additivity. Freedman (2008b) highlighted its dangers under treatment effect heterogeneity. Lin (2013) proposed a simple OLS estimator with treatment-covariate interactions accounting for potential heterogeneity. We establish the consistency and the asymptotic normality of Lin (2013)'s estimator allowing for a growing dimension of the covariates. We further propose a debiased estimator which permits valid inference in broader asymptotic regimes.

In summary, we find that the classical inferential procedure tends to be invalid when the design matrix has many covariates or many influential observations. In these scenarios, the bias blows up and the variance estimation becomes anti-conservative. We suggest using the debiased estimator (4.10) and the HC3 variance estimator for inference. In addition, we suggest trimming the design matrix to reduce the extreme leverage scores before regression adjustment.

4.6 Technical Lemmas

4.6.1 Some general results for sampling without replacement

Completely randomized experiments have deep connections with sampling without replacement because the treatment and control groups are simple random samples from a finite population of n units. Below we use $\mathcal{T}$ to denote a random size-m subset of $\{1, \dots, n\}$ over all $\binom{n}{m}$ subsets, and $\mathcal{S}^{p-1} = \{(\omega_1, \dots, \omega_p)^\mathsf{T} : \omega_1^2 + \cdots + \omega_p^2 = 1\}$ to denote the $(p-1)$-dimensional unit sphere.

The first lemma gives the mean and variance of the sample total from sampling without replacement. See Cochran (2007, Theorem 2.2) for a proof.

Lemma 4.6.1. *Let $(w_1, \dots, w_n)$ be fixed scalars with mean $\bar{w} = n^{-1} \sum_{i=1}^n w_i$. Then $\sum_{i \in \mathcal{T}} w_i$ has mean $m\bar{w}$ and variance*

$$\operatorname{Var}\left(\sum_{i \in \mathcal{T}} w_i \right) = \frac{m(n-m)}{n(n-1)} \sum_{i=1}^n (w_i - \bar{w})^2.$$

The second lemma gives the Berry–Esseen-type bound for the finite population central limit theorem. See Bikelis (1969) and Höglund (1978) for proofs.

Lemma 4.6.2. *Let $(w_1, \dots, w_n)$ be fixed scalars with $\bar{w} = n^{-1} \sum_{i=1}^n w_i$ and $S_w^2 = \sum_{i=1}^n (w_i - \bar{w})^2$. Let $m = nf$ for some $f \in (0, 1)$. Then*

$$\begin{aligned} d_{\mathrm{K}}\left(\frac{\sum_{i \in \mathcal{T}} (w_i - \bar{w})}{S_w \sqrt{f(1-f)}}, N(0,1) \right) &\leq \frac{C}{\sqrt{f(1-f)}} \frac{\sum_{i=1}^n (w_i - \bar{w})^2}{S_w^3} \\ &\leq \frac{C}{\sqrt{f(1-f)}} \frac{\max_{1 \leq i \leq n} |w_i - \bar{w}|}{S_w}, \end{aligned}$$

where d_{K} denotes the Kolmogorov distance between two distributions, and C is a universal constant.

The following two lemmas give novel vector and matrix concentration inequalities for sampling without replacement.

Lemma 4.6.3. *Let $(u_1, \dots, u_n)$ be a finite population of p-dimensional vectors with $\sum_{i=1}^n u_i = 0$. Then for any $\delta \in (0, 1)$, with probability $1 - \delta$*

$$\left\| \sum_{i \in \mathcal{T}} u_i \right\|_2 \leq \|U\|_F \sqrt{\frac{m(n-m)}{n(n-1)}} + \|U\|_{\mathrm{op}} \sqrt{8 \log \frac{1}{\delta}}$$

where u_i^T is the i-th row of the matrix $U \in \mathbb{R}^{n \times p}$.

Lemma 4.6.4. *Let $(V_1, \ldots, V_n)$ be a finite population of $(p \times p)$-dimensional Hermittian matrices with $\sum_{i=1}^n V_i = 0$. Let $C(p) = 4(1 + \lceil 2 \log p \rceil)$, and*

$$\nu^2 = \left\| \frac{1}{n} \sum_{i=1}^n V_i^2 \right\|_{\mathrm{op}}, \quad \nu_-^2 = \sup_{\omega \in \mathcal{S}^{p-1}} \frac{1}{n} \sum_{i=1}^n (\omega^\mathsf{T} V_i \omega)^2, \quad \nu_+ = \max_{1 \le i \le n} \|V_i\|_{\mathrm{op}}.$$

Then for any $\delta \in (0,1)$, with probability $1 - \delta$,

$$\left\| \sum_{i \in \mathcal{T}} V_i \right\|_{\mathrm{op}} \le \sqrt{nC(p)}\nu + C(p)\nu_+ + \sqrt{8n \log \frac{1}{\delta}}\, \nu_-.$$

The following lemma gives the mean and variance of the summation over randomly selected rows and columns from a deterministic matrix $Q \in \mathbb{R}^{n \times n}$.

Lemma 4.6.5. *Let $Q \in \mathbb{R}^{n \times n}$ be a deterministic matrix, and $Q_{\mathcal{T}} \equiv \sum_{i,j \in \mathcal{T}} Q_{ij}$. Assume $n \ge 4$. Then*

$$\mathbb{E} Q_{\mathcal{T}} = \frac{m(n-m)}{n(n-1)} \operatorname{tr}(Q) + \frac{m(m-1)}{n(n-1)} \boldsymbol{1}^\mathsf{T} Q \boldsymbol{1}.$$

If Q further satisfies $\boldsymbol{1}^\mathsf{T} Q = Q\boldsymbol{1} = 0$, then

$$\operatorname{Var}(Q_{\mathcal{T}}) \le \frac{m(n-m)}{n(n-1)} \|Q\|_F^2.$$

Lemmas 4.6.3–4.6.5 are critical for our proofs. The proofs are relegated to Supplementary Material I. They are novel tools to the best of our knowledge and potentially useful in other contexts such as survey sampling, matrix sketching, and transductive learning.

4.6.2 Some results particularly useful for our setting

We first give an implication of Assumption 3, a lower bound on σ_n^2 under Assumption 1.

Lemma 4.6.6. *Under Assumptions 1 and 3, $\sigma_n^2 \ge \eta \min\{n_1/n_0, n_0/n_1\}\, \mathcal{E}_2$.*

Recall $H_t = X_t (X_t^\mathsf{T} X_t)^{-1} X_t^\mathsf{T}$ and define $\Sigma_t = n_t^{-1} X_t^\mathsf{T} X_t$ $(t = 0, 1)$. The following explicit formula is the starting point of our proof.

Lemma 4.6.7. *We have*

$$\hat{\tau}_{\mathrm{adj}} - \tau = \frac{\boldsymbol{1}^\mathsf{T} e_1(1)/n_1 - \boldsymbol{1}^\mathsf{T} H_1 e_1(1)/n_1}{1 - \boldsymbol{1}^\mathsf{T} H_1 \boldsymbol{1}/n_1} - \frac{\boldsymbol{1}^\mathsf{T} e_0(0)/n_0 - \boldsymbol{1}^\mathsf{T} H_0 e_0(0)/n_0}{1 - \boldsymbol{1}^\mathsf{T} H_0 \boldsymbol{1}/n_0}. \tag{4.28}$$

The quantities μ_t, $e(t)$, and our estimators $(\hat{\tau}_{\mathrm{adj}}, \hat{\tau}_{\mathrm{adj}}^{\mathrm{de}})$ are all invariant if X is transformed to XZ for any full rank matrix $Z \in \mathbb{R}^{p \times p}$, provided that (4.2) holds. Thus, without loss of generality, we assume

$$n^{-1} X^\mathsf{T} X = \mathrm{I}. \tag{4.29}$$

Otherwise, suppose X has the singular value decomposition $U\Sigma V^\mathsf{T}$ with $U \in \mathbb{R}^{n\times p}, \Sigma, V \in \mathbb{R}^{p\times p}$, then we can replace X by $n^{1/2}U = X(n^{1/2}V\Sigma^{-1})$ to ensure (4.29). We can verify that the key properties in (4.7) still hold. Assuming (4.29), we can rewrite the hat matrix and the leverage scores as

$$H = n^{-1}XX^\mathsf{T}, \quad H_{ii} = n^{-1}\|x_i\|_2^2, \quad H_{ij} = n^{-1}x_i^\mathsf{T}x_j. \tag{4.30}$$

Note that the invariance property under the standardization (4.29) is a feature of the OLS-based regression adjustment. It does not hold for many other estimators (e.g., Bloniarz et al. 2016; Wager et al. 2016).

We will repeatedly use the following results to obtain the stochastic orders of the terms in (4.28). They are consequences of Lemmas 4.6.3 and 4.6.4.

Lemma 4.6.8. *Under Assumption 1, for $t = 0, 1$,*

$$\frac{\boldsymbol{1}^\mathsf{T}e_t(t)}{n_t} = O_\mathbb{P}\left(\sqrt{\frac{\mathcal{E}_2}{n}}\right), \quad \left\|\frac{X_t^\mathsf{T}\boldsymbol{1}}{n_t}\right\|_2 = O_\mathbb{P}\left(\sqrt{\frac{p}{n}}\right),$$
$$\left\|\frac{X_t^\mathsf{T}e_t(t)}{n_t}\right\|_2 = O_\mathbb{P}\left(\sqrt{\mathcal{E}_2\kappa}\right).$$

Lemma 4.6.9. *Under Assumptions 1, 2 and (4.29), for $t = 0, 1$,*

$$\|\Sigma_t - \mathrm{I}\|_{\mathrm{op}} = O_\mathbb{P}\left(\sqrt{\kappa\log p}\right), \quad \left\|\Sigma_t^{-1}\right\|_{\mathrm{op}} = O_\mathbb{P}(1),$$
$$\left\|\Sigma_t^{-1} - \mathrm{I}\right\|_{\mathrm{op}} = O_\mathbb{P}\left(\sqrt{\kappa\log p}\right).$$

The following lemma states some key properties of an intermediate quantity, which will facilitate our proofs.

Lemma 4.6.10. *Define $Q(t) = H\,\mathrm{diag}\,(e(t)) = (H_{ij}e_j(t))_{i,j=1}^n$. It satisfies*

$$\boldsymbol{1}^\mathsf{T}Q(t) = 0, \quad Q(t)\boldsymbol{1} = 0, \quad \boldsymbol{1}^\mathsf{T}Q(t)\boldsymbol{1} = 0,$$
$$\mathrm{tr}(Q(t)) = n\Delta_t, \quad \|Q(t)\|_F^2 = \sum_{i=1}^n e_i^2(t)H_{ii} \le n\mathcal{E}_2\kappa.$$

4.7 Proofs of The Main Results

4.7.1 Proof of the asymptotic expansions

Proof of Theorem 4.3.3. We need to analyze the terms in (4.28). First, by Lemmas 4.6.8 and 4.6.9,

$$\frac{\mathbf{1}^\mathsf{T}H_t\mathbf{1}}{n_t} = \frac{\mathbf{1}^\mathsf{T}X_t}{n_t}\Sigma_t^{-1}\frac{X_t^\mathsf{T}\mathbf{1}}{n_t} \le \left\|\Sigma_t^{-1}\right\|_{\mathrm{op}}\left\|\frac{X_t^\mathsf{T}\mathbf{1}}{n_t}\right\|_2^2 = O_\mathbb{P}\left(\frac{p}{n}\right).$$

Using (4.17) that $p = o(n)$, we obtain that

$$\frac{1}{1 - \mathbf{1}^\mathsf{T} H_t \mathbf{1}/n_t} = 1 + O_{\mathbb{P}}\left(\frac{p}{n}\right). \tag{4.31}$$

Second,

$$\begin{aligned}
&\frac{\mathbf{1}^\mathsf{T} H_t e_t(t)}{n_t} = \frac{\mathbf{1}^\mathsf{T} X_t}{n_t} \Sigma_t^{-1} \frac{X_t^\mathsf{T} e_t(t)}{n_t} \\
=&\frac{\mathbf{1}^\mathsf{T} X_t}{n_t} \frac{X_t^\mathsf{T} e_t(t)}{n_t} + \frac{\mathbf{1}^\mathsf{T} X_t}{n_t} \left(\Sigma_t^{-1} - \mathrm{I}\right) \frac{X_t^\mathsf{T} e_t(t)}{n_t} \equiv R_{t1} + R_{t2}.
\end{aligned} \tag{4.32}$$

Note that here we do not use the naive bound for $\mathbf{1}^\mathsf{T} H_t e_t(t)/n_t$ as for $\mathbf{1}^\mathsf{T} H_t \mathbf{1}/n_t$ in (4.31) because this gives weaker results. Instead, we bound R_{t1} and R_{t2} separately. Lemmas 4.6.8 and 4.6.9 imply

$$R_{t2} \le \left\|\Sigma_t^{-1} - \mathrm{I}\right\|_{\mathrm{op}} \left\|\frac{X_t^\mathsf{T} \mathbf{1}}{n_t}\right\|_2 \left\|\frac{X_t^\mathsf{T} e_t(t)}{n_t}\right\|_2 = O_{\mathbb{P}}\left(\sqrt{\frac{\mathcal{E}_2 \kappa^2 p \log p}{n}}\right). \tag{4.33}$$

We apply Chebyshev's inequality to obtain that

$$R_{t1} = \mathbb{E} R_{t1} + O_{\mathbb{P}}\left(\sqrt{\mathrm{Var}(R_{t1})}\right). \tag{4.34}$$

Therefore, to bound R_{t1}, we need to calculate its first two moments. Recalling (4.30) and the definition of $Q(t)$ in Lemma 4.6.10, we have

$$\begin{aligned}
R_{t1} &= \frac{1}{n_t^2} \left(\sum_{i \in \mathcal{T}_t} x_i^\mathsf{T}\right) \left(\sum_{i \in \mathcal{T}_t} x_i e_i(t)\right) = \frac{1}{n_t^2} \sum_{i \in \mathcal{T}_t} \sum_{j \in \mathcal{T}_t} x_i^\mathsf{T} x_j e_j(t) \\
&= \frac{1}{n_t^2} \sum_{i \in \mathcal{T}_t} \sum_{j \in \mathcal{T}_t} n H_{ij} e_j(t) = \frac{n}{n_t^2} \sum_{i \in \mathcal{T}_t} \sum_{j \in \mathcal{T}_t} Q_{ij}(t).
\end{aligned} \tag{4.35}$$

Lemmas 4.6.5 and 4.6.10 imply the expectation of R_{t1}:

$$\begin{aligned}
\mathbb{E} R_{t1} &= \frac{n}{n_t^2} \left(\frac{n_1 n_0}{n(n-1)} \mathrm{tr}\left(Q(t)\right) + \frac{n_t(n_t - 1)}{n(n-1)} \mathbf{1}^\mathsf{T} Q(t) \mathbf{1}\right) \\
&= \frac{n n_1 n_0}{n_t^2 (n-1)} \Delta_t = \frac{n_1 n_0}{n_t^2} \Delta_t + O\left(\frac{|\Delta_t|}{n}\right).
\end{aligned} \tag{4.36}$$

We then bound the variance of R_{t1}:

$$\mathrm{Var}(R_{t1}) \;=\; \frac{n^2}{n_t^4} \mathrm{Var}\left(\sum_{i,j \in \mathcal{T}_t} Q_{ij}(t)\right) \le \frac{n^2}{n_t^4} \frac{n_1 n_0}{n(n-1)} \|Q(t)\|_F^2 \tag{4.37}$$

$$\leq \frac{n^2}{n_t^4}\frac{n_1 n_0}{n(n-1)} n\mathcal{E}_2\kappa = O\left(\frac{\mathcal{E}_2\kappa}{n}\right), \tag{4.38}$$

where (4.37) follows from Lemma 4.6.5, (4.38) follows from Lemma 4.6.10 and Assumption 1. Putting (4.32)–(4.36) and (4.38) together, we obtain that

$$\frac{\mathbf{1}^\mathsf{T} H_t e_t(t)}{n_t} = \frac{n_1 n_0}{n_t^2}\Delta_t + O_\mathbb{P}\left(\sqrt{\frac{\mathcal{E}_2\kappa^2 p\log p}{n}} + \frac{|\Delta_t|}{n} + \sqrt{\frac{\mathcal{E}_2\kappa}{n}}\right) \tag{4.39}$$

By (4.20) and (4.17) that $p = o(n)$, (4.39) further simplifies to

$$\frac{\mathbf{1}^\mathsf{T} H_t e_t(t)}{n_t} = \frac{n_1 n_0}{n_t^2}\Delta_t + O_\mathbb{P}\left(\sqrt{\frac{\mathcal{E}_2\kappa^2 p\log p}{n}} + \sqrt{\frac{\mathcal{E}_2\kappa}{n}}\right). \tag{4.40}$$

Using Lemma 4.6.8, (4.40), and the fact that $\kappa \leq 1$, we have

$$\frac{\mathbf{1}^\mathsf{T} e_t(t)}{n_t} - \frac{\mathbf{1}^\mathsf{T} H_t e_t(t)}{n_t} = O_\mathbb{P}\left(\sqrt{\frac{\mathcal{E}_2}{n}} + \Delta + \sqrt{\frac{\mathcal{E}_2\kappa^2 p\log p}{n}}\right). \tag{4.41}$$

Finally, putting (4.31), (4.40) and (4.41) together into (4.28), we obtain that

$$\begin{aligned}
\hat{\tau}_{\text{adj}} - \tau &= \left(\frac{\mathbf{1}^\mathsf{T} e_1(1)}{n_1} - \frac{\mathbf{1}^\mathsf{T} H_1 e_1(1)}{n_1}\right)\left(1 + O_\mathbb{P}\left(\frac{p}{n}\right)\right) \\
&\quad - \left(\frac{\mathbf{1}^\mathsf{T} e_0(0)}{n_0} - \frac{\mathbf{1}^\mathsf{T} H_0 e_0(0)}{n_0}\right)\left(1 + O_\mathbb{P}\left(\frac{p}{n}\right)\right) \\
&= \frac{\mathbf{1}^\mathsf{T} e_1(1)}{n_1} - \frac{\mathbf{1}^\mathsf{T} e_0(0)}{n_0} + \frac{\mathbf{1}^\mathsf{T} H_0 e_0(0)}{n_0} - \frac{\mathbf{1}^\mathsf{T} H_1 e_1(1)}{n_1} \\
&\quad + O_\mathbb{P}\left(\sqrt{\frac{p^2\mathcal{E}_2}{n^3}} + \frac{p\Delta}{n} + \sqrt{\frac{\mathcal{E}_2\kappa^2 p^3\log p}{n^3}}\right) \\
&= \frac{\mathbf{1}^\mathsf{T} e_1(1)}{n_1} - \frac{\mathbf{1}^\mathsf{T} e_0(0)}{n_0} + \frac{n_1}{n_0}\Delta_0 - \frac{n_0}{n_1}\Delta_1 \\
&\quad + O_\mathbb{P}\left(\sqrt{\frac{p^2\mathcal{E}_2}{n^3}} + \frac{p\Delta}{n} + \sqrt{\frac{\mathcal{E}_2\kappa^2 p\log p}{n}} + \sqrt{\frac{\mathcal{E}_2\kappa}{n}}\right).
\end{aligned} \tag{4.42}$$

where (4.42) uses (4.17) that $p = o(n)$. The fourth terms dominates the first term in (4.42) because $p = o(n)$ and $\kappa \geq p/n$. The third term dominates the second term in (4.42) because, by (4.20),

$$\frac{p\Delta}{n} \leq \kappa\Delta \leq \sqrt{\kappa}\Delta = O\left(\sqrt{\frac{\mathcal{E}_2\kappa^2 p}{n}}\right).$$

Deleting the first two terms in (4.42), we complete the proof. □

Proof of Corollary 4.3.4. Assumption 1 implies $\frac{n_1}{n_0}\Delta_0 - \frac{n_0}{n_1}\Delta_1 = O(\Delta)$, which, coupled with Theorem 4.3.3, implies (4.21).

The key is to prove the result for the debiased estimator. By definition,

$$\hat{\tau}^{\text{de}}_{\text{adj}} - \tau = \frac{\mathbf{1}^\mathsf{T} e_1(1)}{n_1} - \frac{\mathbf{1}^\mathsf{T} e_0(0)}{n_0} + \frac{n_1}{n_0}(\Delta_0 - \hat{\Delta}_0) - \frac{n_0}{n_1}(\Delta_1 - \hat{\Delta}_1) + O_{\mathbb{P}}\left(\sqrt{\frac{\mathcal{E}_2\kappa^2 p \log p}{n}} + \sqrt{\frac{\mathcal{E}_2\kappa}{n}}\right),$$

and therefore, the key is to bound $|\Delta_t - \hat{\Delta}_t|$.

We introduce an intermediate quantity $\tilde{\Delta}_t = n_t^{-1}\sum_{i\in\mathcal{T}_t} H_{ii}e_i(t)$. It has mean $\mathbb{E}\tilde{\Delta}_t = \Delta_t$ and variance

$$\text{Var}(\tilde{\Delta}_t) \le \frac{1}{n_t^2}\frac{n_1 n_0}{n(n-1)}\sum_{i=1}^n H_{ii}^2 e_i^2(t) \le \frac{\mathcal{E}_2\kappa^2}{n_t^2} = O\left(\frac{\mathcal{E}_2\kappa^2}{n}\right), \tag{4.43}$$

from Lemma 4.6.1 and Assumption 1. Equipped with the first two moments, we use Chebyshev's inequality to obtain

$$|\tilde{\Delta}_t - \Delta_t| = O_{\mathbb{P}}\left(\sqrt{\frac{\mathcal{E}_2\kappa^2}{n}}\right). \tag{4.44}$$

Next we bound $|\hat{\Delta}_t - \tilde{\Delta}_t|$. The Cauchy–Schwarz inequality implies

$$|\hat{\Delta}_t - \tilde{\Delta}_t| \le \frac{1}{n_t}\sum_{i\in\mathcal{T}_t} H_{ii}|\hat{e}_i - e_i(t)| \le \sqrt{\frac{1}{n_t}\sum_{i\in\mathcal{T}_t} H_{ii}^2}\sqrt{\frac{1}{n_t}\sum_{i\in\mathcal{T}_t}(\hat{e}_i - e_i(t))^2}. \tag{4.45}$$

First,

$$\frac{1}{n_t}\sum_{i\in\mathcal{T}_t} H_{ii}^2 \le \frac{n\kappa}{n_t}\left(\frac{1}{n}\sum_{i=1}^n H_{ii}\right) = O\left(\frac{\kappa p}{n}\right). \tag{4.46}$$

Second, using the fact $\hat{e}_t = (\mathrm{I} - H_t)e_t(t)$, we have

$$\begin{aligned}
&\frac{1}{n_t}\sum_{i\in\mathcal{T}_t}(\hat{e}_i - e_i(t))^2 = \frac{1}{n_t}\|\hat{e}_t - e_t(t)\|_2^2 = \frac{1}{n_t}e_t^\mathsf{T}(t)H_t e_t(t) \\
&= \left(\frac{X_t^\mathsf{T} e_t(t)}{n_t}\right)^\mathsf{T} \Sigma_t^{-1}\frac{X_t^\mathsf{T} e_t(t)}{n_t} \\
&\le \|\Sigma_t\|_{\text{op}}^{-1}\left\|\frac{X_t^\mathsf{T} e_t(t)}{n_t}\right\|_2^2 = O_{\mathbb{P}}(\mathcal{E}_2\kappa),
\end{aligned} \tag{4.47}$$

where the last line follows from Lemma 4.6.8. Putting (4.46) and (4.47) into (4.45), we obtain

$$|\hat{\Delta}_t - \tilde{\Delta}_t| = O_{\mathbb{P}}\left(\sqrt{\frac{\mathcal{E}_2\kappa^2 p}{n}}\right). \tag{4.48}$$

Combining (4.44) and (4.48) together, we have $|\hat{\Delta}_t - \Delta_t| = O_{\mathbb{P}}\left(\sqrt{\mathcal{E}_2\kappa^2 p/n}\right)$. We complete the proof by invoking Theorem 4.3.3. □

4.7.2 Proof of asymptotic normality

Proofs of Theorems 4.3.6 and 4.3.8. We first prove the asymptotic normality of the first term in the expansions:

$$\frac{n^{1/2}}{\sigma_n}\left(\frac{\mathbf{1}^{\mathsf{T}}e_1(1)}{n_1} - \frac{\mathbf{1}^{\mathsf{T}}e_0(0)}{n_0}\right) \xrightarrow{\mathrm{d}} N(0,1). \tag{4.49}$$

Recalling $0 = \mathbf{1}^{\mathsf{T}}e(0) = \mathbf{1}^{\mathsf{T}}e_1(0) + \mathbf{1}^{\mathsf{T}}e_0(0)$, we obtain that

$$\begin{aligned}
&\frac{n^{1/2}}{n_1}\mathbf{1}^{\mathsf{T}}e_1(1) - \frac{n^{1/2}}{n_0}\mathbf{1}^{\mathsf{T}}e_0(0) = \frac{n^{1/2}}{n_1}\mathbf{1}^{\mathsf{T}}e_1(1) + \frac{n^{1/2}}{n_0}\mathbf{1}^{\mathsf{T}}e_1(0) \\
= &\sum_{i\in\mathcal{T}_t}\left(\frac{n^{1/2}}{n_1}e_i(1) + \frac{n^{1/2}}{n_0}e_i(0)\right) \equiv \sum_{i\in\mathcal{T}_t} w_i,
\end{aligned} \tag{4.50}$$

where $w_i = \frac{n^{1/2}}{n_1}e_i(1) + \frac{n^{1/2}}{n_0}e_i(0)$. Based on (4.12), we can verify that

$$S_w^2 \equiv \sum_{i=1}^n (w_i - \bar{w})^2 = \sum_{i=1}^n w_i^2 = n\sum_{i=1}^n \left(\frac{e_i(1)}{n_1} + \frac{e_i(0)}{n_0}\right)^2 = \frac{n^2}{n_1 n_0}\sigma_n^2.$$

Applying Lemma 4.6.2 to the representation (4.50), we have

$$d_{\mathrm{K}}\left(\frac{n^{1/2}}{\sigma_n}\left(\frac{\mathbf{1}^{\mathsf{T}}e_1(1)}{n_1} - \frac{\mathbf{1}^{\mathsf{T}}e_0(0)}{n_0}\right), N(0,1)\right) = O\left(\frac{\max_{1\le i\le n}|w_i|}{S_w}\right).$$

Lemma 4.6.6 and Assumption 4 imply

$$S_w^{-1} = O\left(\mathcal{E}_2^{-1/2}\right), \quad \max_{1\le i\le n}|w_i| = O\left(\frac{\mathcal{E}_\infty}{n^{1/2}}\right) = o\left(\mathcal{E}_2^{1/2}\right).$$

Therefore, (4.49) holds because convergence in Kolmogorov distance implies weak convergence.

We then prove the asymptotic normalities of the two estimators. Corollary 4.3.4 and Lemma 4.6.6 imply

$$\begin{aligned}
&\frac{n^{1/2}(\hat{\tau}_{\mathrm{adj}} - \tau)}{\sigma_n} \\
=&\frac{n^{1/2}}{\sigma_n}\left(\frac{\mathbf{1}^{\mathsf{T}}e_1(1)}{n_1} - \frac{\mathbf{1}^{\mathsf{T}}e_0(0)}{n_0}\right) + O_{\mathbb{P}}\left(\frac{\sqrt{\mathcal{E}_2\kappa^2 p\log p}}{\sigma_n} + \frac{n^{1/2}\Delta}{\sigma_n} + \frac{\sqrt{\mathcal{E}_2\kappa}}{\sigma_n}\right) \\
=&\frac{n^{1/2}}{\sigma_n}\left(\frac{\mathbf{1}^{\mathsf{T}}e_1(1)}{n_1} - \frac{\mathbf{1}^{\mathsf{T}}e_0(0)}{n_0}\right) + O_{\mathbb{P}}\left(\sqrt{\kappa^2 p\log p} + \sqrt{\frac{n}{\mathcal{E}_2}}\Delta + \sqrt{\kappa}\right).
\end{aligned}$$

We complete the proof by noting that $\kappa = o(1)$ in (4.17) under Assumption 2. The same proof carries over to $\hat{\tau}_{\mathrm{adj}}^{\mathrm{de}}$. □

4.7.3 Proof of asymptotic conservatism of variance estimators

Proof of Theorem 4.3.9. First, we prove the result for $j = 0$. Recalling $\hat{e}_t = (\mathrm{I} - H_t)e_t(t)$, we have

$$\begin{aligned}&\frac{1}{n_t}\sum_{i\in\mathcal{T}_t}\hat{e}_i^2 = \frac{1}{n_t}e_t(t)^\mathsf{T}(\mathrm{I}-H_t)e_t(t)\\ =&\frac{1}{n_t}\sum_{i\in\mathcal{T}_t}e_i^2(t) - \left(\frac{X_t^\mathsf{T}e_t(t)}{n_t}\right)^\mathsf{T}\Sigma_t^{-1}\frac{X_t^\mathsf{T}e_t(t)}{n_t} \triangleq S_{t1} - S_{t2}.\end{aligned} \tag{4.51}$$

Lemma 4.6.8 and the fact $\kappa = o(1)$ in (4.17) together imply a bound for S_{t2}:

$$S_{t2} \le \left\|\Sigma_t^{-1}\right\|_{\mathrm{op}}\left\|\frac{X_t^\mathsf{T}e_t(t)}{n_t}\right\|_2^2 = O_{\mathbb{P}}\left(\mathcal{E}_2\kappa\right) = o_{\mathbb{P}}\left(\mathcal{E}_2\right). \tag{4.52}$$

The first term, S_{t1}, has mean $\mathbb{E}S_{t1} = n^{-1}\sum_{i=1}^n e_i^2(t)$ and variance

$$\begin{aligned}\mathrm{Var}(S_{t1}) &\le \frac{1}{n_t^2}\frac{n_1n_0}{n(n-1)}\sum_{i=1}^n e_i^4(t) &(4.53)\\ &\le \frac{n}{n_t^2}\mathcal{E}_\infty^2\mathcal{E}_2 = O\left(\frac{\mathcal{E}_\infty^2\mathcal{E}_2}{n}\right) &(4.54)\\ &= o_{\mathbb{P}}(\mathcal{E}_2^2), &(4.55)\end{aligned}$$

where (4.53) follows from Lemma 4.6.1, (4.54) follows from the definitions of $\mathcal{E}_2$ and $\mathcal{E}_\infty$ and Assumption 1, and (4.55) follows from Assumption 4 that $\mathcal{E}_\infty^2 = o(n\mathcal{E}_2)$. Therefore, applying Chebyshev's inequality, we obtain

$$S_{t1} = \mathbb{E}S_{t1} + O_{\mathbb{P}}\left(\sqrt{\mathrm{Var}(S_{t1})}\right) = \frac{1}{n}\sum_{i=1}^n e_i^2(t) + o_{\mathbb{P}}(\mathcal{E}_2). \tag{4.56}$$

Combining the bounds for S_{t1} in (4.56) and S_{t2} in (4.52), we have

$$\frac{1}{n_t}\sum_{i\in\mathcal{T}_t}\hat{e}_i^2 = \frac{1}{n}\sum_{i=1}^n e_i^2(t) + o_{\mathbb{P}}\left(\mathcal{E}_2\right). \tag{4.57}$$

Using the formula of $\hat{\sigma}^2$ in (4.13) and Assumption 1, we have

$$\begin{aligned}\hat{\sigma}^2_{\mathrm{HC0}} &= \frac{n}{n_1-1}\left(\frac{1}{n}\sum_{i=1}^n e_i^2(1) + o_{\mathbb{P}}\left(\mathcal{E}_2\right)\right) + \frac{n}{n_0-1}\left(\frac{1}{n}\sum_{i=1}^n e_i^2(0) + o_{\mathbb{P}}\left(\mathcal{E}_2\right)\right)\\ &= \frac{1}{n_1}\sum_{i=1}^n e_i^2(1) + \frac{1}{n_0}\sum_{i=1}^n e_i^2(0) + o_{\mathbb{P}}(\mathcal{E}_2).\end{aligned}$$

Using the formula of σ_n^2 in (4.11), we have

$$\hat{\sigma}_{\mathrm{HC0}}^2 \geq \sigma_n^2 + \frac{1}{n}\sum_{i=1}^{n}(e_i(1) - e_i(0))^2 + o_{\mathbb{P}}(\mathcal{E}_2) \geq \sigma_n^2 + o_{\mathbb{P}}(\mathcal{E}_2),$$

which, coupled with Lemma 4.6.6, implies that $\hat{\sigma}_{\mathrm{HC0}}^2/\sigma_n^2 \geq 1 + o_{\mathbb{P}}(1)$.

Next we prove that the $\hat{\sigma}_{\mathrm{HC}j}^2$'s are asymptotically equivalent. It suffices to show that

$$\min_{j=1,2,3}\min_{1\leq i\leq n}|\tilde{e}_{i,j}|/|\hat{e}_i| = 1 + o_{\mathbb{P}}(1). \tag{4.58}$$

The proof for $j = 1$ follows from $p/n = o(1)$ in (4.17). To prove (4.58) for $j = 2, 3$, we need to prove that $\max_{t=0,1}\max_{i\in\mathcal{T}_t} H_{t,ii} = o_{\mathbb{P}}(1)$. This follows from Lemma 4.6.9 and the fact that $\kappa = o(1)$ in (4.17):

$$\max_{i\in\mathcal{T}_t} H_{t,ii} = \max_{i\in\mathcal{T}_t} n_t^{-1} x_i^{\mathsf{T}}\Sigma_t^{-1}x_i = O_{\mathbb{P}}\left(n_t^{-1}\max_{1\leq i\leq n}\|x_i\|_2^2\right) = O_{\mathbb{P}}(\kappa).$$

□

Bibliography

Abadie, A., S. Athey, G. W. Imbens, and J. M. Wooldridge (2017). "Sampling-based vs. design-based uncertainty in regression analysis". In: *arXiv preprint arXiv:1706.01778*.

Adichie, J. N. (1967a). "Asymptotic efficiency of a class of non-parametric tests for regression parameters". In: *The Annals of Mathematical Statistics*, pp. 884–893.

— (1978). "Rank tests of sub-hypotheses in the general linear regression". In: *The Annals of Statistics* 6.5, pp. 1012–1026.

— (1984). "11 Rank tests in linear models". In: *Handbook of statistics* 4, pp. 229–257.

Adichie, J. N. (1967b). "Estimates of regression parameters based on rank tests". In: *The Annals of Mathematical Statistics*, pp. 894–904.

Akritas, M. G. (1990). "The rank transform method in some two-factor designs". In: *Journal of the American Statistical Association* 85.409, pp. 73–78.

Akritas, M. G. and S. Arnold (2000). "Asymptotics for analysis of variance when the number of levels is large". In: *Journal of the American Statistical association* 95.449, pp. 212–226.

Akritas, M. G. and S. F. Arnold (1994). "Fully nonparametric hypotheses for factorial designs I: Multivariate repeated measures designs". In: *Journal of the American Statistical Association* 89.425, pp. 336–343.

Akritas, M. G., S. F. Arnold, and E. Brunner (1997). "Nonparametric hypotheses and rank statistics for unbalanced factorial designs". In: *Journal of the American Statistical Association* 92.437, pp. 258–265.

Alimoradi, S. and A. M. E. Saleh (1998). "9 On some L-estimation in linear regression models". In: *Handbook of Statistics* 17, pp. 237–280.

Anatolyev, S. (2012). "Inference in regression models with many regressors". In: *Journal of Econometrics* 170.2, pp. 368–382.

Anatolyev, S. and P. Yaskov (2017). "Asymptotics of diagonal elements of projection matrices under many instruments/regressors". In: *Econometric Theory* 33, pp. 717–738.

Anderson, M. J. and J. Robinson (2001). "Permutation tests for linear models". In: *Australian & New Zealand Journal of Statistics* 43.1, pp. 75–88.

Anderson, T. W. (1962). *An introduction to multivariate statistical analysis*. Wiley New York.

Angrist, J., D. Lang, and P. Oreopoulos (2009). "Incentives and services for college achievement: Evidence from a randomized trial". In: *American Economic Journal: Applied Economics* 1, pp. 136–63.

Arnold, S. F. (1980). "Asymptotic validity of F tests for the ordinary linear model and the multiple correlation model". In: *Journal of the American Statistical Association* 75.372, pp. 890–894.

Aubuchon, J. C. and T. P. Hettmansperger (1984). "12 On the use of rank tests and estimates in the linear model". In: *Handbook of statistics* 4, pp. 259–274.

Bahr, B. von and C.-G. Esseen (1965). "Inequalities for the r th Absolute Moment of a Sum of Random Variables, $1 \leq r \leq 2$". In: *The Annals of Mathematical Statistics* 36, pp. 299–303.

Bai, Z. and Y. Wu (1994). "Limiting behavior of M-estimators of regression coefficients in high dimensional linear models I. scale dependent case". In: *Journal of Multivariate Analysis* 51.2, pp. 211–239.

Bai, Z. and Y. Yin (1993). "Limit of the smallest eigenvalue of a large dimensional sample covariance matrix". In: *The annals of Probability*, pp. 1275–1294.

Bai, Z. and J. W. Silverstein (2010). *Spectral analysis of large dimensional random matrices*. Vol. 20. Springer.

Baranchik, A. (1973). "Inadmissibility of maximum likelihood estimators in some multiple regression problems with three or more independent variables". In: *The Annals of Statistics*, pp. 312–321.

Barber, R. F., E. J. Candès, et al. (2015). "Controlling the false discovery rate via knockoffs". In: *The Annals of Statistics* 43.5, pp. 2055–2085.

Bardenet, R. and O.-A. Maillard (2015). "Concentration inequalities for sampling without replacement". In: *Bernoulli* 21.3, pp. 1361–1385.

Bartlett, M. (1935). "The effect of non-normality on the t distribution". In: *mathematical proceedings of the cambridge philosophical society*. Vol. 31. Cambridge University Press, pp. 223–231.

Bathke, A. C. and S. W. Harrar (2008). "Nonparametric methods in multivariate factorial designs for large number of factor levels". In: *Journal of Statistical planning and Inference* 138.3, pp. 588–610.

Bathke, A. and D. Lankowski (2005). "Rank procedures for a large number of treatments". In: *Journal of statistical planning and inference* 133.2, pp. 223–238.

Bean, D., P. J. Bickel, N. El Karoui, C. Lim, and B. Yu (2012). "Penalized robust regression in high-dimension". In: *Technical Report 813, Department of Statistics, UC Berkeley*.

Bean, D., P. J. Bickel, N. El Karoui, and B. Yu (2013). "Optimal M-estimation in high-dimensional regression". In: *Proceedings of the National Academy of Sciences* 110.36, pp. 14563–14568.

Benjamini, Y. (1983). "Is the t test really conservative when the parent distribution is long-tailed?" In: *Journal of the American Statistical Association* 78.383, pp. 645–654.

Berk, R., E. Pitkin, L. Brown, A. Buja, E. George, and L. Zhao (2013). "Covariance adjustments for the analysis of randomized field experiments". In: *Evaluation review* 37.3-4, pp. 170–196.

Berry, K. J., J. E. Johnston, and P. W. Mielke (2013). *A chronicle of permutation statistical methods. 1920-2000, and beyond*. Springer. DOI: 10.1007/978-3-319-02744-9.

Bhattacharya, R. N. and J. K. Ghosh (1978). "On the validity of the formal Edgeworth expansion". In: *Ann. Statist* 6.2, pp. 434–451.

Bickel, P. J. and E. L. Lehmann (1975). "Descriptive Statistics for Nonparametric Models II. Location". In: *The Annals of Statistics* 3.5, pp. 1045–1069.

Bickel, P. J. (1965). "On some robust estimates of location". In: *The Annals of Mathematical Statistics* 36.3, pp. 847–858.

— (1973). "On some analogues to linear combinations of order statistics in the linear model". In: *The Annals of Statistics*, pp. 597–616.

— (1975). "One-step Huber estimates in the linear model". In: *Journal of the American Statistical Association* 70.350, pp. 428–434.

Bickel, P. J. and K. A. Doksum (2015). *Mathematical Statistics: Basic Ideas and Selected Topics, volume I*. Vol. 117. CRC Press.

Bickel, P. J. and D. A. Freedman (1981). "Some asymptotic theory for the bootstrap". In: *The Annals of Statistics*, pp. 1196–1217.

— (1983a). "Bootstrapping regression models with many parameters". In: *Festschrift for Erich L. Lehmann*, pp. 28–48.

— (1983b). "Bootstrapping regression models with many parameters". In: *Festschrift for Erich L. Lehmann*, pp. 28–48.

Bickel, P. J. and A. Sakov (2008). "On the choice of m in the m out of n bootstrap and confidence bounds for extrema". In: *Statistica Sinica* 18.3, pp. 967–985.

Bikelis, A. (1969). "On the estimation of the remainder term in the central limit theorem for samples from finite populations". In: *Studia Sci. Math. Hungar* 4, pp. 345–354.

Bloniarz, A., H. Liu, C.-H. Zhang, J. S. Sekhon, and B. Yu (2016). "Lasso adjustments of treatment effect estimates in randomized experiments". In: *Proceedings of the National Academy of Sciences* 113, pp. 7383–7390.

Bobkov, S. G. (2004). "Concentration of normalized sums and a central limit theorem for noncorrelated random variables". In: *Annals of probability* 32, pp. 2884–2907.

Boos, D. D. (1992). "On generalized score tests". In: *The American Statistician* 46.4, pp. 327–333.

Boos, D. D. and C. Brownie (1995). "ANOVA and rank tests when the number of treatments is large". In: *Statistics & Probability Letters* 23.2, pp. 183–191.

Boucheron, S., G. Lugosi, and P. Massart (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press.

Box, G. E. (1953). "Non-normality and tests on variances". In: *Biometrika* 40.3/4, pp. 318–335.

Box, G. E. and S. L. Andersen (1955). "Permutation theory in the derivation of robust criteria and the study of departures from assumption". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 17.1, pp. 1–26.

Box, G. E. and G. S. Watson (1962). "Robustness to non-normality of regression tests". In: *Biometrika* 49.1-2, pp. 93–106.

Brown, B. and J. Maritz (1982). "Distribution-Free Methods in Regression". In: *Australian Journal of Statistics* 24.3, pp. 318–331.

Brown, G. W. and A. M. Mood (1951). "On median tests for linear hypotheses". In: *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*. The Regents of the University of California.

Brownie, C. and D. D. Boos (1994). "Type I error robustness of ANOVA and ANOVA on ranks when the number of treatments is large". In: *Biometrics*, pp. 542–549.

Brunner, E. and M. Denker (1994). "Rank statistics under dependent observations and applications to factorial designs". In: *Journal of Statistical planning and Inference* 42.3, pp. 353–378.

Cai, T. T. and Z. Guo (2017). "Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity". In: *The Annals of statistics* 45, pp. 615–646.

Calhoun, G. (2011). "Hypothesis testing in linear regression when k/n is large". In: *Journal of econometrics* 165.2, pp. 163–174.

Campi, M. C., S. Ko, and E. Weyer (2009). "Non-asymptotic confidence regions for model parameters in the presence of unmodelled dynamics". In: *Automatica* 45.10, pp. 2175–2186.

Campi, M. C. and E. Weyer (2005). "Guaranteed non-asymptotic confidence regions in system identification". In: *Automatica* 41.10, pp. 1751–1764.

Cattaneo, M. D., M. Jansson, and W. K. Newey (2018). "Inference in linear regression models with many covariates and heteroscedasticity". In: *Journal of the American Statistical Association* 113.523, pp. 1350–1361.

Chatterjee, S. B. (1999). "Generalised bootstrap techniques". PhD thesis. Indian Statistical Institute, Kolkata.

Chatterjee, S. (2009). "Fluctuations of eigenvalues and second order Poincaré inequalities". In: *Probability Theory and Related Fields* 143.1-2, pp. 1–40.

Chernoff, H. (1956). "Large-sample theory: Parametric case". In: *The Annals of Mathematical Statistics* 27.1, pp. 1–22.

— (1981). "A note on an inequality involving the normal distribution". In: *The Annals of Probability*, pp. 533–535.

Chernozhukov, V., C. Hansen, and M. Jansson (2009). "Finite sample inference for quantile regression models". In: *Journal of Econometrics* 152.2, pp. 93–103.

Chung, E. and J. P. Romano (2013). "Exact and asymptotically robust permutation tests". In: *The Annals of Statistics* 41.2, pp. 484–507.

Cizek, P., W. K. Härdle, and R. Weron (2005). *Statistical tools for finance and insurance*. Springer Science & Business Media.

Cochran, W. G. (1937). "The Efficiencies of the Binomial Series Tests of Significance of a Mean and of a Correlation Coefficient." In: *Journal of the Royal Statistical Society* 100.1, pp. 69–73.

— (1977). *Sampling Techniques*. John Wiley & Sons.

— (2007). *Sampling Techniques*. 3rd. New York: John Wiley & Sons.

Collins, M. F. (1987). "A permutation test for planar regression". In: *Australian Journal of Statistics* 29.3, pp. 303–308.

Conover, W. J. and R. L. Iman (1981). "Rank transformations as a bridge between parametric and nonparametric statistics". In: *The American Statistician* 35.3, pp. 124–129.

Conover, W. and R. L. Iman (1976). "On some alternative procedures using ranks for the analysis of experimental designs". In: *Communications in Statistics-Theory and Methods* 5.14, pp. 1349–1368.

Cornfield, J. (1944). "On samples from finite populations". In: *Journal of the American Statistical Association* 39.226, pp. 236–239.

Cortes, C., M. Mohri, D. Pechyony, and A. Rastogi (2009). "Stability analysis and learning bounds for transductive regression algorithms". In: *arXiv preprint arXiv:0904.0814*.

Cressie, N. (1980). "Relaxing assumptions in the one sample t-test". In: *Australian Journal of Statistics* 22.2, pp. 143–153.

Daniels, H. (1954). "A distribution-free test for regression parameters". In: *The Annals of Mathematical Statistics*, pp. 499–513.

Das, D. and S. N. Lahiri (2019). "Second order correctness of perturbation bootstrap M-estimator of multiple linear regression parameter". In: *Bernoulli* 25.1, pp. 654–682.

David, F. N. and N. Johnson (1951a). "The effect of non-normality on the power function of the F-test in the analysis of variance". In: *Biometrika* 38.1/2, pp. 43–57.

David, F. and N. Johnson (1951b). "A method of investigating the effect of nonnormality and heterogeneity of variance on tests of the general linear hypothesis". In: *The Annals of Mathematical Statistics*, pp. 382–392.

David, H. A. and H. N. Nagaraja (1981). *Order statistics*. Wiley Online Library.

Dehejia, R. H. and S. Wahba (1999). "Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs". In: *Journal of the American Statistical Association* 94, pp. 1053–1062.

Diaconis, P. and M. Shahshahani (1987). "Time to reach stationarity in the Bernoulli–Laplace diffusion model". In: *SIAM Journal on Mathematical Analysis* 18, pp. 208–218.

DiCiccio, C. J. and J. P. Romano (2017). "Robust permutation tests for correlation and regression coefficients". In: *Journal of the American Statistical Association* 112.519, pp. 1211–1220.

Donoho, D. L. and X. Huo (2001). "Uncertainty principles and ideal atomic decomposition". In: *IEEE transactions on information theory* 47, pp. 2845–2862.

Donoho, D. and A. Montanari (2016). "High dimensional robust m-estimation: Asymptotic variance via approximate message passing". In: *Probability Theory and Related Fields* 166, pp. 935–969.

Donoho, D. and A. Montanari (2015). "Variance breakdown of Huber (M)-estimators: $n/p \to m \in (1, \infty)$. arXiv preprint". In: *arXiv preprint arXiv:1503.02106*.

Doob, J. L. (1935). "The limiting distributions of certain statistics". In: *The Annals of Mathematical Statistics* 6.3, pp. 160–169.

Draper, D. (1988). "Rank-based robust analysis of linear models. I. Exposition and review". In: *Statistical Science*, pp. 239–257.

Durrett, R. (2010). *Probability: theory and examples*. Cambridge university press.

Eden, T. and F. Yates (1933). "On the validity of Fisher's z test when applied to an actual example of non-normal data". In: *The Journal of Agricultural Science* 23.1, pp. 6–17.

Eeden, C. van (1972). "An analogue, for signed rank statistics, of Jureckova's asymptotic linearity theorem for rank statistics". In: *The Annals of Mathematical Statistics* 43.3, pp. 791–802.

Efron, B. (1979). "Bootstrap Methods: Another Look at the Jackknife". In: *The Annals of Statistics* 7.1, pp. 1–26.

Efron, B. (1969). "Student's t-test under symmetry conditions". In: *Journal of the American Statistical Association* 64.328, pp. 1278–1302.

Efron, B. and B. Efron (1982). *The jackknife, the bootstrap and other resampling plans*. Vol. 38. SIAM.

Eicker, F. (1963). "Asymptotic normality and consistency of the least squares estimators for families of linear regressions". In: *The Annals of Mathematical Statistics* 34.2, pp. 447–456.

— (1967). "Limit theorems for regressions with unequal and dependent errors". In: *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. Vol. 1, pp. 59–82.

El Karoui, N. (2009). "Concentration of measure and spectra of random matrices: applications to correlation matrices, elliptical distributions and beyond". In: *The Annals of Applied Probability* 19.6, pp. 2362–2405.

— (2010). "High-dimensionality effects in the Markowitz problem and other quadratic programs with linear constraints: Risk underestimation". In: *The Annals of Statistics* 38.6, pp. 3487–3566.

— (2013). "Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators: rigorous results". In: *arXiv preprint arXiv:1311.2445*.

— (2015). "On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators". In: *Technical Report 826, Department of Statistics, UC Berkeley*.

— (2018). "On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators". In: *Probability Theory and Related Fields* 170.1-2, pp. 95–175.

El Karoui, N., D. Bean, P. J. Bickel, C. Lim, and B. Yu (2013). "On robust regression with high-dimensional predictors". In: *Proceedings of the National Academy of Sciences* 110.36, pp. 14557–14562.

El Karoui, N., D. Bean, P. J. Bickel, C. Lim, and B. Yu (2011). "On robust regression with high-dimensional predictors". In: *Technical Report 811, Department of Statistics, UC Berkeley*.

El Karoui, N. and E. Purdom (2015). "Can we trust the bootstrap in high-dimension?" In: *Technical Report 824, Department of Statistics, UC Berkeley*.

— (2018). "Can we trust the bootstrap in high-dimensions? the case of linear models". In: *The Journal of Machine Learning Research* 19.1, pp. 170–235.

Esseen, C.-G. (1945a). "Fourier analysis of distribution functions. A mathematical study of the Laplace-Gaussian law". In: *Acta Mathematica* 77.1, pp. 1–125.
— (1945b). "Fourier analysis of distribution functions. A mathematical study of the Laplace-Gaussian law". In: *Acta Mathematica* 77.1, pp. 1–125.
Evans, R. D. and R. Evans (1955). *Appendix G: The atomic nucleus*. McGraw-Hill New York.
Feng, L., C. Zou, Z. Wang, and B. Chen (2013). "Rank-based score tests for high-dimensional regression coefficients". In: *Electronic Journal of Statistics* 7, pp. 2131–2149.
Feng, X., X. He, and J. Hu (2011). "Wild bootstrap for quantile regression". In: *Biometrika* 98.4, pp. 995–999.
Fisher, R. A. (1915). "Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population". In: *Biometrika* 10.4, pp. 507–521.
— (1922). "The goodness of fit of regression formulae, and the distribution of regression coefficients". In: *Journal of the Royal Statistical Society* 85.4, pp. 597–612.
— (1926). "The arrangement of field experiments". In: *Journal of the Ministry of Agriculture* 33, pp. 503–513.
— (1935a). "The logic of inductive inference". In: *Journal of the royal statistical society* 98.1, pp. 39–82.
Fisher, R. A. (1924). "036: On a Distribution Yielding the Error Functions of Several Well Known Statistics." In: *Proceedings of the International Congress of Mathematics* 2, pp. 805–813.
— (1925). *Statistical methods for research workers*. Oliver, Boyd, Edinburgh, and London.
— (1935b). *The Design of Experiments*. 1st. Edinburgh: Oliver and Boyd.
Fogarty, C. B. (2018). "Regression assisted inference for the average treatment effect in paired experiments". In: *Biometrika*, in press.
Fogel, F., R. Jenatton, F. Bach, and A. d'Aspremont (2013). "Convex relaxations for permutation problems". In: *Advances in Neural Information Processing Systems*, pp. 1016–1024.
Freedman, D. A. (1981). "Bootstrapping regression models". In: *The Annals of Statistics* 9.6, pp. 1218–1228.
— (2008a). "On regression adjustments in experiments with several treatments". In: *The annals of applied statistics*, pp. 176–196.
— (2008b). "On regression adjustments to experimental data". In: *Advances in Applied Mathematics* 40, pp. 180–193.
Freedman, D. A. and D. Lane (1983). "A nonstochastic interpretation of reported significance levels". In: *Journal of Business & Economic Statistics* 1.4, pp. 292–298.
Friedman, M. (1937). "The use of ranks to avoid the assumption of normality implicit in the analysis of variance". In: *Journal of the american statistical association* 32.200, pp. 675–701.
— (1940). "A comparison of alternative tests of significance for the problem of m rankings". In: *The Annals of Mathematical Statistics* 11.1, pp. 86–92.
Galton, F. (1894). *Natural inheritance*. Macmillan and Company.

Gastwirth, J. L. (1966). "On robust procedures". In: *Journal of the American Statistical Association* 61.316, pp. 929–948.

Gayen, A. K. (1949). "The distribution of Student's t in random samples of any size drawn from non-normal universes". In: *Biometrika* 36.3/4, pp. 353–369.

— (1950). "The distribution of the variance ratio in random samples of any size drawn from non-normal universes". In: *Biometrika* 37.3/4, pp. 236–255.

Geary, R. (1927). "Some properties of correlation and regression in a limited universe". In: *Metron* 7, pp. 83–119.

Geary, R. C. (1947). "Testing for normality". In: *Biometrika* 34.3/4, pp. 209–242.

Geman, S. (1980). "A limit theorem for the norm of random matrices". In: *The Annals of Probability*, pp. 252–261.

Gutenbrunner, C. and J. Jurecková (1992). "Regression quantile and regression rank score process in the linear model and derived statistics". In: *Annals of Statistics* 20, pp. 305–330.

Gutenbrunner, C., J. Jurečková, R. Koenker, and S. Portnoy (1993). "Tests of linear hypotheses based on regression rank scores". In: *Journaltitle of Nonparametric Statistics* 2.4, pp. 307–331.

Hájek, J. (1960). "Limiting distributions in simple random sampling from a finite population". In: *Publications of the Mathematics Institute of the Hungarian Academy of Science* 5, pp. 361–74.

Hájek, J. (1962). "Asymptotically most powerful rank-order tests". In: *The Annals of Mathematical Statistics*, pp. 1124–1147.

Hájek, J. and Z. Šidák (1967). *Theory of rank tests. Academia.*

Hall, P. (1989). "Unusual properties of bootstrap confidence intervals in regression problems". In: *Probability Theory and Related Fields* 81.2, pp. 247–273.

— (1992). *The bootstrap and Edgeworth expansion.* Springer Science & Business Media.

Hanson, D. L. and F. T. Wright (1971). "A bound on tail probabilities for quadratic forms in independent random variables". In: *The Annals of Mathematical Statistics* 42.3, pp. 1079–1083.

Hartigan, J. (1970). "Exact confidence intervals in regression problems with independent symmetric errors". In: *The Annals of Mathematical Statistics* 41.6, pp. 1992–1998.

Hastie, T., A. Montanari, S. Rosset, and R. J. Tibshirani (2019). "Surprises in High Dimensional Ridgeless Least Squares Interpolation". In: *arXiv preprint arXiv:1903.08560.*

Hastings, C., F. Mosteller, J. W. Tukey, and C. P. Winsor (1947). "Low moments for small samples: a comparative study of order statistics". In: *The Annals of Mathematical Statistics* 18.3, pp. 413–426.

Hettmansperger, T. P. and J. W. McKean (1978). "Statistical inference based on ranks". In: *Psychometrika* 43.1, pp. 69–79.

Hinkelmann, K. and O. Kempthorne (2007). *Design and Analysis of Experiments, Introduction to Experimental Design.* Vol. 1. New York: John Wiley & Sons.

Hinkley, D. V. (1977). "Jackknifing in unbalanced situations". In: *Technometrics* 19.3, pp. 285–292.

Hodges, J. L. and E. L. Lehmann (1962). "Rank methods for combination of independent experiments in analysis of variance". In: *The Annals of Mathematical Statistics* 33.2, pp. 482–497.

Hodges, J. L. and E. L. Lehmann (1963). "Estimates of location based on rank tests". In: *The Annals of Mathematical Statistics*, pp. 598–611.

Hoeffding, W. (1952). "The large-sample power of tests based on permutations of observations". In: *The Annals of Mathematical Statistics* 23.2, pp. 169–192.

— (1963). "Probability inequalities for sums of bounded random variables". In: *Journal of the American Statistical Association* 58, pp. 13–30.

Höglund, T. (1978). "Sampling from a finite population. A remainder term estimate". In: *Scandinavian Journal of Statistics* 5.1, pp. 69–71.

Horn, R. A. and C. R. Johnson (2012). *Matrix analysis*. Cambridge university press.

Hotelling, H. and M. R. Pabst (1936). "Rank correlation and tests of significance involving no assumption of normality". In: *The Annals of Mathematical Statistics* 7.1, pp. 29–43.

Hu, F. and J. D. Kalbfleisch (2000). "The estimating function bootstrap". In: *Canadian Journal of Statistics* 28.3, pp. 449–481.

Hu, F. and J. V. Zidek (1995). "A bootstrap based on the estimating equations of the linear model". In: *Biometrika* 82.2, pp. 263–275.

Huber, P. J. (1964). "Robust estimation of a location parameter". In: *The Annals of Mathematical Statistics* 35.1, pp. 73–101.

Huber, P. J. (1972). "The 1972 wald lecture robust statistics: A review". In: *The Annals of Mathematical Statistics*, pp. 1041–1067.

Huber, P. J. (1973a). "Robust regression: asymptotics, conjectures and Monte Carlo". In: *The Annals of Statistics*, pp. 799–821.

— (1973b). "Robust regression: asymptotics, conjectures and Monte Carlo". In: *The Annals of Statistics* 1.5, pp. 799–821.

— (1981). *Robust statistics*. John Wiley & Sons, Inc., New York.

Imbens, G. W. and D. B. Rubin (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Jaeckel, L. A. (1972). "Estimating regression coefficients by minimizing the dispersion of the residuals". In: *The Annals of Mathematical Statistics*, pp. 1449–1458.

Jensen, D. (1979). "Linear models without moments". In: *Biometrika* 66.3, pp. 611–617.

Jin, Z., Z. Ying, and L. Wei (2001). "A simple resampling method by perturbing the minimand". In: *Biometrika* 88.2, pp. 381–390.

Johnstone, I. M. (2001). "On the distribution of the largest eigenvalue in principal components analysis". In: *Annals of statistics*, pp. 295–327.

Johnstone, I. M. and P. F. Velleman (1985). "The resistant line and related regression methods". In: *Journal of the American Statistical Association* 80.392, pp. 1041–1054.

Jung, J. (1956). "On linear estimates defined by a continuous weight function". In: *Arkiv för matematik* 3.3, pp. 199–209.

Jureckova, J. (1983). "Winsorized least squares estimator and its M-estimator counterpart". In: *Contributions to Statistics: Essays in Honour of Norman L. Johnson*, pp. 237–245.

Jureckova, J. (1969). "Asymptotic linearity of a rank statistic in regression parameter". In: *The Annals of Mathematical Statistics* 40.6, pp. 1889–1900.

— (1971). "Nonparametric estimate of regression coefficients". In: *The Annals of Mathematical Statistics*, pp. 1328–1338.

— (1977). "Asymptotic Relations of M-Estimates and R-Estimates in Linear Regression Model". In: *The Annals of Statistics* 5.3, pp. 464–472.

Jurečková, J. (1984). "Regression quantiles and trimmed least squares estimator under a general design". In: *Kybernetika* 20.5, pp. 345–357.

Jurečkovà, J. and L. Klebanov (1997). "Inadmissibility of robust estimators with respect to L1 norm". In: *Lecture Notes-Monograph Series*, pp. 71–78.

Kallenberg, O. (2006). *Foundations of Modern Probability*. New York: Springer.

Kempthorne, O. (1952). *The Design and Analysis of Experiments*. Wiley.

Kendall, M. G. and B. B. Smith (1939). "The problem of m rankings." In: *Annals of mathematical statistics*.

Kennedy, F. E. (1995). "Randomization tests in econometrics". In: *Journal of Business & Economic Statistics* 13.1, pp. 85–94.

Kennedy, P. E. and B. S. Cade (1996). "Randomization tests for multiple regression". In: *Communications in Statistics-Simulation and Computation* 25.4, pp. 923–936.

Kildea, D. (1981). "Brown-Mood type median estimators for simple regression models". In: *The Annals of Statistics*, pp. 438–442.

Kline, P. and A. Santos (2012). "A score based approach to wild bootstrap inference". In: *Journal of Econometric Methods* 1.1, pp. 23–41.

Koenker, R. (1997). "8 Rank tests for linear models". In: *Handbook of statistics* 15, pp. 175–199.

Koenker, R. and G. Bassett (1978). "Regression quantiles". In: *Econometrica: journal of the Econometric Society*, pp. 33–50.

Koenker, R. and S. Portnoy (1987). "L-estimation for linear models". In: *Journal of the American statistical Association* 82.399, pp. 851–857.

Koenker, R. and Q. Zhao (1994). "L-estimatton for linear heteroscedastic models". In: *Journal of Nonparametric Statistics* 3.3-4, pp. 223–235.

Koul, H. L. (1970). "A class of ADF tests for subhypothesis in the multiple linear regression". In: *The Annals of Mathematical Statistics*, pp. 1273–1281.

Koul, H. L. (1969). "Asymptotic behavior of Wilcoxon type confidence regions in multiple linear regression". In: *The Annals of Mathematical Statistics* 40.6, pp. 1950–1979.

Kraft, C. H. and C. Van Eeden (1972). "Linearized rank estimates and signed-rank estimates for the general linear hypothesis". In: *The Annals of Mathematical Statistics* 43.1, pp. 42–57.

Kruskal, W. H. and W. A. Wallis (1952). "Use of ranks in one-criterion variance analysis". In: *Journal of the American statistical Association* 47.260, pp. 583–621.

Lahiri, S. N. (1992). "Bootstrapping M-estimators of a multiple linear regression parameter". In: *The Annals of Statistics*, pp. 1548–1570.

LaLonde, R. J. (1986). "Evaluating the econometric evaluations of training programs with experimental data". In: *The American Economic Review* 76, pp. 604–620.

Lancaster, J. and D. Quade (1985). "A nonparametric test for linear regression based on combining Kendall's tau with the sign test". In: *Journal of the American Statistical Association* 80.390, pp. 393–397.

Latała, R. (2005). "Some estimates of norms of random matrices". In: *Proceedings of the American Mathematical Society* 133.5, pp. 1273–1282.

Ledoux, M. (2001). *The concentration of measure phenomenon*. 89. American Mathematical Soc.

Lee, T.-Y. and H.-T. Yau (1998). "Logarithmic Sobolev inequality for some models of random walks". In: *The Annals of Probability* 26, pp. 1855–1873.

Lehmann, E. L. and J. P. Romano (2006). *Testing statistical hypotheses*. Springer Science & Business Media.

Lei, L. and P. J. Bickel (2019). "An Assumption-Free Exact Test For Fixed-Design Linear Models With Exchangeable Errors". In: *arXiv preprint arXiv:1907.06133*.

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

Lei, L., P. J. Bickel, and N. E. Karoui (2016). "Asymptotics For High Dimensional Regression M-Estimates: Fixed Design Results". In: *arXiv preprint arXiv:1612.06358*.

Lei, L. and P. Ding (2018). "Regression adjustment in completely randomized experiments with a diverging number of covariates". In: *arXiv preprint arXiv:1806.07585*.

Li, X. and P. Ding (2017). "General forms of finite population central limit theorems with applications to causal inference". In: *Journal of the American Statistical Association* 112, pp. 1759–1769.

Lin, W. (2013). "Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique". In: *The Annals of Applied Statistics* 7, pp. 295–318.

Litvak, A. E., A. Pajor, M. Rudelson, and N. Tomczak-Jaegermann (2005). "Smallest singular value of random matrices and geometry of random polytopes". In: *Advances in Mathematics* 195.2, pp. 491–523.

Liu, R. Y. (1988). "Bootstrap procedures under some non-iid models". In: *The Annals of Statistics* 16.4, pp. 1696–1708.

Liu, R. Y. and K. Singh (1992). "Efficiency and robustness in resampling". In: *The Annals of Statistics* 20.1, pp. 370–384.

Lloyd, E. (1952). "Least-squares estimation of location and scale parameters using order statistics". In: *Biometrika* 39.1/2, pp. 88–95.

MacKinnon, J. G. (2013). "Thirty years of heteroskedasticity-robust inference". In: *Recent advances and future directions in causality, prediction, and specification analysis*. Springer, pp. 437–461.

Mallows, C. (1972). "A note on asymptotic joint normality". In: *The Annals of Mathematical Statistics*, pp. 508–515.

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

— (1993). "Bootstrap and wild bootstrap for high dimensional linear models". In: *The annals of statistics* 21.1, pp. 255–285.

Manly, B. F. (1991). *Randomization, bootstrap and Monte Carlo methods in biology*. Chapman and Hall/CRC.

Mann, H. B. and A. Wald (1943). "On stochastic limit and order relationships". In: *The Annals of Mathematical Statistics* 14.3, pp. 217–226.

Marčenko, V. A. and L. A. Pastur (1967). "Distribution of eigenvalues for some sets of random matrices". In: *Mathematics of the USSR-Sbornik* 1.4, p. 457.

Markatou, M. and E. Ronchetti (1997). "3 Robust inference: The approach based on influence functions". In: *Handbook of statistics* 15, pp. 49–75.

Maxwell, J. C. (1860). "V. Illustrations of the dynamical theory of gases. Part I. On the motions and collisions of perfectly elastic spheres". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 19.124, pp. 19–32.

McKean, J. W. and T. P. Hettmansperger (1976). "Tests of hypotheses based on ranks in the general linear model". In: *Communications in statistics-theory and methods* 5.8, pp. 693–709.

— (1978). "A robust analysis of the general linear model based on one step R-estimates". In: *Biometrika* 65.3, pp. 571–579.

Mehra, K. and P. Sen (1969). "On a class of conditionally distribution-free tests for interactions in factorial experiments". In: *The Annals of Mathematical Statistics* 40.2, pp. 658–664.

Meinshausen, N. (2015). "Group bound: confidence intervals for groups of variables in sparse high dimensional regression without assumptions on the design". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 77.5, pp. 923–945.

Michalewicz, Z. (2013). *Genetic algorithms+ data structures= evolution programs*. Springer Science & Business Media.

Middleton, J. A. (2018). "A Unified Theory of Regression Adjustment for Design-based Inference". In: *arXiv preprint arXiv:1803.06011*.

Miller, R. G. (1974). "An unbalanced jackknife". In: *The Annals of Statistics*, pp. 880–891.

Mosteller, F. (1946). "On Some Useful" Inefficient" Statistics". In: *The Annals of Mathematical Statistics* 17.4, pp. 377–408.

Muirhead, R. J. (1982). *Aspects of multivariate statistical theory*. Vol. 197. John Wiley & Sons.

Mukerjee, R., T. Dasgupta, and D. B. Rubin (2018). "Using Standard Tools from Finite Population Sampling to Improve Causal Inference for Complex Experiments". In: *Journal of the American Statistical Association*, in press.

Navidi, W. (1989). "Edgeworth expansions for bootstrapping regression models". In: *The Annals of Statistics* 17.4, pp. 1472–1478.

Neyman, J. (1923/1990). "On the application of probability theory to agricultural experiments. Essay on principles. Section 9. Translated by Dabrowska, D. M. and Speed, T. P." In: *Statistical Science* 5, pp. 465–472.

— (1934). "On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection". In: *Journal of the Royal Statistical Society* 97.4, pp. 558–625.

— (1935). "Statistical problems in agricultural experimentation". In: *Supplement to the Journal of the Royal Statistical Society* 2, pp. 107–180.

— (1959). "Optimal asymptotic tests of composite hypotheses". In: *Probability and statsitics*, pp. 213–234.

Neyman, J. S. (1923). "On the application of probability theory to agricultural experiments. essay on principles. section 9. (translated and edited by dm dabrowska and tp speed, statistical science (1990), 5, 465-480)". In: *Annals of Agricultural Sciences* 10, pp. 1–51.

Oja, H. (1987). "On permutation tests in multiple regression and analysis of covariance problems". In: *Australian Journal of Statistics* 29.1, pp. 91–100.

Parzen, M., L. Wei, and Z. Ying (1994). "A resampling method based on pivotal estimating functions". In: *Biometrika* 81.2, pp. 341–350.

Pearson, E. S. (1929). "Some notes on sampling tests with two variables". In: *Biometrika*, pp. 337–360.

— (1931). "The analysis of variance in cases of non-normal variation". In: *Biometrika*, pp. 114–133.

Pearson, E. S. and N. Adyanthāya (1929). "The distribution of frequency constants in small samples from non-normal symmetrical and skew populations". In: *Biometrika* 21.1/4, pp. 259–286.

Pearson, E. and N. Please (1975). "Relation between the shape of population distribution and the robustness of four simple test statistics". In: *Biometrika* 62.2, pp. 223–241.

Pearson, K. (1907). *On further methods of determining correlation*. Dulau and Company.

Peddada, S. D. and G. Patwardhan (1992). "Jackknife variance estimators in linear models". In: *Biometrika* 79.3, pp. 654–657.

Pinelis, I. (1994). "Extremal probabilistic problems and Hotelling's T^2 test under a symmetry condition". In: *The Annals of Statistics* 22.1, pp. 357–368.

Pitkin, E., R. Berk, L. Brown, A. Buja, E. George, K. Zhang, and L. Zhao (2017). *An asymptotically powerful test for the average treatment effect*.

Pitman, E. J. G. (1937a). "Significance tests which may be applied to samples from any populations". In: *Supplement to the Journal of the Royal Statistical Society* 4.1, pp. 119–130.

— (1937b). "Significance tests which may be applied to samples from any populations. II. The correlation coefficient test". In: *Supplement to the Journal of the Royal Statistical Society* 4.2, pp. 225–232.

Pitman, E. J. G. (1938). "Significance tests which may be applied to samples from any populations: III. The analysis of variance test". In: *Biometrika* 29.3/4, pp. 322–335.

Pollard, D. (1991). "Asymptotics for least absolute deviation regression estimators". In: *Econometric Theory* 7.2, pp. 186–199.

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

— (1985). "Asymptotic behavior of M estimators of p regression parameters when p^2/n is large; II. Normal approximation". In: *The Annals of Statistics*, pp. 1403–1417.

— (1986). "On the central limit theorem in $\mathbb{R}^p$ when $p \to \infty$". In: *Probability theory and related fields* 73.4, pp. 571–583.

— (1987). "A central limit theorem applicable to robust regression estimators". In: *Journal of multivariate analysis* 22.1, pp. 24–50.

Portnoy, S. and R. Koenker (1989). "Adaptive L-estimation for linear models". In: *The Annals of Statistics* 17.1, pp. 362–381.

Posekany, A., K. Felsenstein, and P. Sykacek (2011). "Biological assessment of robust noise models in microarray data analysis". In: *Bioinformatics* 27.6, pp. 807–814.

Puri, M. L. and P. Sen (1973). "A note on asymptotically distribution free tests for subhypotheses in multiple linear regression". In: *The Annals of Statistics* 1.3, pp. 553–556.

Puri, M. L. and P. K. Sen (1966). "On a class of multivariate multisample rank-order tests". In: *Sankhyā: The Indian Journal of Statistics, Series A*, pp. 353–376.

Quade, D. (1979). "Regression analysis based on the signs of the residuals". In: *Journal of the American Statistical Association* 74.366a, pp. 411–417.

Quenouille, M. H. (1949). "Problems in plane sampling". In: *The Annals of Mathematical Statistics* 20.3, pp. 355–375.

— (1956). "Notes on bias in estimation". In: *Biometrika* 43.3/4, pp. 353–360.

Qumsiyeh, M. B. (1994). "Bootstrapping and empirical Edgeworth expansions in multiple linear regression models". In: *Communications in Statistics-Theory and Methods* 23.11, pp. 3227–3239.

Rao, C. R. and L. Zhao (1992). "Approximation to the distribution of M-estimates in linear models by randomly weighted bootstrap". In: *Sankhyā: The Indian Journal of Statistics, Series A*, pp. 323–331.

Relles, D. A. (1968). *Robust regression by modified least-squares.* Tech. rep. DTIC Document.

Romano, J. P. (1989). "Bootstrap and randomization tests of some nonparametric hypotheses". In: *The Annals of Statistics*, pp. 141–159.

— (1990). "On the behavior of randomization tests without a group invariance assumption". In: *Journal of the American Statistical Association* 85.411, pp. 686–692.

Rosenthal, H. P. (1970). "On the subspaces of $L^p(p > 2)$ spanned by sequences of independent random variables". In: *Israel Journal of Mathematics* 8.3, pp. 273–303.

Rousseeuw, P. J. (1984). "Least median of squares regression". In: *Journal of the American statistical association* 79.388, pp. 871–880.

Rousseeuw, P. J. and M. Hubert (1999). "Regression depth". In: *Journal of the American Statistical Association* 94.446, pp. 388–402.

Rubin, D. B. (1974). "Estimating causal effects of treatments in randomized and nonrandomized studies." In: *Journal of educational Psychology* 66.5, p. 688.

Rubin, D. B. (1981). "The bayesian bootstrap". In: *The annals of statistics*, pp. 130–134.
Rudelson, M. and R. Vershynin (2009). "Smallest singular value of a random rectangular matrix". In: *Communications on Pure and Applied Mathematics* 62.12, pp. 1707–1739.
— (2010). "Non-asymptotic theory of random matrices: extreme singular values". In: *arXiv preprint arXiv:1003.2990*.
— (2013). "Hanson-Wright inequality and sub-gaussian concentration". In: *Electron. Commun. Probab* 18.82, pp. 1–9.
Ruppert, D. and R. J. Carroll (1980). "Trimmed least squares estimation in the linear model". In: *Journal of the American Statistical Association* 75.372, pp. 828–838.
Särndal, C.-E., I. Thomsen, J. M. Hoem, D. Lindley, O. Barndorff-Nielsen, and T. Dalenius (1978). "Design-based and model-based inference in survey sampling [with discussion and reply]". In: *Scandinavian Journal of Statistics*, pp. 27–52.
Scheffe, H. (1999). *The analysis of variance*. Vol. 72. John Wiley & Sons.
Scheffé, H. (1959). *The Analysis of Variance*. New York: John Wiley & Sons.
Schrader, R. M. and T. P. Hettmansperger (1980). "Robust analysis of variance based upon a likelihood ratio criterion". In: *Biometrika* 67.1, pp. 93–101.
Sen, P. K. (1968a). "Estimates of the regression coefficient based on Kendall's tau". In: *Journal of the American statistical association* 63.324, pp. 1379–1389.
— (1968b). "On a class of aligned rank order tests in two-way layouts". In: *The Annals of Mathematical Statistics* 39.4, pp. 1115–1124.
— (1969). "On a class of rank order tests for the parallelism of several regression lines". In: *The Annals of Mathematical Statistics*, pp. 1668–1683.
— (1982). "On M Test in Linear Models". In: *Biometrika*, pp. 245–248.
Serfling, R. J. (1974). "Probability inequalities for the sum in sampling without replacement". In: *The Annals of Statistics* 2, pp. 39–48.
Shao, J. (1988). "On resampling methods for variance and bias estimation in linear models". In: *The Annals of Statistics*, pp. 986–1008.
— (1989). "Jackknifing weighted least squares estimators". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 51.1, pp. 139–156.
Shao, J. and C. Wu (1987). "Heteroscedasticity-robustness of jackknife variance estimators in linear models". In: *The Annals of Statistics*, pp. 1563–1579.
Shorack, G. R. (1982). "Bootstrapping robust regression". In: *Communications in Statistics-Theory and Methods* 11.9, pp. 961–972.
Siegel, A. F. (1982). "Robust regression using repeated medians". In: *Biometrika* 69.1, pp. 242–244.
Sievers, G. L. (1978). "Weighted rank statistics for simple linear regression". In: *Journal of the American Statistical Association* 73.363, pp. 628–631.
— (1983). "A weighted dispersion function for estimation in linear models". In: *Communications in Statistics-Theory and Methods* 12.10, pp. 1161–1179.
Silvapulle, M. J. (1992). "Robust tests of inequality constraints and one-sided hypotheses in the linear model". In: *Biometrika* 79.3, pp. 621–630.

Silverstein, J. W. (1985). "The smallest eigenvalue of a large dimensional Wishart matrix". In: *The Annals of Probability*, pp. 1364–1368.

Singer, J. M. and P. K. Sen (1985). "M-methods in multivariate linear models". In: *Journal of multivariate Analysis* 17.2, pp. 168–184.

Snedecor, G. W. (1934). *Calculation and interpretation of analysis of varianceand covariance*. Collegiate Press, Inc,; Ames Iowa.

Spearman, C. (1904). "The proof and measurement of association between two things". In: *American journal of Psychology* 15.1, pp. 72–101.

Srivastava, M. (1972). "Asymptotically most powerful rank tests for regression parameters in MANOVA". In: *Annals of the Institute of Statistical Mathematics* 24.1, pp. 285–297.

Stone, M. (1974). "Cross-validatory choice and assessment of statistical predictions". In: *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 111–147.

Student (1908a). "Probable error of a correlation coefficient". In: *Biometrika*, pp. 302–310.

— (1908b). "The probable error of a mean". In: *Biometrika*, pp. 1–25.

Sur, P. and E. J. Candès (2019). "A modern maximum-likelihood theory for high-dimensional logistic regression". In: *Proceedings of the National Academy of Sciences*, p. 201810420.

Sur, P., Y. Chen, and E. J. Candès (2017). "The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled chi-square". In: *Probability Theory and Related Fields*, pp. 1–72.

Ter Braak, C. J. (1992). "Permutation versus bootstrap significance tests in multiple regression and ANOVA". In: *Bootstrapping and related techniques*. Springer, pp. 79–85.

Theil, H. (1950a). "A rank-invariant method of linear and polynomial regression analysis, I". In: *Nederl. Akad. Wetensch. Proc.* Vol. 53, pp. 386–392.

— (1950b). "A rank-invariant method of linear and polynomial regression analysis, II". In: *Nederl. Akad. Wetensch. Proc.* Vol. 53, pp. 521–525.

— (1950c). "A rank-invariant method of linear and polynomial regression analysis, III". In: *Nederl. Akad. Wetensch. Proc.* Vol. 53, pp. 1397–1412.

Tikhomirov, K. (2017). "Sample covariance matrices of heavy-tailed distributions". In: *International Mathematics Research Notices*, in press.

Tolstikhin, I. (2017). "Concentration Inequalities for Samples without Replacement". In: *Theory of Probability and Its Applications* 61, pp. 462–481.

Tropp, J. A. (2016). "The expected norm of a sum of independent random matrices: An elementary approach". In: *High Dimensional Probability VII*. Springer, pp. 173–202.

Tsiatis, A. A., M. Davidian, M. Zhang, and X. Lu (2008). "Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: a principled yet flexible approach". In: *Statistics in Medicine* 27, pp. 4658–4677.

Tukey, J. (1958). "Bias and confidence in not quite large samples". In: *Ann. Math. Statist.* 29, p. 614.

Tukey, J. W. (1960). "A survey of sampling from contaminated distributions". In: *Contributions to probability and statistics*, pp. 448–485.

— (1962). "The future of data analysis". In: *The annals of mathematical statistics* 33.1, pp. 1–67.

Tyler, D. E. (1987). "A distribution-free M-estimator of multivariate scatter". In: *The Annals of Statistics*, pp. 234–251.

Van Aelst, S., P. J. Rousseeuw, M. Hubert, and A. Struyf (2002). "The deepest regression method". In: *Journal of Multivariate Analysis* 81.1, pp. 138–166.

Van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge university press.

Vershynin, R. (2010). "Introduction to the non-asymptotic analysis of random matrices". In: *arXiv preprint arXiv:1011.3027*.

Wachter, K. W. (1976). "Probability plotting points for principal components". In: *Ninth Interface Symposium Computer Science and Statistics*. Prindle, Weber and Schmidt, Boston, pp. 299–308.

— (1978). "The strong limits of random matrix spectra for sample matrices of independent elements". In: *The Annals of Probability*, pp. 1–18.

Wager, S., W. Du, J. Taylor, and R. J. Tibshirani (2016). "High-dimensional regression adjustments in randomized experiments". In: *Proceedings of the National Academy of Sciences* 113, pp. 12673–12678.

Wald, A. (1949). "Note on the consistency of the maximum likelihood estimate". In: *The Annals of Mathematical Statistics* 20.4, pp. 595–601.

Wallace, D. L. (1958). "Asymptotic approximations to distributions". In: *The Annals of Mathematical Statistics* 29.3, pp. 635–654.

Wang, H. and M. G. Akritas (2004). "Rank tests for ANOVA with large number of factor levels". In: *Journal of Nonparametric Statistics* 16.3-4, pp. 563–589.

Wasserman, L. and K. Roeder (2009). "High dimensional variable selection". In: *Annals of statistics* 37.5A, p. 2178.

Welch, B. L. (1937). "On the z-test in randomized blocks and Latin squares". In: *Biometrika* 29.1/2, pp. 21–52.

Welch, W. J. (1990). "Construction of permutation tests". In: *Journal of the American Statistical Association* 85.411, pp. 693–698.

Welsh, A. (1987). "One-Step L-Estimators for the Linear Model". In: *The Annals of Statistics* 15.2, pp. 626–641.

— (1989). "On M-processes and M-estimation". In: *The Annals of Statistics* 17.1, pp. 337–361.

— (1991). "Asymptotically Efficient Adaptive L-Estimators in Linear Models". In: *Statistica Sinica*, pp. 203–228.

Wilks, S. S. (1938). "The large-sample distribution of the likelihood ratio for testing composite hypotheses". In: *The Annals of Mathematical Statistics* 9.1, pp. 60–62.

Woodruff, D. P. (2014). "Sketching as a tool for numerical linear algebra". In: *Foundations and Trends® in Theoretical Computer Science* 10.1–2, pp. 1–157.

Wu, C. F. (1990). "On the asymptotic properties of the jackknife histogram". In: *The Annals of Statistics*, pp. 1438–1452.

Wu, C.-F. J. (1986). "Jackknife, bootstrap and other resampling methods in regression analysis". In: *the Annals of Statistics* 14.4, pp. 1261–1295.

El-Yaniv, R. and D. Pechyony (2009). "Transductive Rademacher complexity and its applications". In: *Journal of Artificial Intelligence Research* 35, p. 193.

Yaskov, P. (2014). "Lower bounds on the smallest eigenvalue of a sample covariance matrix". In: *Electronic Communications in Probability* 19, pp. 1–10.

Yohai, V. J. (1972). *Robust M estimates for the general linear model*. Universidad Nacional de la Plata. Departamento de Matematica.

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

— (1979b). "Asymptotic behavior of M-estimators for the linear model". In: *The Annals of Statistics*, pp. 258–268.

Zellner, A. (1976). "Bayesian and non-Bayesian analysis of the regression model with multivariate Student-t error terms". In: *Journal of the American Statistical Association* 71.354, pp. 400–405.

Zhong, P.-S. and S. X. Chen (2011). "Tests for high-dimensional regression coefficients with factorial designs". In: *Journal of the American Statistical Association* 106.493, pp. 260–274.

Appendix A

Appendix for Chapter 2

A.1 Proof Sketch of Lemma 2.4.5

In this Appendix, we provide a roadmap for proving Lemma 2.4.5 by considering a special case where X is one realization of a random matrix Z with i.i.d. mean-zero σ^2-sub-gaussian entries. Random matrix theory (Geman 1980; Silverstein 1985; Bai and Yin 1993) implies that $\lambda_+ = (1+\sqrt{\kappa})^2 + o_p(1) = O_p(1)$ and $\lambda_- = (1-\sqrt{\kappa})^2 + o_p(1) = \Omega_p(1)$. Thus, the assumption **A**3 is satisfied with high probability. Thus, the Lemma 2.4.4 in p. 28 holds with high probability. It remains to prove the following lemma to obtain Theorem 2.3.1.

Lemma A.1.1. *Let Z be a random matrix with i.i.d. mean-zero σ^2-sub-gaussian entries and X be one realization of Z. Then under assumptions **A**1 and **A**2,*

$$\max_{1\le j\le p} M_j = O_p\left(\frac{\text{polyLog}(n)}{n}\right), \quad \min_{1\le j\le p} \text{Var}(\hat{\beta}_j) = \Omega_p\left(\frac{1}{n\cdot\text{polyLog}(n)}\right),$$

where M_j is defined in (2.11) *in p.28 and the randomness in $o_p(\cdot)$ and $O_p(\cdot)$ comes from Z.*

Note that we prove in Proposition 2.3.4 that assump3tions **A**4 and **A**5 are satisfied with high probability in this case. However, we will not use them directly but prove Lemma A.1.1 from the scratch instead, in order to clarify why assump3tions in forms of **A**4 and **A**5 are needed in the proof.

A.1.1 Upper Bound of M_j

First by Proposition A.5.3,

$$\lambda_+ = O_p(1), \quad \lambda_- = \Omega_p(1).$$

In the rest of the proof, the symbol $\mathbb{E}$ and Var denotes the expectation and the variance conditional on Z. Let $\tilde{Z} = D^{\frac{1}{2}}Z$, then $M_j = \mathbb{E}\|e_j^T(\tilde{Z}^T\tilde{Z})^{-1}\tilde{Z}^T\|_\infty$. Let $\tilde{H}_j = I - \tilde{Z}_{[j]}(\tilde{Z}_{[j]}^T\tilde{Z}_{[j]})^{-1}\tilde{Z}_{[j]}^T$,

then by block matrix inversion formula (see Proposition A.5.1), which we state as Proposition A.5.1 in Appendix A.5.

$$\begin{aligned}(\tilde{Z}^T\tilde{Z})^{-1}\tilde{Z}^T &= \begin{pmatrix} \tilde{Z}_1^T\tilde{Z}_1 & \tilde{Z}_1^T\tilde{Z}_{[1]} \\ \tilde{Z}_{[1]}^T\tilde{Z}_1 & \tilde{Z}_{[1]}^T\tilde{Z}_{[1]} \end{pmatrix}^{-1} \begin{pmatrix} \tilde{Z}_1 \\ \tilde{Z}_{[1]} \end{pmatrix} \\ &= \frac{1}{\tilde{Z}_1^T(I-\tilde{H}_1)\tilde{Z}_1} \begin{pmatrix} 1 & -\tilde{Z}_1^T\tilde{Z}_{[1]}(\tilde{Z}_{[1]}^T\tilde{Z}_{[1]})^{-1} \\ * & * \end{pmatrix} \begin{pmatrix} \tilde{Z}_1 \\ \tilde{Z}_{[1]} \end{pmatrix} \\ &= \frac{1}{\tilde{Z}_1^T(I-\tilde{H}_1)\tilde{Z}_1} \begin{pmatrix} \tilde{Z}_1^T(I-\tilde{H}_1) \\ * \end{pmatrix}.\end{aligned}$$

This implies that

$$M_1 = \mathbb{E}\frac{\|\tilde{Z}_1^T(I-\tilde{H}_1)\|_\infty}{\tilde{Z}_1^T(I-\tilde{H}_1)\tilde{Z}_1}. \tag{A.1}$$

Since $Z^TDZ/n \succeq K_0\lambda_- I$, we have

$$\frac{1}{\tilde{Z}_1^T(I-\tilde{H}_1)\tilde{Z}_1} = e_1^T(\tilde{Z}^T\tilde{Z})^{-1}e_1 = e_1^T(Z^TDZ)^{-1}e_1 = \frac{1}{n}e_1^T\left(\frac{Z^TDZ}{n}\right)^{-1}e_1 \le \frac{1}{nK_0\lambda_-}$$

and we obtain a bound for M_1 as

$$M_1 \le \frac{\mathbb{E}\|\tilde{Z}_1^T(I-\tilde{H}_1)\|_\infty}{nK_0\lambda_-} = \frac{\mathbb{E}\|Z_1^TD^{\frac{1}{2}}(I-\tilde{H}_1)\|_\infty}{nK_0\lambda_-}.$$

Similarly,

$$M_j \le \frac{\mathbb{E}\|Z_j^TD^{\frac{1}{2}}(I-\tilde{H}_j)\|_\infty}{nK_0\lambda_-} = \frac{\mathbb{E}\|Z_j^TD^{\frac{1}{2}}(I-D^{\frac{1}{2}}Z_{[j]}^T(Z_{[j]}^TDZ_{[j]})^{-1}Z_{[j]}D^{\frac{1}{2}})\|_\infty}{nK_0\lambda_-}. \tag{A.2}$$

The vector in the numerator is a linear contrast of Z_j and Z_j has mean-zero i.i.d. sub-gaussian entries. For any fixed matrix $A \in \mathbb{R}^{n\times n}$, denote A_k by its k-th column, then $A_k^TZ_j$ is $\sigma^2\|A_k\|_2^2$-sub-gaussian (see Section 5.2.3 of Vershynin (2010) for a detailed discussion) and hence by definition of sub-Gaussianity,

$$P(|A_k^TZ_j| \ge \sigma\|A_k\|_2 t) \le 2e^{-\frac{t^2}{2}}.$$

Therefore, by a simple union bound, we conclude that

$$P(\|A^TZ_j\|_\infty \ge \sigma\max_k\|A_k\|_2 t) \le 2ne^{-\frac{t^2}{2}}.$$

Let $t = 2\sqrt{\log n}$,

$$P(\|A^TZ_j\|_\infty \ge 2\sigma\max_k\|A_k\|_2\sqrt{\log n}) \le \frac{2}{n} = o(1).$$

This entails that

$$\|A^T Z_j\|_\infty = O_p\left(\max_k \|A_k\|_2 \cdot \text{polyLog}(n)\right) = O_p\left(\|A\|_{\text{op}} \cdot \text{polyLog}(n)\right). \tag{A.3}$$

with high probability. In M_j, the coefficient matrix $(I - H_j)D^{\frac{1}{2}}$ depends on Z_j through D and hence we cannot use (A.3) directly. However, the dependence can be removed by replacing D by $D_{[j]}$ since $r_{i,[j]}$ does not depend on Z_j.

Since Z has i.i.d. sub-gaussian entries, no column is highly influential. In other words, the estimator will not change drastically after removing j-th column. This would suggest $R_i \approx r_{i,[j]}$. It is proved by El Karoui (2013) that

$$\sup_{i,j} |R_i - r_{i,[j]}| = O_p\left(\frac{\text{polyLog}(n)}{\sqrt{n}}\right).$$

It can be rigorously proved that

$$\left|\|Z_j^T D(I - \tilde{H}_j)\|_\infty - \|Z_j^T D_{[j]}(I - H_j)\|_\infty\right| = O_p\left(\frac{\text{polyLog}(n)}{n}\right),$$

where $H_j = I - D_{[j]}^{\frac{1}{2}} Z_{[j]}(Z_{[j]}^T D_{[j]} Z_{[j]})^{-1} Z_{[j]}^T D_{[j]}^{\frac{1}{2}}$; see Appendix A.1.1 for details. Since $D_{[j]}(I - H_j)$ is independent of Z_j and

$$\|D_{[j]}(I - H_j)\|_{\text{op}} \le \|D_{[j]}\|_{\text{op}} \le K_1 = O\left(\text{polyLog}(n)\right),$$

it follows from (A.2) and (A.3) that

$$\|Z_j^T D_{[j]}(I - H_j)\|_\infty = O_p\left(\frac{\text{polyLog}(n)}{n}\right).$$

In summary,

$$M_j = O_p\left(\frac{\text{polyLog}(n)}{n}\right). \tag{A.4}$$

A.1.2 Lower Bound of $\text{Var}(\hat{\beta}_j)$

Approximating $\text{Var}(\hat{\beta}_j)$ by $\text{Var}(b_j)$

It is shown by El Karoui (2013)[1] that

$$\hat{\beta}_j \approx b_j \triangleq \frac{1}{\sqrt{n}} \frac{N_j}{\xi_j} \tag{A.5}$$

[1] El Karoui (2013) considers a ridge regularized M estimator, which is different from our setting. However, this argument still holds in our case and proved in Appendix A.2.

where

$$N_j = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_{ij}\psi(r_{i,[j]}), \quad \xi_j = \frac{1}{n} Z_j^T (D_{[j]} - D_{[j]} Z_{[j]} (X_{[j]}^T D_{[j]} X_{[j]})^{-1} Z_{[j]}^T D_{[j]}) Z_j.$$

It has been shown by El Karoui (2013) that

$$\max_j |\hat{\beta}_j - b_j| = O_p\left(\frac{\text{polyLog}(n)}{n}\right).$$

Thus, $\text{Var}(\hat{\beta}_j) \approx \text{Var}(b_j)$ and a more refined calculation in Appendix A.1.2 shows that

$$|\,\text{Var}(\hat{\beta}_j) - \text{Var}(b_j)| = O_p\left(\frac{\text{polyLog}(n)}{n^{\frac{3}{2}}}\right).$$

It is left to show that

$$\text{Var}(b_j) = \Omega_p\left(\frac{1}{n \cdot \text{polyLog}(n)}\right). \tag{A.6}$$

Bounding $\text{Var}(b_j)$ **via** $\text{Var}(N_j)$

By definition of b_j,

$$\text{Var}(b_j) = \Omega_p\left(\frac{\text{polyLog}(n)}{n}\right) \Longleftrightarrow \text{Var}\left(\frac{N_j}{\xi_j}\right) = \Omega_p\left(\text{polyLog}(n)\right).$$

As will be shown in Appendix A.2.6,

$$\text{Var}(\xi_j) = O_p\left(\frac{\text{polyLog}(n)}{n}\right).$$

As a result, $\xi_j \approx \mathbb{E}\xi_j$ and

$$\text{Var}\left(\frac{N_j}{\xi_j}\right) \approx \text{Var}\left(\frac{N_j}{\mathbb{E}\xi_j}\right) = \frac{\text{Var}(N_j)}{(\mathbb{E}\xi_j)^2}.$$

As in the previous paper (El Karoui 2013), we rewrite ξ_j as

$$\xi_j = \frac{1}{n} Z_j^T D_{[j]}^{\frac{1}{2}} (I - D_{[j]}^{\frac{1}{2}} Z_{[j]} (X_{[j]}^T D_{[j]} X_{[j]})^{-1} Z_{[j]}^T D_{[j]}^{\frac{1}{2}}) D_{[j]}^{\frac{1}{2}} Z_j.$$

The middle matrix is idempotent and hence positive semi-definite. Thus,

$$\xi_j \leq \frac{1}{n} Z_j^T D_{[j]} Z_j \leq K_1 \lambda_+ = O_p\left(\text{polyLog}(n)\right).$$

Then we obtain that

$$\frac{\text{Var}(N_j)}{(\mathbb{E}\xi_j)^2} = \Omega_p\left(\frac{\text{Var}(N_j)}{\text{polyLog}(n)}\right),$$

and it is left to show that

$$\text{Var}(N_j) = \Omega_p\left(\frac{1}{\text{polyLog}(n)}\right). \tag{A.7}$$

Bounding $\mathrm{Var}(N_j)$ **via** $\mathrm{tr}(Q_j)$

Recall the definition of N_j (A.5), and that of Q_j (see Section 2.3.1 in p.16), we have

$$\mathrm{Var}(N_j) = \frac{1}{n} Z_j^T Q_j Z_j$$

Notice that Z_j is independent of $r_{i,[j]}$ and hence the conditional distribution of Z_j given Q_j remains the same as the marginal distribution of Z_j. Since Z_j has i.i.d. sub-gaussian entries, the Hanson-Wright inequality ((Hanson and Wright 1971; Rudelson and Vershynin 2013); see Proposition A.5.2), shown in Proposition A.5.2, implies that any quadratic form of Z_j, denoted by $Z_j^T Q_j Z_j$ is concentrated on its mean, i.e.

$$Z_j^T Q_j Z_j \approx \mathbb{E}_{Z_j,\epsilon} Z_j^T Q_j Z_j = (\mathbb{E} Z_{1j}^2) \cdot \mathrm{tr}(Q_j).$$

As a consequence, it is left to show that

$$\mathrm{tr}(Q_j) = \Omega_p \left(\frac{n}{\mathrm{polyLog}(n)} \right). \tag{A.8}$$

Lower Bound of $\mathrm{tr}(Q_j)$

By definition of Q_j,

$$\mathrm{tr}(Q_j) = \sum_{i=1}^{n} \mathrm{Var}(\psi(r_{i,[j]})).$$

To lower bounded the variance of $\psi(r_{i,[j]})$, recall that for any random variable W,

$$\mathrm{Var}(W) = \frac{1}{2} \mathbb{E}(W - W')^2. \tag{A.9}$$

where W' is an independent copy of W. Suppose $g : \mathbb{R} \to \mathbb{R}$ is a function such that $|g'(x)| \geq c$ for all x, then (A.9) implies that

$$\mathrm{Var}(g(W)) = \frac{1}{2} \mathbb{E}(g(W) - g(W'))^2 \geq \frac{c^2}{2} \mathbb{E}(W - W')^2 = c^2 \, \mathrm{Var}(W). \tag{A.10}$$

In other words, (A.10) entails that $\mathrm{Var}(W)$ is a lower bound for $\mathrm{Var}(g(W))$ provided that the derivative of g is bounded away from 0. As an application, we see that

$$\mathrm{Var}(\psi(r_{i,[j]})) \geq K_0^2 \, \mathrm{Var}(r_{i,[j]})$$

and hence

$$\mathrm{tr}(Q_j) \geq K_0^2 \sum_{i=1}^{n} \mathrm{Var}(r_{i,[j]}).$$

By the variance decomposition formula,

$$\mathrm{Var}(r_{i,[j]}) = \mathbb{E}\left(\mathrm{Var}\left(r_{i,[j]}\middle|\epsilon_{(i)}\right)\right) + \mathrm{Var}\left(\mathbb{E}\left(r_{i,[j]}\middle|\epsilon_{(i)}\right)\right) \geq \mathbb{E}\left(\mathrm{Var}\left(r_{i,[j]}\middle|\epsilon_{[i]}\right)\right),$$

where $\epsilon_{(i)}$ includes all but i-th entry of ϵ. Given $\epsilon_{(i)}$, $r_{i,[j]}$ is a function of ϵ_i. Using (A.10), we have

$$\mathrm{Var}(r_{i,[j]}|\epsilon_{(i)}) \geq \inf_{\epsilon_i}\left|\frac{\partial r_{i,[j]}}{\partial \epsilon_i}\right|^2 \cdot \mathrm{Var}(\epsilon_i|\epsilon_{(i)}) \geq \inf_{\epsilon_i}\left|\frac{\partial r_{i,[j]}}{\partial \epsilon_i}\right|^2 \cdot \mathrm{Var}(\epsilon_i).$$

This implies that

$$\mathrm{Var}(r_{i,[j]}) \geq \mathbb{E}\left(\mathrm{Var}\left(r_{i,[j]}\middle|\epsilon_{[i]}\right)\right) \geq \mathbb{E}\inf_{\epsilon}\left|\frac{\partial r_{i,[j]}}{\partial \epsilon_i}\right|^2 \cdot \min_i \mathrm{Var}(\epsilon_i).$$

Summing $\mathrm{Var}(r_{i,[j]})$ over $i = 1, \ldots, n$, we obtain that

$$\mathrm{tr}(Q_j) = \sum_{i=1}^{n}\mathrm{Var}(r_{i,[j]}) \geq \mathbb{E}\left(\sum_i \inf_{\epsilon}\left|\frac{\partial r_{i,[j]}}{\partial \epsilon_i}\right|^2\right) \cdot \min_i \mathrm{Var}(\epsilon_i).$$

It will be shown in Appendix A.2.6 that under assumptions **A**1-**A**3,

$$\mathbb{E}\sum_i \inf_{\epsilon}\left|\frac{\partial r_{i,[j]}}{\partial \epsilon_i}\right|^2 = \Omega_p\left(\frac{n}{\mathrm{polyLog}(n)}\right). \tag{A.11}$$

This proves (A.8) and as a result,

$$\min_j \mathrm{Var}(\hat{\beta}_j) = \Omega_p\left(\frac{1}{n \cdot \mathrm{polyLog}(n)}\right).$$

A.2 Proof of Theorem 2.3.1

A.2.1 Notation

To be self-contained, we summarize our notations in this subsection. The model we considered here is

$$y = X\beta^* + \epsilon$$

where $X \in \mathbb{R}^{n\times p}$ be the design matrix and ϵ is a random vector with independent entries. Notice that the target quantity $\frac{\hat{\beta}_j - \mathbb{E}\hat{\beta}_j}{\sqrt{\mathrm{Var}(\hat{\beta}_j)}}$ is shift invariant, we can assume $\beta^* = 0$ without loss of generality provided that X has full column rank; see Section 2.3.1 for details.

Let $x_i^T \in \mathbb{R}^{1\times p}$ denote the i-th row of X and $X_j \in \mathbb{R}^{n\times 1}$ denote the j-th column of X. Throughout the Chapter we will denote by $X_{ij} \in \mathbb{R}$ the (i,j)-th entry of X, by $X_{(i)} \in \mathbb{R}^{(n-1)\times p}$ the design matrix X after removing the i-th row, by $X_{[j]} \in \mathbb{R}^{n\times(p-1)}$ the design matrix X after removing the j-th column, by $X_{(i),[j]} \in \mathbb{R}^{(n-1)\times(p-1)}$ the design matrix after removing

both i-th row and j-th column, and by $x_{i,[j]} \in \mathbb{R}^{1\times(p-1)}$ the vector x_i after removing j-th entry. The M-estimator $\hat{\beta}$ associated with the loss function ρ is defined as

$$\hat{\beta} = \underset{\beta\in\mathbb{R}^p}{\arg\min} \frac{1}{n}\sum_{k=1}^{n} \rho(\epsilon_k - x_k^T\beta). \tag{A.12}$$

Similarly we define the leave-j-th-predictor-out version as

$$\hat{\beta}_{[j]} = \underset{\beta\in\mathbb{R}^p}{\arg\min} \frac{1}{n}\sum_{k=1}^{n} \rho(\epsilon_k - x_{k,[j]}^T\beta). \tag{A.13}$$

Based on these notation we define the full residual R_k as

$$R_k = \epsilon_k - x_k^T\hat{\beta}, \quad k = 1, 2, \ldots, n \tag{A.14}$$

the leave-j-th-predictor-out residual as

$$r_{k,[j]} = \epsilon_k - x_{k,[j]}^T\hat{\beta}_{[j]}, \quad k = 1, 2, \ldots, n,\ j \in J_n. \tag{A.15}$$

Four diagonal matrices are defined as

$$D = \mathrm{diag}(\psi'(R_k)), \quad \tilde{D} = \mathrm{diag}(\psi''(R_k)), \tag{A.16}$$

$$D_{[j]} = \mathrm{diag}(\psi'(r_{k,[j]})), \quad \tilde{D}_{[j]} = \mathrm{diag}(\psi''(r_{k,[j]})). \tag{A.17}$$

Further we define G and $G_{[j]}$ as

$$G = I - X(X^TDX)^{-1}X^TD, \quad G_{[j]} = I - X_{[j]}(X_{[j]}^TD_{[j]}X_{[j]})^{-1}X_{[j]}^TD_{[j]}. \tag{A.18}$$

Let J_n denote the indices of coefficients of interest. We say $a \in]a_1, a_2[$ if and only if $a \in [\min\{a_1, a_2\}, \max\{a_1, a_2\}]$. Regarding the technical assumptions, we need the following quantities

$$\lambda_+ = \lambda_{\max}\left(\frac{X^TX}{n}\right), \quad \lambda_- = \lambda_{\min}\left(\frac{X^TX}{n}\right) \tag{A.19}$$

be the largest (resp. smallest) eigenvalue of the matrix $\frac{X^TX}{n}$. Let $e_i \in \mathbb{R}^n$ be the i-th canonical basis vector and

$$h_{j,0} = (\psi(r_{1,[j]}), \ldots, \psi(r_{n,[j]}))^T, \quad h_{j,1,i} = G_{[j]}^Te_i. \tag{A.20}$$

Finally, let

$$\Delta_C = \max\left\{\max_{j\in J_n}\frac{|h_{j,0}^TX_j|}{\|h_{j,0}\|}, \max_{i\le n, j\in J_n}\frac{|h_{j,1,i}^TX_j|}{\|h_{j,1,i}\|}\right\}, \tag{A.21}$$

$$Q_j = \mathrm{Cov}(h_{j,0}). \tag{A.22}$$

We adopt Landau's notation $(O(\cdot), o(\cdot), O_p(\cdot), o_p(\cdot))$. In addition, we say $a_n = \Omega(b_n)$ if $b_n = O(a_n)$ and similarly, we say $a_n = \Omega_p(b_n)$ if $b_n = O_p(a_n)$. To simplify the logarithm factors, we use the symbol $\text{polyLog}(n)$ to denote any factor that can be upper bounded by $(\log n)^\gamma$ for some $\gamma > 0$. Similarly, we use $\frac{1}{\text{polyLog}(n)}$ to denote any factor that can be lower bounded by $\frac{1}{(\log n)^{\gamma'}}$ for some $\gamma' > 0$.

Finally we restate all the technical assumptions:

A1 $\rho(0) = \psi(0) = 0$ and there exists $K_0 = \Omega\left(\frac{1}{\text{polyLog}(n)}\right)$, $K_1, K_2 = O\left(\text{polyLog}(n)\right)$, such that for any $x \in \mathbb{R}$,

$$K_0 \le \psi'(x) \le K_1, \quad \left|\frac{d}{dx}(\sqrt{\psi'}(x))\right| = \frac{|\psi''(x)|}{\sqrt{\psi'(x)}} \le K_2;$$

A2 $\epsilon_i = u_i(W_i)$ where $(W_1, \ldots, W_n) \sim N(0, I_{n\times n})$ and u_i are smooth functions with $\|u_i'\|_\infty \le c_1$ and $\|u_i''\|_\infty \le c_2$ for some $c_1, c_2 = O(\text{polyLog}(n))$. Moreover, assume $\min_i \text{Var}(\epsilon_i) = \Omega\left(\frac{1}{\text{polyLog}(n)}\right)$.

A3 $\lambda_+ = O(\text{polyLog}(n))$ and $\lambda_- = \Omega\left(\frac{1}{\text{polyLog}(n)}\right)$;

A4 $\min_{j\in J_n} \frac{X_j^T Q_j X_j}{\text{tr}(Q_j)} = \Omega\left(\frac{1}{\text{polyLog}(n)}\right)$;

A5 $\mathbb{E}\Delta_C^8 = O\left(\text{polyLog}(n)\right)$.

A.2.2 Deterministic Approximation Results

In Appendix A.1, we use several approximations under random designs, e.g. $R_i \approx r_{i,[j]}$. To prove them, we follow the strategy of El Karoui (2013) which establishes the deterministic results and then apply the concentration inequalities to obtain high probability bounds. Note that $\hat{\beta}$ is the solution of

$$0 = f(\beta) \triangleq \frac{1}{n}\sum_{i=1}^n x_i \psi(\epsilon_i - x_i^T\beta),$$

we need the following key lemma to bound $\|\beta_1 - \beta_2\|_2$ by $\|f(\beta_1) - f(\beta_2)\|_2$, which can be calculated explicily.

Lemma A.2.1. *[El Karoui (2013), Proposition 2.1] For any β_1 and β_2,*

$$\|\beta_1 - \beta_2\|_2 \le \frac{1}{K_0\lambda_-}\|f(\beta_1) - f(\beta_2)\|_2.$$

Proof. By the mean value theorem, there exists $\nu_i \in]\epsilon_i - x_i^T\beta_1, \epsilon_i - x_i^T\beta_2[$ such that

$$\psi(\epsilon_i - x_i^T\beta_1) - \psi(\epsilon_i - x_i^T\beta_2) = \psi'(\nu_i)\cdot x_i^T(\beta_2 - \beta_1).$$

Then

$$\begin{aligned}\|f(\beta_1) - f(\beta_2)\|_2 &= \left\| \frac{1}{n} \sum_{i=1}^{n} \psi'(\nu_i) x_i x_i^T (\beta_1 - \beta_2) \right\|_2 \\ &\geq \lambda_{\min} \left(\frac{1}{n} \sum_{i=1}^{n} \psi'(\nu_i) x_i x_i^T \right) \cdot \|\beta_1 - \beta_2\|_2 \\ &\geq K_0 \lambda_- \|\beta_1 - \beta_2\|_2 .\end{aligned}$$

□

Based on Lemma A.2.1, we can derive the deterministic results informally stated in Appendix A.1. Such results are shown by El Karoui (2013) for ridge-penalized M-estimates and here we derive a refined version for unpenalized M-estimates. Throughout this subsection, we only assume assumption **A**1. This implies the following lemma,

Lemma A.2.2. *Under assumption **A**1, for any x and y,*

$$|\psi(x)| \leq K_1|x|, \quad |\sqrt{\psi'}(x) - \sqrt{\psi'}(y)| \leq K_2|x-y|,$$

and

$$|\psi'(x) - \psi'(y)| \leq 2\sqrt{K_1}K_2|x-y| \triangleq K_3|x-y|.$$

To state the result, we define the following quantities.

$$T = \frac{1}{\sqrt{n}} \max \left\{ \max_i \|x_i\|_2, \max_{j \in J_n} \|X_j\|_2 \right\}, \quad \mathcal{E} = \frac{1}{n} \sum_{i=1}^{n} \rho(\epsilon_i), \tag{A.23}$$

$$U = \left\| \frac{1}{n} \sum_{i=1}^{n} x_i (\psi(\epsilon_i) - \mathbb{E}\psi(\epsilon_i)) \right\|_2, \quad U_0 = \left\| \frac{1}{n} \sum_{i=1}^{n} x_i \mathbb{E}\psi(\epsilon_i) \right\|_2 . \tag{A.24}$$

The following proposition summarizes all deterministic results which we need in the proof.

Proposition A.2.3. *Under Assumption **A**1,*

(i) The norm of M estimator is bounded by

$$\|\hat{\beta}\|_2 \leq \frac{1}{K_0 \lambda_-}(U + U_0);$$

(ii) Define b_j as

$$b_j = \frac{1}{\sqrt{n}} \frac{N_j}{\xi_j}$$

where

$$N_j = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_{ij}\psi(r_{i,[j]}), \quad \xi_j = \frac{1}{n} X_j^T (D_{[j]} - D_{[j]}X_{[j]}(X_{[j]}^T D_{[j]} X_{[j]})^{-1} X_{[j]}^T D_{[j]}) X_j,$$

Then

$$\max_{j \in J_n} |b_j| \leq \frac{1}{\sqrt{n}} \cdot \frac{\sqrt{2K_1}}{K_0 \lambda_-} \cdot \Delta_C \cdot \sqrt{\mathcal{E}},$$

(iii) The difference between $\hat{\beta}_j$ and b_j is bounded by

$$\max_{j \in J_n} |\hat{\beta}_j - b_j| \leq \frac{1}{n} \cdot \frac{2K_1^2 K_3 \lambda_+ T}{K_0^4 \lambda_-^{\frac{7}{2}}} \cdot \Delta_C^3 \cdot \mathcal{E}.$$

(iv) The difference between the full and the leave-one-predictor-out residual is bounded by

$$\max_{j \in J_n} \max_i |R_i - r_{i,[j]}| \leq \frac{1}{\sqrt{n}} \left(\frac{2K_1^2 K_3 \lambda_+ T^2}{K_0^4 \lambda_-^{\frac{7}{2}}} \cdot \Delta_C^3 \cdot \mathcal{E} + \frac{\sqrt{2} K_1}{K_0^{\frac{3}{2}} \lambda_-} \cdot \Delta_C^2 \cdot \sqrt{\mathcal{E}} \right).$$

Proof. (i) By Lemma A.2.1,

$$\|\hat{\beta}\|_2 \leq \frac{1}{K_0 \lambda_-} \|f(\hat{\beta}) - f(0)\|_2 = \frac{\|f(0)\|_2}{K_0 \lambda_-},$$

since $\hat{\beta}$ is a zero of $f(\beta)$. By definition,

$$f(0) = \frac{1}{n} \sum_{i=1}^{n} x_i \psi(\epsilon_i) = \frac{1}{n} \sum_{i=1}^{n} x_i (\psi(\epsilon_i) - \mathbb{E}\psi(\epsilon_i)) + \frac{1}{n} \sum_{i=1}^{n} x_i \mathbb{E}\psi(\epsilon_i).$$

This implies that

$$\|f(0)\|_2 \leq U + U_0.$$

(ii) First we prove that

$$\xi_j \geq K_0 \lambda_-. \tag{A.25}$$

Since all diagonal entries of $D_{[j]}$ is lower bounded by K_0, we conclude that

$$\lambda_{\min} \left(\frac{X^T D_{[j]} X}{n} \right) \geq K_0 \lambda_-.$$

Note that ξ_j is the Schur's complement ((Horn and Johnson 2012), chapter 0.8) of $\frac{X^T D_{[j]} X}{n}$, we have

$$\xi_j^{-1} = e_j^T \left(\frac{X^T D_{[j]} X}{n} \right)^{-1} e_j \leq \frac{1}{K_0 \lambda_-},$$

which implies (A.25). As for N_j, we have

$$N_j = \frac{X_j^T h_{j,0}}{\sqrt{n}} = \frac{\|h_{j,0}\|_2}{\sqrt{n}} \cdot \frac{X_j^T h_{j,0}}{\|h_{j,0}\|_2}. \tag{A.26}$$

The the second term is bounded by Δ_C by definition, see (A.21). For the first term, the assumption **A**1 that $\psi'(x) \leq K_1$ implies that

$$\rho(x) = \rho(x) - \rho(0) = \int_0^x \psi(y)dy \geq \int_0^x \frac{\psi'(y)}{K_1} \cdot \psi(y)dy = \frac{1}{2K_1}\psi^2(x).$$

Here we use the fact that $\text{sign}(\psi(y)) = \text{sign}(y)$. Recall the definition of $h_{j,0}$, we obtain that

$$\frac{\|h_{j,0}\|_2}{\sqrt{n}} = \sqrt{\frac{\sum_{i=1}^n \psi(r_{i,[j]})^2}{n}} \leq \sqrt{2K_1} \cdot \sqrt{\frac{\sum_{i=1}^n \rho(r_{i,[j]})}{n}}.$$

Since $\hat{\beta}_{[j]}$ is the minimizer of the loss function $\sum_{i=1}^n \rho(\epsilon_i - x_{i,[j]}^T \beta_{[j]})$, it holds that

$$\frac{1}{n}\sum_{i=1}^n \rho(r_{i,[j]}) \leq \frac{1}{n}\sum_{i=1}^n \rho(\epsilon_i) = \mathcal{E}.$$

Putting together the pieces, we conclude that

$$|N_j| \leq \sqrt{2K_1} \cdot \Delta_C \sqrt{\mathcal{E}}. \tag{A.27}$$

By definition of b_j,

$$|b_j| \leq \frac{1}{\sqrt{n}} \cdot \frac{\sqrt{2K_1}}{K_0 \lambda_-} \Delta_C \sqrt{\mathcal{E}}.$$

(iii) The proof of this result is almost the same as El Karoui (2013). We state it here for the sake of completeness. Let $\tilde{\mathbf{b}}_\mathbf{j} \in \mathbb{R}^p$ with

$$(\tilde{\mathbf{b}}_\mathbf{j})_j = b_j, \quad (\tilde{\mathbf{b}}_\mathbf{j})_{[j]} = \hat{\beta}_{[j]} - b_j (X_{[j]}^T D_{[j]} X_{[j]})^{-1} X_{[j]}^T D_{[j]} X_j \tag{A.28}$$

where the subscript j denotes the j-th entry and the subscript $[j]$ denotes the sub-vector formed by all but j-th entry. Furthermore, define γ_j with

$$(\gamma_j)_j = -1, \quad (\gamma_j)_{[j]} = (X_{[j]}^T D_{[j]} X_{[j]})^{-1} X_{[j]}^T D_{[j]} X_j. \tag{A.29}$$

Then we can rewrite $\tilde{\mathbf{b}}_\mathbf{j}$ as

$$(\tilde{\mathbf{b}}_\mathbf{j})_j = -b_j (\gamma_j)_j, \quad (\tilde{\mathbf{b}}_\mathbf{j})_{[j]} = \hat{\beta}_{[j]} - b_j (\gamma_j)_{[j]}.$$

By definition of $\hat{\beta}_{[j]}$, we have $[f(\hat{\beta}_{[j]})]_{[j]} = 0$ and hence

$$[f(\tilde{\mathbf{b}}_\mathbf{j})]_{[j]} = [f(\tilde{\mathbf{b}}_\mathbf{j})]_{[j]} - [f(\hat{\beta}_{[j]})]_{[j]} = \frac{1}{n}\sum_{i=1}^{n} x_{i,[j]} \left[\psi(\epsilon_i - x_i^T\tilde{\mathbf{b}}_\mathbf{j}) - \psi(\epsilon_i - x_{i,[j]}^T\hat{\beta}_{[j]})\right]. \tag{A.30}$$

By mean value theorem, there exists $\nu_{i,j} \in]\epsilon_i - x_i^T\tilde{\mathbf{b}}_\mathbf{j}, \epsilon_i - x_{i,[j]}^T\hat{\beta}_{[j]}[$ such that

$$\begin{aligned}
&\psi(\epsilon_i - x_i^T\tilde{\mathbf{b}}_\mathbf{j}) - \psi(\epsilon_i - x_{i,[j]}^T\hat{\beta}_{[j]}) = \psi'(\nu_{i,j})(x_{i,[j]}^T\hat{\beta}_{[j]} - x_i^T\tilde{\mathbf{b}}_\mathbf{j}) \\
=&\psi'(\nu_{i,j})(x_{i,[j]}^T\hat{\beta}_{[j]} - x_{i,[j]}^T(\tilde{\mathbf{b}}_\mathbf{j})_{[j]} - X_{ij}b_j) \\
=&\psi'(\nu_{i,j}) \cdot b_j \cdot \left[x_{i,[j]}^T(X_{[j]}^T D_{[j]} X_{[j]})^{-1} X_{[j]}^T D_{[j]} X_j - X_{ij}\right]
\end{aligned}$$

Let

$$d_{i,j} = \psi'(\nu_{i,j}) - \psi'(r_{i,[j]}) \tag{A.31}$$

and plug the above result into (A.30), we obtain that

$$\begin{aligned}
[f(\tilde{\mathbf{b}}_\mathbf{j})]_{[j]} &= \frac{1}{n}\sum_{i=1}^{n} x_{i,[j]} \cdot \left(\psi'(r_{i,[j]}) + d_{i,j}\right) \cdot b_j \cdot \left[x_{i,[j]}^T(X_{[j]}^T D_{[j]} X_{[j]})^{-1} X_{[j]}^T D_{[j]} X_j - X_{ij}\right] \\
&= b_j \cdot \frac{1}{n}\sum_{i=1}^{n} \psi'(r_{i,[j]}) x_{i,[j]} \left[x_{i,[j]}^T(X_{[j]}^T D_{[j]} X_{[j]})^{-1} X_{[j]}^T D_{[j]} X_j - X_{ij}\right] \\
&\quad + b_j \cdot \frac{1}{n}\sum_{i=1}^{n} d_{i,j} x_{i,[j]} (x_{i,[j]}^T(X_{[j]}^T D_{[j]} X_{[j]})^{-1} X_{[j]}^T D_{[j]} X_j - X_{ij}) \\
&= b_j \cdot \frac{1}{n}\left[X_{[j]}^T D_{[j]} X_{[j]}(X_{[j]}^T D_{[j]} X_{[j]})^{-1} X_{[j]}^T D_{[j]} X_j - X_{[j]}^T D_{[j]} X_j\right] \\
&\quad + b_j \cdot \frac{1}{n}\sum_{i=1}^{n} d_{i,j} x_{i,[j]} \cdot x_i^T\gamma_j \\
&= b_j \cdot \frac{1}{n}\left(\sum_{i=1}^{n} d_{i,j} x_{i,[j]} x_i^T\right)\gamma_j.
\end{aligned}$$

Now we calculate $[f(\tilde{\mathbf{b}}_\mathbf{j})]_j$, the j-th entry of $f(\tilde{\mathbf{b}}_\mathbf{j})$. Note that

$$\begin{aligned}
&[f(\tilde{\mathbf{b}}_\mathbf{j})]_j = \frac{1}{n}\sum_{i=1}^{n} X_{ij}\psi(\epsilon_i - x_i^T\tilde{\mathbf{b}}_\mathbf{j}) \\
=&\frac{1}{n}\sum_{i=1}^{n} X_{ij}\psi(r_{i,[j]}) + b_j \cdot \frac{1}{n}\sum_{i=1}^{n} X_{ij}(\psi'(r_{i,[j]}) + d_{i,j}) \cdot \left[x_{i,[j]}^T(X_{[j]}^T D_{[j]} X_{[j]})^{-1} X_{[j]}^T D_{[j]} X_j - X_{ij}\right] \\
=&\frac{1}{n}\sum_{i=1}^{n} X_{ij}\psi(r_{i,[j]}) + b_j \cdot \frac{1}{n}\sum_{i=1}^{n} \psi'(r_{i,[j]}) X_{ij} \left[x_{i,[j]}^T(X_{[j]}^T D_{[j]} X_{[j]})^{-1} X_{[j]}^T D_{[j]} X_j - X_{ij}\right]
\end{aligned}$$

$$
\begin{aligned}
&+ b_j \cdot \left(\frac{1}{n} \sum_{i=1}^{n} d_{i,j} X_{ij} x_i^T \right) \gamma_j \\
=&\frac{1}{\sqrt{n}} N_j + b_j \cdot \left(\frac{1}{n} X_j^T D_{[j]} X_{[j]} (X_{[j]}^T D_{[j]} X_{[j]})^{-1} X_{[j]}^T D_{[j]} X_j - \frac{1}{n} \sum_{i=1}^{n} \psi'(r_{i,[j]}) X_{ij}^2 \right) \\
&+ b_j \cdot \left(\frac{1}{n} \sum_{i=1}^{n} d_{i,j} X_{ij} x_i^T \right) \gamma_j \\
=& \frac{1}{\sqrt{n}} N_j - b_j \cdot \xi_j + b_j \cdot \left(\frac{1}{n} \sum_{i=1}^{n} d_{i,j} X_{ij} x_i^T \right) \gamma_j \\
=& b_j \cdot \left(\frac{1}{n} \sum_{i=1}^{n} d_{i,j} X_{ij} x_i^T \right) \gamma_j
\end{aligned}
$$

where the second last line uses the definition of b_j. Putting the results together, we obtain that

$$
f(\tilde{\mathbf{b}}_\mathbf{j}) = b_j \cdot \left(\frac{1}{n} \sum_{i=1}^{n} d_{i,j} x_i x_i^T \right) \cdot \gamma_j.
$$

This entails that

$$
\|f(\tilde{\mathbf{b}}_\mathbf{j})\|_2 \leq |b_j| \cdot \max_i |d_{i,j}| \cdot \lambda_+ \cdot \|\gamma_j\|_2. \tag{A.32}
$$

Now we derive a bound for $\max_i |d_{i,j}|$, where $d_{i,j}$ is defined in (A.31). By Lemma A.2.2,

$$
|d_{i,j}| = |\psi'(\nu_{i,j}) - \psi'(r_{i,[j]})| \leq K_3 |\nu_{i,j} - r_{i,[j]}| = K_3 |x_{i,[j]}^T \hat{\beta}_{[j]} - x_i^T \tilde{\mathbf{b}}_\mathbf{j}|.
$$

By definition of $\tilde{\mathbf{b}}_\mathbf{j}$ and $h_{j,1,i}$,

$$
\begin{aligned}
&|x_{i,[j]}^T \hat{\beta}_{[j]} - x_i^T \tilde{\mathbf{b}}_\mathbf{j}| = |b_j| \cdot \left| x_{i,[j]}^T (X_{[j]}^T D_{[j]} X_{[j]})^{-1} X_{[j]}^T D_{[j]} X_j - X_{ij} \right| \\
=&|b_j| \cdot |e_i^T (I - X_{[j]} (X_{[j]}^T D_{[j]} X_{[j]})^{-1} X_{[j]}^T D_{[j]}) X_j| \\
=&|b_j| \cdot |h_{j,1,i}^T X_j| \leq |b_j| \cdot \Delta_C \left\| h_{j,1,i} \right\|_2,
\end{aligned} \tag{A.33}
$$

where the last inequality is derived by definition of Δ_C, see (A.21). Since $h_{j,1,i}$ is the i-th column of matrix $I - D_{[j]} X_{[j]} (X_{[j]}^T D_{[j]} X_{[j]})^{-1} X_{[j]}^T$, its L_2 norm is upper bounded by the operator norm of this matrix. Notice that

$$
I - D_{[j]} X_{[j]} (X_{[j]}^T D_{[j]} X_{[j]})^{-1} X_{[j]}^T = D_{[j]}^{\frac{1}{2}} \left(I - D_{[j]}^{\frac{1}{2}} X_{[j]} (X_{[j]}^T D_{[j]} X_{[j]})^{-1} X_{[j]}^T D_{[j]}^{\frac{1}{2}} \right) D_{[j]}^{-\frac{1}{2}}.
$$

The middle matrix in RHS of the displayed atom is an orthogonal projection matrix and hence

$$
\|I - D_{[j]} X_{[j]} (X_{[j]}^T D_{[j]} X_{[j]})^{-1} X_{[j]}^T\|_{\text{op}} \leq \|D_{[j]}^{\frac{1}{2}}\|_{\text{op}} \cdot \|D_{[j]}^{-\frac{1}{2}}\|_{\text{op}} \leq \left(\frac{K_1}{K_0} \right)^{\frac{1}{2}}. \tag{A.34}
$$

Therefore,

$$\max_{i,j} \|h_{j,1,i}\|_2 \leq \max_{j \in J_n} \|I - D_{[j]} X_{[j]} (X_{[j]}^T D_{[j]} X_{[j]})^{-1} X_{[j]}^T\|_{\text{op}} \leq \left(\frac{K_1}{K_0}\right)^{\frac{1}{2}}, \tag{A.35}$$

and thus

$$\max_i |d_{i,j}| \leq K_3 \sqrt{\frac{K_1}{K_0}} \cdot |b_j| \cdot \Delta_C. \tag{A.36}$$

As for γ_j, we have

$$\begin{aligned} K_0 \lambda_- \|\gamma_j\|_2^2 \leq & \gamma_j^T \left(\frac{X^T D_{[j]} X}{n}\right) \gamma_j \\ = & (\gamma_j)_j^2 \cdot \frac{X_j^T D_j X_j}{n} + (\gamma_j)_{[j]}^T \left(\frac{X_{[j]}^T D_{[j]} X_{[j]}}{n}\right) (\gamma_j)_{[j]} + 2\gamma_j \frac{X_j^T D_{[j]} X_{[j]}}{n} (\gamma_j)_{[j]} \end{aligned}$$

Recall the definition of γ_j in (A.29), we have

$$(\gamma_j)_{[j]}^T \left(\frac{X_{[j]}^T D_{[j]} X_{[j]}}{n}\right) (\gamma_j)_{[j]} = \frac{1}{n} X_j^T D_{[j]} X_{[j]} (X_{[j]}^T D_{[j]} X_{[j]})^{-1} X_{[j]}^T D_{[j]} X_j$$

and

$$\gamma_j \frac{X_j^T D_{[j]} X_{[j]}}{n} (\gamma_j)_{[j]} = -\frac{1}{n} X_j^T D_{[j]} X_{[j]} (X_{[j]}^T D_{[j]} X_{[j]})^{-1} X_{[j]}^T D_{[j]} X_j.$$

As a result,

$$\begin{aligned} K_0 \lambda_- \|\gamma_j\|_2^2 \leq & \frac{1}{n} X_j^T D_{[j]}^{\frac{1}{2}} (I - D_{[j]}^{\frac{1}{2}} X_{[j]} (X_{[j]}^T D_{[j]} X_{[j]})^{-1} X_{[j]}^T D_{[j]}^{\frac{1}{2}}) D_{[j]}^{\frac{1}{2}} X_j \\ \leq & \frac{\|D_{[j]}^{\frac{1}{2}} X_j\|_2^2}{n} \cdot \left\| I - D_{[j]}^{\frac{1}{2}} X_{[j]} (X_{[j]}^T D_{[j]} X_{[j]})^{-1} X_{[j]}^T D_{[j]}^{\frac{1}{2}} \right\|_{op} \\ \leq & \frac{\|D_{[j]}^{\frac{1}{2}} X_j\|_2^2}{n} \leq \frac{K_1 \|X_j\|_2^2}{n} \leq T^2 K_1, \end{aligned}$$

where T is defined in (A.23). Therefore we have

$$\|\gamma_j\|_2 \leq \sqrt{\frac{K_1}{K_0 \lambda_-}} T. \tag{A.37}$$

Putting (A.32), (A.36), (A.37) and part (ii) together, we obtain that

$$\|f(\tilde{\mathbf{b}}_\mathbf{j})\|_2 \leq \lambda_+ \cdot |b_j| \cdot K_3 \sqrt{\frac{K_1}{K_0}} \Delta_C |b_j| \cdot \sqrt{\frac{K_1}{K_0 \lambda_-}} T$$

$$\leq \lambda_+ \cdot \frac{1}{n}\frac{2K_1}{(K_0\lambda_-)^2}\Delta_C^2\mathcal{E} \cdot K_3\sqrt{\frac{K_1}{K_0}}\Delta_C \cdot \sqrt{\frac{K_1}{K_0\lambda_-}}T$$
$$= \frac{1}{n} \cdot \frac{2K_1^2K_3\lambda_+T}{K_0^3\lambda_-^{\frac{5}{2}}} \cdot \Delta_C^3 \cdot \mathcal{E}.$$

By Lemma A.2.1,

$$\|\hat{\beta} - \tilde{\mathbf{b}}_\mathbf{j}\|_2 \leq \frac{\|f(\hat{\beta}) - f(\tilde{\mathbf{b}}_\mathbf{j})\|_2}{K_0\lambda_-} = \frac{\|f(\tilde{\mathbf{b}}_\mathbf{j})\|_2}{K_0\lambda_-} \leq \frac{1}{n} \cdot \frac{2K_1^2K_3\lambda_+T}{K_0^4\lambda_-^{\frac{7}{2}}} \cdot \Delta_C^3 \cdot \mathcal{E}.$$

Since $\hat{\beta}_j - b_j$ is the j-th entry of $\hat{\beta} - \tilde{\mathbf{b}}_\mathbf{j}$, we have

$$|\hat{\beta}_j - b_j| \leq \|\hat{\beta} - \tilde{\mathbf{b}}_\mathbf{j}\|_2 \leq \frac{1}{n} \cdot \frac{2K_1^2K_3\lambda_+T}{K_0^4\lambda_-^{\frac{7}{2}}} \cdot \Delta_C^3 \cdot \mathcal{E}.$$

(iv) Similar to part (iii), this result has been shown by El Karoui (2013). Here we state a refined version for the sake of completeness. Let $\tilde{\mathbf{b}}_\mathbf{j}$ be defined as in (A.28), then

$$|R_i - r_{i,[j]}| = |x_i^T\hat{\beta} - x_{i,[j]}^T\hat{\beta}_{[j]}| = |x_i^T(\hat{\beta} - \tilde{\mathbf{b}}_\mathbf{j}) + x_i^T\tilde{\mathbf{b}}_\mathbf{j} - x_{i,[j]}^T\hat{\beta}_{[j]}|$$
$$\leq \|x_i\|_2 \cdot \|\hat{\beta} - \tilde{\mathbf{b}}_\mathbf{j}\|_2 + |x_i^T\tilde{\mathbf{b}}_\mathbf{j} - x_{i,[j]}^T\hat{\beta}_{[j]}|.$$

Note that $\|x_i\|_2 \leq \sqrt{n}T$, by part (iii), we have

$$\|x_i\|_2 \cdot \|\hat{\beta} - \tilde{\mathbf{b}}_\mathbf{j}\|_2 \leq \frac{1}{\sqrt{n}}\frac{2K_1^2K_3\lambda_+T^2}{K_0^4\lambda_-^{\frac{7}{2}}} \cdot \Delta_C^3 \cdot \mathcal{E}. \tag{A.38}$$

On the other hand, similar to (A.36), by (A.33),

$$|x_i^T\tilde{\mathbf{b}}_\mathbf{j} - x_{i,[j]}^T\hat{\beta}_{[j]}| \leq \sqrt{\frac{K_1}{K_0}} \cdot |b_j| \cdot \Delta_C \leq \frac{1}{\sqrt{n}} \cdot \frac{\sqrt{2}K_1}{K_0^{\frac{3}{2}}\lambda_-} \cdot \Delta_C^2 \cdot \sqrt{\mathcal{E}}. \tag{A.39}$$

Therefore,

$$|R_i - r_{i,[j]}| \leq \frac{1}{\sqrt{n}}\left(\frac{2K_1^2K_3\lambda_+T^2}{K_0^4\lambda_-^{\frac{7}{2}}} \cdot \Delta_C^3 \cdot \mathcal{E} + \frac{\sqrt{2}K_1}{K_0^{\frac{3}{2}}\lambda_-} \cdot \Delta_C^2 \cdot \sqrt{\mathcal{E}}\right).$$

□

A.2.3 Summary of Approximation Results

Under our technical assumptions, we can derive the rate for approximations via Proposition A.2.3. This justifies all approximations in Appendix A.1.

Theorem A.2.4. *Under the assumptions **A**1 - **A**5,*

(i)
$$T \leq \lambda_+ = O\left(\text{polyLog}(n)\right);$$

(ii)
$$\max_{j\in J_n} |\hat{\beta}_j| \leq \|\hat{\beta}\|_2 = O_{L^4}\left(\text{polyLog}(n)\right);$$

(iii)
$$\max_{j\in J_n} |b_j| = O_{L^2}\left(\frac{\text{polyLog}(n)}{\sqrt{n}}\right);$$

(iv)
$$\max_{j\in J_n} |\hat{\beta}_j - b_j| = O_{L^2}\left(\frac{\text{polyLog}(n)}{n}\right);$$

(v)
$$\max_{j\in J_n} \max_i |R_i - r_{i,[j]}| = O_{L^2}\left(\frac{\text{polyLog}(n)}{\sqrt{n}}\right).$$

Proof. (i) Notice that $X_j = Xe_j$, where e_j is the j-th canonical basis vector in $\mathbb{R}^p$, we have
$$\frac{\|X_j\|^2}{n} = e_j^T \frac{X^T X}{n} e_j \leq \lambda_+.$$
Similarly, consider the X^T instead of X, we conclude that
$$\frac{\|x_i\|^2}{n} \leq \lambda_{\max}\left(\frac{XX^T}{n}\right) = \lambda_+.$$
Recall the definition of T in (A.23), we conclude that
$$T \leq \sqrt{\lambda_+} = O\left(\text{polyLog}(n)\right).$$

(ii) Since $\epsilon_i = u_i(W_i)$ with $\|u_i'\|_\infty \leq c_1$, the gaussian concentration property ((Ledoux 2001), chapter 1.3) implies that ϵ_i is c_1^2-sub-gaussian and hence $\mathbb{E}|\epsilon_i|^k = O(c_1^k)$ for any finite $k > 0$. By Lemma A.2.2, $|\psi(\epsilon_i)| \leq K_1|\epsilon_i|$ and hence for any finite k,
$$\mathbb{E}|\psi(\epsilon_i)|^k \leq K_1^k \mathbb{E}|\epsilon_i|^k = O(c_1^k).$$

By part (i) of Proposition A.2.3, using the convexity of x^4 and hence $\left(\frac{a+b}{2}\right)^4 \leq \frac{a^4+b^4}{2}$,

$$\mathbb{E}\|\hat{\beta}\|_2^4 \leq \frac{1}{(K_0\lambda_-)^4}\mathbb{E}(U+U_0)^4 \leq \frac{8}{(K_0\lambda_-)^4}(\mathbb{E}U^4+U_0^4).$$

Recall (A.24) that $U = \left\|\frac{1}{n}\sum_{i=1}^n x_i(\psi(\epsilon_i)-\mathbb{E}\psi(\epsilon_i))\right\|_2$,

$$\begin{aligned}
U^4 = (U^2)^2 =& \frac{1}{n^4}\left(\sum_{i,i'=1}^n x_i^T x_{i'}(\psi(\epsilon_i)-\mathbb{E}\psi(\epsilon_i))(\psi(\epsilon_{i'})-\mathbb{E}\psi(\epsilon_{i'}))\right)^2 \\
=& \frac{1}{n^4}\left(\sum_{i=1}^n \|x_i\|_2^2(\psi(\epsilon_i)-\mathbb{E}\psi(\epsilon_i))^2 + \sum_{i\neq i'}|x_i^T x_{i'}|(\psi(\epsilon_i)-\mathbb{E}\psi(\epsilon_i))(\psi(\epsilon_{i'})-\mathbb{E}\psi(\epsilon_{i'}))\right)^2 \\
=& \frac{1}{n^4}\Bigg\{\sum_{i=1}^n \|x_i\|_2^4(\psi(\epsilon_i)-\mathbb{E}\psi(\epsilon_i))^4 \\
&+ \sum_{i\neq i'}(2|x_i^T x_{i'}|^2 + \|x_i\|_2^2\|x_{i'}\|_2^2)(\psi(\epsilon_i)-\mathbb{E}\psi(\epsilon_i))^2(\psi(\epsilon_{i'})-\mathbb{E}\psi(\epsilon_{i'}))^2 \\
&+ \sum_{\text{others}} |x_i^T x_{i'}|(\psi(\epsilon_i)-\mathbb{E}\psi(\epsilon_i))(\psi(\epsilon_{i'})-\mathbb{E}\psi(\epsilon_{i'})) \\
&\qquad \cdot |x_k^T x_{k'}|(\psi(\epsilon_k)-\mathbb{E}\psi(\epsilon_k))(\psi(\epsilon_{k'})-\mathbb{E}\psi(\epsilon_{k'}))\Bigg\}
\end{aligned}$$

Since $\psi(\epsilon_i)-\mathbb{E}\psi(\epsilon_i)$ has a zero mean, we have

$$\mathbb{E}(\psi(\epsilon_i)-\mathbb{E}\psi(\epsilon_i))(\psi(\epsilon_{i'})-\mathbb{E}\psi(\epsilon_{i'}))(\psi(\epsilon_k)-\mathbb{E}\psi(\epsilon_k))(\psi(\epsilon_{k'})-\mathbb{E}\psi(\epsilon_{k'})) = 0$$

for any $(i,i') \neq (k,k')$ or (k',k) and $i \neq i'$. As a consequence,

$$\begin{aligned}
\mathbb{E}U^4 =& \frac{1}{n^4}\Bigg(\sum_{i=1}^n \|x_i\|_2^4\mathbb{E}(\psi(\epsilon_i)-\mathbb{E}\psi(\epsilon_i))^4 \\
&+ \sum_{i\neq i'}(2|x_i^T x_{i'}|_2^2 + \|x_i\|_2^2\|x_{i'}\|_2^2)\mathbb{E}(\psi(\epsilon_i)-\mathbb{E}\psi(\epsilon_i))^2\mathbb{E}(\psi(\epsilon_{i'})-\mathbb{E}\psi(\epsilon_{i'}))^2\Bigg) \\
\leq& \frac{1}{n^4}\Bigg(\sum_{i=1}^n \|x_i\|_2^4\mathbb{E}(\psi(\epsilon_i)-\mathbb{E}\psi(\epsilon_i))^4 \\
&+ 3\sum_{i\neq i'}\|x_i\|_2^2\|x_{i'}\|_2^2\mathbb{E}(\psi(\epsilon_i)-\mathbb{E}\psi(\epsilon_i))^2\mathbb{E}(\psi(\epsilon_{i'})-\mathbb{E}\psi(\epsilon_{i'}))^2\Bigg).
\end{aligned}$$

For any i, using the convexity of x^4, hence $(\frac{a+b}{2})^4 \leq \frac{a^4+b^4}{2}$, we have

$$\mathbb{E}(\psi(\epsilon_i)-\mathbb{E}\psi(\epsilon_i))^4 \leq 8\mathbb{E}\left(\psi(\epsilon_i)^4 + (\mathbb{E}\psi(\epsilon_i))^4\right) \leq 16\mathbb{E}\psi(\epsilon_i)^4 \leq 16\max_i \mathbb{E}\psi(\epsilon_i)^4.$$

By Cauchy-Schwartz inequality,

$$\mathbb{E}(\psi(\epsilon_i) - \mathbb{E}\psi(\epsilon_i))^2 \leq \mathbb{E}\psi(\epsilon_i)^2 \leq \sqrt{\mathbb{E}\psi(\epsilon_i)^4} \leq \sqrt{\max_i \mathbb{E}\psi(\epsilon_i)^4}.$$

Recall (A.23) that $\|x_i\|_2^2 \leq nT^2$ and thus,

$$\begin{aligned}\mathbb{E}U^4 &\leq \frac{1}{n^4}\left(16n \cdot n^2T^4 + 3n^2 \cdot n^2T^4\right) \cdot \max_i \mathbb{E}\psi(\epsilon_i)^4 \\ &\leq \frac{1}{n^4} \cdot (16n^3 + 3n^4)T^4 \max_i \mathbb{E}\psi(\epsilon_i)^4 = O\left(\text{polyLog}(n)\right).\end{aligned}$$

On the other hand, let $\mu^T = (\mathbb{E}\psi(\epsilon_1), \ldots, \mathbb{E}\psi(\epsilon_n))$, then $\|\mu\|_2^2 = O(n \cdot \text{polyLog}(n))$ and hence by definition of U_0 in (A.24),

$$U_0 = \frac{\|\mu^T X\|_2}{n} = \frac{1}{n}\sqrt{\mu^T X X^T \mu} \leq \sqrt{\frac{\|\mu\|_2^2}{n} \cdot \lambda_+} = O\left(\text{polyLog}(n)\right).$$

In summary,

$$\mathbb{E}\|\hat{\beta}\|_2^4 = O\left(\text{polyLog}(n)\right).$$

(iii) By mean-value theorem, there exists $a_x \in (0, x)$ such that

$$\rho(x) = \rho(0) + x\psi(0) + \frac{x^2}{2}\psi'(a_x).$$

By assumption **A**1 and Lemma A.2.2, we have

$$\rho(x) = \frac{x^2}{2}\psi'(a_x) \leq \frac{x^2}{2}\|\psi'\|_\infty \leq \frac{K_3 x^2}{2},$$

where K_3 is defined in Lemma A.2.2. As a result,

$$\mathbb{E}\rho(\epsilon_i)^8 \leq \left(\frac{K_3}{2}\right)^8 \mathbb{E}\epsilon_i^{16} = O(c_1^{16}).$$

Recall the definition of $\mathcal{E}$ in (A.23) and the convexity of x^8, we have

$$\mathbb{E}\mathcal{E}^8 \leq \frac{1}{n}\sum_{i=1}^n \mathbb{E}\rho(\epsilon_i)^8 = O(c_1^{16}) = O\left(\text{polyLog}(n)\right). \tag{A.40}$$

Under assumption **A**5, by Cauchy-Schwartz inequality,

$$\mathbb{E}(\Delta_C\sqrt{\mathcal{E}})^2 = \mathbb{E}\Delta_C^2\mathcal{E} \leq \sqrt{\mathbb{E}\Delta_C^4} \cdot \sqrt{\mathbb{E}\mathcal{E}^2} = O\left(\text{polyLog}(n)\right).$$

Under assumptions **A**1 and **A**3,

$$\frac{\sqrt{2K_1}}{K_0\lambda_-} = O\left(\text{polyLog}(n)\right).$$

Putting all the pieces together, we obtain that

$$\max_{j \in J_n} |b_j| = O_{L^2}\left(\frac{\text{polyLog}(n)}{\sqrt{n}}\right).$$

(iv) Similarly, by Holder's inequality,

$$\mathbb{E}(\Delta_C^3\mathcal{E})^2 = \mathbb{E}\Delta_C^6\mathcal{E}^2 \leq \left(\mathbb{E}\Delta_C^8\right)^{\frac{3}{4}} \cdot \left(\mathbb{E}\mathcal{E}^8\right)^{\frac{1}{4}} = O\left(\text{polyLog}(n)\right),$$

and under assumptions **A**1 and **A**3,

$$\frac{2K_1^2K_3\lambda_+T}{K_0^4\lambda_-^{\frac{7}{2}}} = O\left(\text{polyLog}(n)\right).$$

Therefore,

$$\max_{j\in J_n}|\hat{\beta}_j - b_j| = O_{L^2}\left(\frac{\text{polyLog}(n)}{n}\right).$$

(v) It follows from the previous part that

$$\mathbb{E}(\Delta_C^2\cdot\sqrt{\mathcal{E}})^2 = O\left(\text{polyLog}(n)\right).$$

Under assumptions **A**1 and **A**3, the multiplicative factors are also $O\left(\text{polyLog}(n)\right)$, i.e.

$$\frac{2K_1^2K_3\lambda_+T^2}{K_0^4\lambda_-^{\frac{7}{2}}} = O\left(\text{polyLog}(n)\right), \quad \frac{\sqrt{2}K_1}{K_0^{\frac{3}{2}}\lambda_-} = O\left(\text{polyLog}(n)\right).$$

Therefore,

$$\max_{j\in J_n}\max_i|R_i - r_{i,[j]}| = O_{L^2}\left(\frac{\text{polyLog}(n)}{\sqrt{n}}\right).$$

□

A.2.4 Controlling Gradient and Hessian

Proof of Lemma 2.4.2. Recall that $\hat{\beta}$ is the solution of the following equation

$$\frac{1}{n}\sum_{i=1}^n x_i\psi(\epsilon_i - x_i^T\hat{\beta}) = 0. \tag{A.41}$$

Taking derivative of (A.41), we have

$$X^TD\left(I - X\frac{\partial\hat{\beta}}{\partial\epsilon^T}\right) = 0 \Longrightarrow \frac{\partial\hat{\beta}}{\partial\epsilon^T} = (X^TDX)^{-1}X^TD.$$

This establishes (2.9). To establishes (2.10), note that (2.9) can be rewritten as

$$(X^TDX)\frac{\partial\hat{\beta}}{\partial\epsilon^T} = X^TD. \tag{A.42}$$

Fix $k \in \{1, \cdots, n\}$. Note that

$$\frac{\partial R_i}{\partial \epsilon_k} = \frac{\partial \epsilon_i}{\partial \epsilon_k} - x_i^T \frac{\partial \hat{\beta}}{\partial \epsilon_k} = I(i=k) - x_i^T (X^T D X)^{-1} X^T D.$$

Recall that $G = I - X(X^T D X)^{-1} X^T D$, we have

$$\frac{\partial R_i}{\partial \epsilon_k} = e_i^T G e_k, \tag{A.43}$$

where e_i is the i-th canonical basis of $\mathbb{R}^n$. As a result,

$$\frac{\partial D}{\partial \epsilon_k} = \tilde{D} \operatorname{diag}(G e_k). \tag{A.44}$$

Taking derivative of (A.42), we have

$$\begin{aligned}
& X^T \frac{\partial D}{\partial \epsilon_k} X \frac{\partial \hat{\beta}}{\partial \epsilon^T} + (X^T D X) \frac{\partial \hat{\beta}}{\partial \epsilon_k \partial \epsilon^T} = X^T \frac{\partial D}{\partial \epsilon_k} \\
\Longrightarrow & \frac{\partial \hat{\beta}}{\partial \epsilon_k \partial \epsilon^T} = (X^T D X)^{-1} X^T \frac{\partial D}{\partial \epsilon_k} \left(I - X (X^T D X)^{-1} X^T D \right) \\
\Longrightarrow & \frac{\partial \hat{\beta}}{\partial \epsilon_k \partial \epsilon^T} = (X^T D X)^{-1} X^T \tilde{D} \operatorname{diag}(G e_k) G,
\end{aligned}$$

where $G = I - X(X^T D X)^{-1} X^T D$ is defined in (A.18) in p.119. Then for each $j \in \{1, \cdots, p\}$ and $k \in \{1, \ldots, n\}$,

$$\frac{\partial \hat{\beta}_j}{\partial \epsilon_k \partial \epsilon^T} = e_j^T (X^T D X)^{-1} X^T \tilde{D} \operatorname{diag}(G e_k) G = e_k^T G^T \operatorname{diag}(e_j^T (X^T D X)^{-1} X^T \tilde{D}) G$$

where we use the fact that $a^T \operatorname{diag}(b) = b^T \operatorname{diag}(a)$ for any vectors a, b. This implies that

$$\frac{\partial \hat{\beta}_j}{\partial \epsilon \partial \epsilon^T} = G^T \operatorname{diag}(e_j^T (X^T D X)^{-1} X^T \tilde{D}) G$$

□

Proof of Lemma 2.4.3. Throughout the proof we are using the simple fact that $\|a\|_\infty \leq \|a\|_2$. Based on it, we found that

$$\begin{aligned}
& \left\| e_j^T (X^T D X)^{-1} X^T D^{\frac{1}{2}} \right\|_\infty \leq \left\| e_j^T (X^T D X)^{-1} X^T D^{\frac{1}{2}} \right\|_2 \\
= & \sqrt{e_j^T (X^T D X)^{-1} X^T D X (X^T D X)^{-1} e_j} \\
= & \sqrt{e_j^T (X^T D X)^{-1} e_j} \leq \frac{1}{(n K_0 \lambda_-)^{\frac{1}{2}}}.
\end{aligned} \tag{A.45}$$

Thus for any $m > 1$, recall that $M_j = \mathbb{E}\left\|e_j^T(X^TDX)^{-1}X^TD^{\frac{1}{2}}\right\|_\infty$,

$$\begin{aligned}
&\mathbb{E}\left\|e_j^T(X^TDX)^{-1}X^TD^{\frac{1}{2}}\right\|_\infty^m \\
\leq&\mathbb{E}\left\|e_j^T(X^TDX)^{-1}X^TD^{\frac{1}{2}}\right\|_\infty \cdot \left\|e_j^T(X^TDX)^{-1}X^TD^{\frac{1}{2}}\right\|_2^{m-1} \\
\leq&\frac{M_j}{(nK_0\lambda_-)^{\frac{m-1}{2}}}.
\end{aligned} \tag{A.46}$$

We should emphasize that we cannot use the naive bound that

$$\begin{aligned}
&\mathbb{E}\left\|e_j^T(X^TDX)^{-1}X^TD^{\frac{1}{2}}\right\|_\infty^m \leq \mathbb{E}\left\|e_j^T(X^TDX)^{-1}X^TD^{\frac{1}{2}}\right\|_2^m \leq \frac{1}{(nK_0\lambda_-)^{\frac{m}{2}}}, \\
\Longrightarrow&\left\|e_j^T(X^TDX)^{-1}X^TD^{\frac{1}{2}}\right\|_\infty = O_{L^m}\left(\frac{\text{polyLog}(n)}{\sqrt{n}}\right)
\end{aligned} \tag{A.47}$$

since it fails to guarantee the convergence of TV distance. We will address this issue after deriving Lemma 2.4.4.

By contrast, as proved below,

$$\left\|e_j^T(X^TDX)^{-1}X^TD^{\frac{1}{2}}\right\|_\infty = O_p(M_j) = O_p\left(\frac{\text{polyLog}(n)}{n}\right) \ll \frac{1}{\sqrt{nK_0\lambda_-}}. \tag{A.48}$$

Thus (A.46) produces a slightly tighter bound

$$\left\|e_j^T(X^TDX)^{-1}X^TD^{\frac{1}{2}}\right\|_\infty = O_{L^m}\left(\frac{\text{polyLog}(n)}{n^{\frac{m+1}{2m}}}\right).$$

It turns out that the above bound suffices to prove the convergence. Although (A.48) implies the possibility to sharpen the bound from $n^{-\frac{m+1}{2m}}$ to n^{-1} using refined analysis, we do not explore this to avoid extra conditions and notation.

• **Bound for κ_{0j}**

First we derive a bound for κ_{0j}. By definition,

$$\kappa_{0j}^2 = \mathbb{E}\left\|\frac{\partial\hat{\beta}_j}{\partial\epsilon^T}\right\|_4^4 \leq \mathbb{E}\left(\left\|\frac{\partial\hat{\beta}_j}{\partial\epsilon^T}\right\|_\infty^2 \cdot \left\|\frac{\partial\hat{\beta}_j}{\partial\epsilon^T}\right\|_2^2\right).$$

By Lemma 2.4.2 and (A.46) with $m = 2$,

$$\mathbb{E}\left\|\frac{\partial\hat{\beta}_j}{\partial\epsilon^T}\right\|_\infty^2 \leq \mathbb{E}\left\|e_j^T(X^TDX)^{-1}X^TD^{\frac{1}{2}}\right\|_\infty^2 \cdot K_1 = \frac{K_1M_j}{(nK_0\lambda_-)^{\frac{1}{2}}}.$$

On the other hand, it follows from (A.45) that

$$\left\|\frac{\partial\hat{\beta}_j}{\partial\epsilon^T}\right\|_2^2 = \left\|e_j^T(X^TDX)^{-1}X^TD\right\|_2^2 \leq K_1\cdot\left\|e_j^T(X^TDX)^{-1}X^TD^{\frac{1}{2}}\right\|_2^2 \leq \frac{K_1}{nK_0\lambda_-}. \tag{A.49}$$

Putting the above two bounds together we have

$$\kappa_{0j}^2 \leq \frac{K_1^2}{(nK_0\lambda_-)^{\frac{3}{2}}}\cdot M_j. \tag{A.50}$$

• **Bound for** κ_{1j}

As a by-product of (A.49), we obtain that

$$\kappa_{1j}^4 = \mathbb{E}\left\|\frac{\partial\hat{\beta}_j}{\partial\epsilon^T}\right\|_2^4 \leq \frac{K_1^2}{(nK_0\lambda_-)^2}. \tag{A.51}$$

• **Bound for** κ_{2j}

Finally, we derive a bound for κ_{2j}. By Lemma 2.4.2, κ_{2j} involves the operator norm of a symmetric matrix with form G^TMG where M is a diagonal matrix. Then by the triangle inequality,

$$\left\|G^TMG\right\|_{op} \leq \|M\|_{\text{op}}\cdot\left\|G^TG\right\|_{op} = \|M\|_{\text{op}}\cdot\|G\|_{op}^2.$$

Note that

$$D^{\frac{1}{2}}GD^{-\frac{1}{2}} = I - D^{\frac{1}{2}}X(X^TDX)^{-1}X^TD^{\frac{1}{2}}$$

is a projection matrix, which is idempotent. This implies that

$$\left\|D^{\frac{1}{2}}GD^{-\frac{1}{2}}\right\|_{op} = \lambda_{\max}\left(D^{\frac{1}{2}}GD^{-\frac{1}{2}}\right) \leq 1.$$

Write G as $D^{-\frac{1}{2}}(D^{\frac{1}{2}}GD^{-\frac{1}{2}})D^{\frac{1}{2}}$, then we have

$$\|G\|_{op} \leq \left\|D^{-\frac{1}{2}}\right\|_{op}\cdot\left\|D^{\frac{1}{2}}GD^{-\frac{1}{2}}\right\|_{op}\cdot\left\|D^{\frac{1}{2}}\right\|_{op} \leq \sqrt{\frac{K_1}{K_0}}.$$

Returning to κ_{2j}, we obtain that

$$\begin{aligned}
\kappa_{2j}^4 &= \mathbb{E}\left\|G^T\operatorname{diag}(e_j^T(X^TDX)^{-1}X^T\tilde{D})G\right\|_{op}^4 \\
&\leq \mathbb{E}\left(\left\|e_j^T(X^TDX)^{-1}X^T\tilde{D}\right\|_\infty^4\cdot\|G\|_{op}^8\right) \\
&\leq \mathbb{E}\left(\left\|e_j^T(X^TDX)^{-1}X^T\tilde{D}\right\|_\infty^4\right)\left(\frac{K_1}{K_0}\right)^4
\end{aligned}$$

$$= \mathbb{E}\left(\left\|e_j^T(X^TDX)^{-1}X^TD^{\frac{1}{2}}D^{-\frac{1}{2}}\tilde{D}\right\|_\infty^4\right)\cdot\left(\frac{K_1}{K_0}\right)^4$$

Assumption **A**1 implies that

$$\forall i,\ \frac{|\psi''(R_i)|}{\sqrt{\psi'(R_i)}} \le K_2 \ \&\ \text{hence}\ \|D^{-\frac{1}{2}}\tilde{D}\|_{\text{op}} \le K_2.$$

Therefore,

$$\left\|e_j^T(X^TDX)^{-1}X^TD^{\frac{1}{2}}D^{-\frac{1}{2}}\tilde{D}\right\|_\infty^4 \le K_2^4\cdot\left\|e_j^T(X^TDX)^{-1}X^TD^{\frac{1}{2}}\right\|_\infty^4.$$

By (A.46) with $m=4$,

$$\kappa_{2j}^4 \le \frac{K_2^4}{(n\lambda_-)^{\frac{3}{2}}}\cdot\left(\frac{K_1}{K_0}\right)^4\cdot M_j. \tag{A.52}$$

□

Proof of Lemma 2.4.4. By Theorem A.2.4, for any j,

$$\mathbb{E}\hat{\beta}_j^4 \le \mathbb{E}\|\hat{\beta}\|_2^4 < \infty.$$

Then using the second-order Poincaré inequality (Proposition 2.4.1),

$$\max_{j\in J_n} d_{TV}\left(\mathcal{L}\left(\frac{\hat{\beta}_j-\mathbb{E}\hat{\beta}_j}{\sqrt{\text{Var}(\hat{\beta}_j)}}\right),N(0,1)\right) = O\left(\frac{c_1c_2\kappa_{0j}+c_1^3\kappa_{1j}\kappa_{2j}}{\text{Var}(\hat{\beta}_j)}\right)$$

$$=O\left(\frac{\frac{M_j^{\frac{1}{2}}}{n^{\frac{3}{4}}}+\frac{M_j^{\frac{1}{4}}}{n^{\frac{7}{8}}}}{\text{Var}(\hat{\beta}_j)}\cdot\text{polyLog}(n)\right) = O\left(\frac{(nM_j^2)^{\frac{1}{4}}+(nM_j^2)^{\frac{1}{8}}}{n\,\text{Var}(\hat{\beta}_j)}\cdot\text{polyLog}(n)\right).$$

It follows from (A.45) that $nM_j^2 = O\left(\text{polyLog}(n)\right)$ and the above bound can be simplified as

$$\max_{j\in J_n} d_{TV}\left(\mathcal{L}\left(\frac{\hat{\beta}_j-\mathbb{E}\hat{\beta}_j}{\sqrt{\text{Var}(\hat{\beta}_j)}}\right),N(0,1)\right) = O\left(\frac{(nM_j^2)^{\frac{1}{8}}}{n\,\text{Var}(\hat{\beta}_j)}\cdot\text{polyLog}(n)\right).$$

□

Remark A.2.5. *If we use the naive bound* (A.47)*, by repeating the above derivation, we obtain a worse bound for* $\kappa_{0,j} = O(\frac{\text{polyLog}(n)}{n})$ *and* $\kappa_2 = O(\frac{\text{polyLog}(n)}{\sqrt{n}})$*, in which case,*

$$\max_{j\in J_n} d_{TV}\left(\mathcal{L}\left(\frac{\hat{\beta}_j-\mathbb{E}\hat{\beta}_j}{\sqrt{\text{Var}(\hat{\beta}_j)}}\right),N(0,1)\right) = O\left(\frac{\text{polyLog}(n)}{n\,\text{Var}(\hat{\beta}_j)}\right).$$

However, we can only prove that $\mathrm{Var}(\hat{\beta}_j) = \Omega(\frac{1}{n})$. *Without the numerator* $(nM_j^2)^{\frac{1}{8}}$, *which will be shown to be* $O(n^{-\frac{1}{8}}\mathrm{polyLog}(n))$ *in the next subsection, the convergence cannot be proved.*

A.2.5 Upper Bound of M_j

As mentioned in Appendix A.1, we should approximate D by $D_{[j]}$ to remove the functional dependence on X_j . To achieve this, we introduce two terms, $M_j^{(1)}$ and $M_j^{(2)}$, defined as

$$M_j^{(1)} = \mathbb{E}(\|e_j^T(X^TDX)^{-1}X^TD_{[j]}^{\frac{1}{2}}\|_\infty), \quad M_j^{(2)} = \mathbb{E}(\|e_j^T(X^TD_{[j]}X)^{-1}X^TD_{[j]}^{\frac{1}{2}}\|_\infty).$$

We will first prove that both $|M_j - M_j^{(1)}|$ and $|M_j^{(1)} - M_j^{(2)}|$ are negligible and then derive an upper bound for $M_j^{(2)}$.

Controlling $|M_j - M_j^{(1)}|$

By Lemma A.2.2,

$$\|D^{\frac{1}{2}} - D_{[j]}^{\frac{1}{2}}\|_\infty \le K_2 \max_i |R_i - r_{i,[j]}| \triangleq K_2\mathcal{R}_j,$$

and by Theorem A.2.4,

$$\sqrt{\mathbb{E}\mathcal{R}_j^2} = O\left(\frac{\mathrm{polyLog}(n)}{\sqrt{n}}\right).$$

Then we can bound $|M_j - M_j^{(1)}|$ via the fact that $\|a\|_\infty \le \|a\|_2$ and algebra as follows.

$$\begin{aligned}
|M_j - M_j^{(1)}| &\le \mathbb{E}(\|e_j^T(X^TDX)^{-1}X^T(D^{\frac{1}{2}} - D_{[j]}^{\frac{1}{2}})\|_\infty) \\
&\le \mathbb{E}(\|e_j^T(X^TDX)^{-1}X^T(D^{\frac{1}{2}} - D_{[j]}^{\frac{1}{2}})\|_2) \\
&\le \sqrt{\mathbb{E}(\|e_j^T(X^TDX)^{-1}X^T(D^{\frac{1}{2}} - D_{[j]}^{\frac{1}{2}})\|_2^2)} \\
&= \sqrt{\mathbb{E}(e_j^T(X^TDX)^{-1}X^T(D^{\frac{1}{2}} - D_{[j]}^{\frac{1}{2}})^2X(X^TDX)^{-1}e_j)}.
\end{aligned}$$

By Lemma A.2.2,

$$|\sqrt{\psi'(R_i)} - \sqrt{\psi'(r_{i,[j]})}| \le K_2|R_i - r_{i,[j]}| \le K_2\mathcal{R}_j,$$

thus

$$(D^{\frac{1}{2}} - D_{[j]}^{\frac{1}{2}})^2 \preceq K_2^2\mathcal{R}_j^2 I \preceq \frac{K_2^2}{K_0}\mathcal{R}_j^2 D.$$

This entails that

$$|M_j - M_j^{(1)}| \le K_2K_0^{-\frac{1}{2}}\sqrt{\mathbb{E}(\mathcal{R}_j^2 \cdot e_j^T(X^TDX)^{-1}X^TDX(X^TDX)^{-1}e_j)}$$

$$= K_2 K_0^{-\frac{1}{2}} \sqrt{\mathbb{E}(\mathcal{R}_j^2 \cdot e_j^T (X^T D X)^{-1} e_j)}$$
$$\leq \frac{K_2}{\sqrt{n} K_0 \sqrt{\lambda_-}} \sqrt{\mathbb{E}(\mathcal{R}_j^2)} = O\left(\frac{\text{polyLog}(n)}{n}\right).$$

Bound of $|M_j^{(1)} - M_j^{(2)}|$

First we prove a useful lemma.

Lemma A.2.6. *For any symmetric matrix N with $\|N\|_{\text{op}} < 1$,*

$$(I - (I+N)^{-1})^2 \preceq \frac{N^2}{(1 - \|N\|_{\text{op}})^2}.$$

Proof. First, notice that

$$I - (I+N)^{-1} = (I + N - I)(I+N)^{-1} = N(I+N)^{-1},$$

and therefore

$$(I - (I+N)^{-1})^2 = N(I+N)^{-2}N.$$

Since $\|N\|_{\text{op}} < 1$, $I + N$ is positive semi-definite and

$$(I+N)^{-2} \preceq \frac{1}{(1 - \|N\|_{\text{op}})^2} I.$$

Therefore,

$$N(I+N)^{-2}N \preceq \frac{N^2}{(1 - \|N\|_{\text{op}})^2}.$$

□

We now back to bounding $|M_j^{(1)} - M_j^{(2)}|$. Let $A_j = X^T D_{[j]} X$, $B_j = X^T (D - D_{[j]}) X$. By Lemma A.2.2,

$$\|D - D_{[j]}\|_\infty \leq K_3 \max_i |R_i - r_{i,[j]}| = K_3 \mathcal{R}_j$$

and hence

$$\|B_j\|_{\text{op}} \leq K_3 \mathcal{R}_j \cdot n\lambda_+ I \triangleq n\eta_j.$$

where $\eta_j = K_3 \lambda_+ \cdot \mathcal{R}_j$. Then by Theorem A.2.4.(v),

$$\mathbb{E}(\eta_j^2) = O\left(\frac{\text{polyLog}(n)}{n}\right).$$

Using the fact that $\|a\|_\infty \leq \|a\|_2$, we obtain that

$$|M_j^{(1)} - M_j^{(2)}| \leq \mathbb{E}(\|e_j^T A_j^{-1} X^T D_{[j]}^{\frac{1}{2}} - e_j^T (A_j + B_j)^{-1} X^T D_{[j]}^{\frac{1}{2}}\|_\infty)$$

$$\begin{aligned}
&\leq \sqrt{\mathbb{E}(\|e_j^T A_j^{-1} X^T D_{[j]}^{\frac{1}{2}} - e_j^T (A_j+B_j)^{-1} X^T D_{[j]}^{\frac{1}{2}}\|_2^2)} \\
&= \sqrt{\mathbb{E}\left[e_j^T (A_j^{-1} - (A_j+B_j)^{-1}) X^T D_{[j]} X (A_j^{-1} - (A_j+B_j)^{-1}) e_j\right]} \\
&= \sqrt{\mathbb{E}\left[e_j^T (A_j^{-1} - (A_j+B_j)^{-1}) A_j (A_j^{-1} - (A_j+B_j)^{-1}) e_j\right]}
\end{aligned}$$

The inner matrix can be rewritten as

$$\begin{aligned}
&(A_j^{-1} - (A_j+B_j)^{-1}) A_j (A_j^{-1} - (A_j+B_j)^{-1}) \\
=&A_j^{-\frac{1}{2}} (I - (I + A_j^{-\frac{1}{2}} B_j A_j^{-\frac{1}{2}})^{-1}) A_j^{-\frac{1}{2}} A_j A_j^{-\frac{1}{2}} (I - (I + A_j^{-\frac{1}{2}} B_j A_j^{-\frac{1}{2}})^{-1}) A_j^{-\frac{1}{2}} \\
=&A_j^{-\frac{1}{2}} (I - (I + A_j^{-\frac{1}{2}} B_j A_j^{-\frac{1}{2}})^{-1})^2 A_j^{-\frac{1}{2}}. \qquad \text{(A.53)}
\end{aligned}$$

Let $N_j = A_j^{-\frac{1}{2}} B_j A_j^{-\frac{1}{2}}$, then

$$\|N_j\|_{\mathrm{op}} \leq \|A_j^{-\frac{1}{2}}\|_{\mathrm{op}} \cdot \|B_j\|_{\mathrm{op}} \cdot \|A_j^{-\frac{1}{2}}\|_{\mathrm{op}} \leq (nK_0\lambda_-)^{-\frac{1}{2}} \cdot n\eta_j \cdot (nK_0\lambda_-)^{-\frac{1}{2}} = \frac{\eta_j}{K_0\lambda_-}.$$

On the event $\{\eta_j \leq \frac{1}{2}K_0\lambda_-\}$, $\|N_j\|_{\mathrm{op}} \leq \frac{1}{2}$. By Lemma A.2.6,

$$(I - (I+N_j)^{-1})^2 \preceq 4N_j^2.$$

This together with (A.53) entails that

$$\begin{aligned}
&e_j^T (A_j^{-1} - (A_j+B_j)^{-1}) A_j (A_j^{-1} - (A_j+B_j)^{-1}) e_j = e_j^T A_j^{-\frac{1}{2}} (I - (I+N_j)^{-1})^2 A_j^{-\frac{1}{2}} e_j \\
\leq& 4 e_j^T A_j^{-\frac{1}{2}} N_j^2 A_j^{-\frac{1}{2}} e_j = e_j^T A_j^{-1} B_j A_j^{-1} B_j A_j^{-1} e_j \leq \|A_j^{-1} B_j A_j^{-1} B_j A_j^{-1}\|_{\mathrm{op}}.
\end{aligned}$$

Since $A_j \succeq nK_0\lambda_- I$, and $\|B_j\|_{\mathrm{op}} \leq n\eta_j$, we have

$$\|A_j^{-1} B_j A_j^{-1} B_j A_j^{-1}\|_{\mathrm{op}} \leq \|A_j^{-1}\|_{\mathrm{op}}^3 \cdot \|B_j\|_{\mathrm{op}}^2 \leq \frac{1}{n} \cdot \frac{1}{(K_0\lambda_-)^3} \cdot \eta_j^2.$$

Thus,

$$\begin{aligned}
&\mathbb{E}\left[e_j^T (A_j^{-1} - (A_j+B_j)^{-1}) A_j (A_j^{-1} - (A_j+B_j)^{-1}) e_j \cdot I\left(\eta_j \leq \frac{K_0\lambda_-}{2}\right)\right] \\
\leq& \mathbb{E}\left[e_j^T A_j^{-1} B_j A_j^{-1} B_j A_j^{-1} e_j\right] \leq \frac{1}{n} \cdot \frac{1}{(K_0\lambda_-)^3} \cdot \mathbb{E}\eta_j^2 = O\left(\frac{\mathrm{polyLog}(n)}{n^2}\right).
\end{aligned}$$

On the event $\{\eta_j > \frac{1}{2}K_0\lambda_-\}$, since $nK_0\lambda_- I \preceq A_j \preceq nK_1\lambda_+ I$ and $A_j + B_j \succeq nK_0\lambda_- I$,

$$\begin{aligned}
&|e_j^T (A_j^{-1} - (A_j+B_j)^{-1}) A_j (A_j^{-1} - (A_j+B_j)^{-1}) e_j| \\
\leq& nK_1\lambda_+ \cdot |e_j^T (A_j^{-1} - (A_j+B_j)^{-1})^2 e_j|
\end{aligned}$$

$$\begin{aligned}
\leq & nK_1\lambda_+ \cdot \left(2|e_j^T A_j^{-2} e_j| + 2|e_j^T (A_j+B_j)^{-2} e_j|\right) \\
\leq & \frac{4nK_1\lambda_+}{(nK_0\lambda_-)^2} = \frac{1}{n}\cdot\frac{4K_1\lambda_+}{(K_0\lambda_-)^2}.
\end{aligned}$$

This together with Markov inequality implies htat

$$\begin{aligned}
& \mathbb{E}\left[e_j^T(A_j^{-1}-(A_j+B_j)^{-1})A_j(A_j^{-1}-(A_j+B_j)^{-1})e_j \cdot I\left(\eta_j > \frac{K_0\lambda_-}{2}\right)\right] \\
\leq & \frac{1}{n}\cdot\frac{4K_1\lambda_+}{(K_0\lambda_-)^2}\cdot P\left(\eta_j > \frac{K_0\lambda_-}{2}\right) \\
\leq & \frac{1}{n}\cdot\frac{4K_1\lambda_+}{(K_0\lambda_-)^2}\cdot\frac{4}{(K_0\lambda_-)^2}\cdot\mathbb{E}\eta_j^2 \\
= & O\left(\frac{\mathrm{polyLog}(n)}{n^2}\right).
\end{aligned}$$

Putting pieces together, we conclude that

$$\begin{aligned}
|M_j^{(1)} - M_j^{(2)}| \leq & \sqrt{\mathbb{E}\left[e_j^T(A_j^{-1}-(A_j+B_j)^{-1})A_j(A_j^{-1}-(A_j+B_j)^{-1})e_j\right]} \\
\leq & \sqrt{\mathbb{E}\left[e_j^T(A_j^{-1}-(A_j+B_j)^{-1})A_j(A_j^{-1}-(A_j+B_j)^{-1})e_j \cdot I\left(\eta_j > \frac{K_0\lambda_-}{2}\right)\right]} \\
& + \sqrt{\mathbb{E}\left[e_j^T(A_j^{-1}-(A_j+B_j)^{-1})A_j(A_j^{-1}-(A_j+B_j)^{-1})e_j \cdot I\left(\eta_j \leq \frac{K_0\lambda_-}{2}\right)\right]} \\
= & O\left(\frac{\mathrm{polyLog}(n)}{n}\right).
\end{aligned}$$

Bound of $M_j^{(2)}$

Similar to (A.1), by block matrix inversion formula (See Proposition A.5.1),

$$e_j^T(X^T D_{[j]} X)^{-1} X^T D_{[j]}^{\frac{1}{2}} = \frac{X_j^T D_{[j]}^{\frac{1}{2}}(I-H_j)}{X_j^T D_{[j]}^{\frac{1}{2}}(I-H_j)D_{[j]}^{\frac{1}{2}}X_j},$$

where $H_j = D_{[j]}^{\frac{1}{2}} X_{[j]}(X_{[j]}^T D_{[j]} X_{[j]})^{-1} X_{[j]}^T D_{[j]}^{\frac{1}{2}}$. Recall that $\xi_j \geq K_0\lambda_-$ by (A.25), so we have

$$X_j^T D_{[j]}^{\frac{1}{2}}(I-H_j)D_{[j]}^{\frac{1}{2}}X_j = n\xi_j \geq n\lambda_-.$$

As for the numerator, recalling the definition of $h_{j,1,i}$, we obtain that

$$\|X_j^T D_{[j]}^{\frac{1}{2}}(I-H_j)\|_\infty = \left\|\frac{1}{n}X_j^T(I - D_{[j]}X_{[j]}(X_{[j]}^T D_{[j]} X_{[j]})^{-1}X_{[j]})\cdot D_{[j]}^{\frac{1}{2}}\right\|_\infty$$

$$\leq \sqrt{K_1} \cdot \left\| \frac{1}{n} X_j^T (I - D_{[j]} X_{[j]} (X_{[j]}^T D_{[j]} X_{[j]})^{-1} X_{[j]}) \right\|_\infty$$
$$= \sqrt{K_1} \max_i \left| h_{j,1,i}^T X_j \right| \leq \sqrt{K_1} \Delta_C \max_i \|h_{j,1,i}\|_2.$$

As proved in (A.35),

$$\max_i \|h_{j,1,i}\|_2 \leq \left(\frac{K_1}{K_0} \right)^{\frac{1}{2}}.$$

This entails that

$$\|X_j^T D_{[j]}^{\frac{1}{2}} (I - H_j)\|_\infty \leq \frac{K_1}{\sqrt{K_0}} \cdot \Delta_C = O_{L^1}(\text{polyLog}(n)).$$

Putting the pieces together we conclude that

$$M_j^{(2)} \leq \frac{\mathbb{E}\|X_j^T D_{[j]}^{\frac{1}{2}} (I - H_j)\|_\infty}{n\lambda_-} = O\left(\frac{\text{polyLog}(n)}{n} \right).$$

Summary

Based on results from Section B.5.1 - Section B.5.3, we have

$$M_j = O\left(\frac{\text{polyLog}(n)}{n} \right).$$

Note that the bounds we obtained do not depend on j, so we conclude that

$$\max_{j \in J_n} M_j = O\left(\frac{\text{polyLog}(n)}{n} \right).$$

A.2.6 Lower Bound of $\text{Var}(\hat{\beta}_j)$

Approximating $\text{Var}(\hat{\beta}_j)$ by $\text{Var}(b_j)$

By Theorem A.2.4,

$$\max_j \mathbb{E}(\hat{\beta}_j - b_j)^2 = O\left(\frac{\text{polyLog}(n)}{n^2} \right), \quad \max_j \mathbb{E} b_j^2 = O\left(\frac{\text{polyLog}(n)}{n} \right).$$

Using the fact that

$$\hat{\beta}_j^2 - b_j^2 = (\hat{\beta}_j - b_j + b_j)^2 - b_j^2 = (\hat{\beta}_j - b_j)^2 + 2(\hat{\beta}_j - b_j) b_j,$$

we can bound the difference between $\mathbb{E}\hat{\beta}_j^2$ and $\mathbb{E} b_j^2$ by

$$\left| \mathbb{E}\hat{\beta}_j^2 - \mathbb{E} b_j^2 \right| \leq \mathbb{E}(\hat{\beta}_j - b_j)^2 + 2|\mathbb{E}(\hat{\beta}_j - b_j) b_j|$$

$$\leq \mathbb{E}(\hat{\beta}_j - b_j)^2 + 2\sqrt{\mathbb{E}(\hat{\beta}_j - b_j)^2}\sqrt{\mathbb{E}b_j^2} = O\left(\frac{\text{polyLog}(n)}{n^{\frac{3}{2}}}\right).$$

Similarly, since $|a^2 - b^2| = |a - b| \cdot |a + b| \leq |a - b|(|a - b| + 2|b|)$,

$$|(\mathbb{E}\hat{\beta}_j)^2 - (\mathbb{E}b_j)^2| \leq \mathbb{E}|\hat{\beta}_j - b_j| \cdot \left(\mathbb{E}|\hat{\beta}_j - b_j| + 2\mathbb{E}|b_j|\right) = O\left(\frac{\text{polyLog}(n)}{n^{\frac{3}{2}}}\right).$$

Putting the above two results together, we conclude that

$$\left|\operatorname{Var}(\hat{\beta}_j) - \operatorname{Var}(b_j)\right| = O\left(\frac{\text{polyLog}(n)}{n^{\frac{3}{2}}}\right). \tag{A.54}$$

Then it is left to show that

$$\operatorname{Var}(b_j) = \Omega\left(\frac{1}{n \cdot \text{polyLog}(n)}\right).$$

Controlling $\operatorname{Var}(b_j)$ **by** $\operatorname{Var}(N_j)$

Recall that

$$b_j = \frac{1}{\sqrt{n}}\frac{N_j}{\xi_j}$$

where

$$N_j = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_{ij}\psi(r_{i,[j]}), \quad \xi_j = \frac{1}{n}X_j^T(D_{[j]} - D_{[j]}X_{[j]}(X_{[j]}^T D_{[j]}X_{[j]})^{-1}X_{[j]}^T D_{[j]})X_j.$$

Then

$$n\operatorname{Var}(b_j) = \mathbb{E}\left(\frac{N_j}{\xi_j} - \mathbb{E}\frac{N_j}{\xi_j}\right)^2 = \mathbb{E}\left(\frac{N_j - \mathbb{E}N_j}{\xi_j} + \frac{\mathbb{E}N_j}{\xi_j} - \mathbb{E}\frac{N_j}{\xi_j}\right)^2.$$

Using the fact that $(a + b)^2 - (\frac{1}{2}a^2 - b^2) = \frac{1}{2}(a + 2b)^2 \geq 0$, we have

$$n\operatorname{Var}(b_j) \geq \frac{1}{2}\mathbb{E}\left(\frac{N_j - \mathbb{E}N_j}{\xi_j}\right)^2 - \mathbb{E}\left(\frac{\mathbb{E}N_j}{\xi_j} - \mathbb{E}\frac{N_j}{\xi_j}\right)^2 \triangleq \frac{1}{2}I_1 - I_2. \tag{A.55}$$

Controlling I_1

The Assumption **A**4 implies that

$$\operatorname{Var}(N_j) = \frac{1}{n}X_j^T Q_j X_j = \Omega\left(\frac{\operatorname{tr}(\operatorname{Cov}(h_{j,0}))}{n \cdot \text{polyLog}(n)}\right).$$

It is left to show that $\operatorname{tr}(\operatorname{Cov}(h_{j,0}))/n = \Omega\left(\frac{1}{\text{polyLog}(n)}\right)$. Since this result will also be used later in Appendix A.3, we state it in the following the lemma.

Lemma A.2.7. *Under assumptions **A**1 - **A**3,*

$$\frac{\operatorname{tr}(\operatorname{Cov}(\psi(h_{j,0})))}{n} \geq \frac{K_0^4}{K_1^2} \cdot \left(\frac{n-p+1}{n}\right)^2 \cdot \min_i \operatorname{Var}(\epsilon_i) = \Omega\left(\frac{1}{\operatorname{polyLog}(n)}\right).$$

Proof. The (A.10) implies that

$$\operatorname{Var}(\psi(r_{i,[j]})) \geq K_0^2 \operatorname{Var}(r_{i,[j]}). \tag{A.56}$$

Note that $r_{i,[j]}$ is a function of ϵ, we can apply (A.10) again to obtain a lower bound for $\operatorname{Var}(r_{i,[j]})$. In fact, by variance decomposition formula, using the independence of ϵ_i's,

$$\operatorname{Var}(r_{i,[j]}) = \mathbb{E}\left(\operatorname{Var}\left(r_{i,[j]}\middle|\epsilon_{(i)}\right)\right) + \operatorname{Var}\left(\mathbb{E}\left(r_{i,[j]}\middle|\epsilon_{(i)}\right)\right) \geq \mathbb{E}\left(\operatorname{Var}\left(r_{i,[j]}\middle|\epsilon_{(i)}\right)\right),$$

where $\epsilon_{(i)}$ includes all but the i-th entry of ϵ. Apply A.10 again,

$$\operatorname{Var}\left(r_{i,[j]}\middle|\epsilon_{(i)}\right) \geq \inf_{\epsilon_i}\left|\frac{\partial r_{i,[j]}}{\partial \epsilon_i}\right|^2 \cdot \operatorname{Var}(\epsilon_i),$$

and hence

$$\operatorname{Var}(r_{i,[j]}) \geq \mathbb{E}\operatorname{Var}\left(r_{i,[j]}\middle|\epsilon_{(i)}\right) \geq \mathbb{E}\inf_{\epsilon}\left|\frac{\partial r_{i,[j]}}{\partial \epsilon_i}\right|^2 \cdot \operatorname{Var}(\epsilon_i). \tag{A.57}$$

Now we compute $\frac{\partial r_{i,[j]}}{\partial \epsilon_i}$. Similar to (A.43) in p.132, we have

$$\frac{\partial r_{k,[j]}}{\partial \epsilon_i} = e_i^T G_{[j]} e_k, \tag{A.58}$$

where $G_{[j]}$ is defined in (A.18) in p.119. When $k = i$,

$$\frac{\partial r_{i,[j]}}{\partial \epsilon_i} = e_i^T G_{[j]} e_i = e_i^T D_{[j]}^{-\frac{1}{2}} D_{[j]}^{\frac{1}{2}} G_{[j]} D_{[j]}^{-\frac{1}{2}} D_{[j]}^{\frac{1}{2}} e_i = e_i^T D_{[j]}^{\frac{1}{2}} G_{[j]} D_{[j]}^{-\frac{1}{2}} e_i. \tag{A.59}$$

By definition of $G_{[j]}$,

$$D_{[j]}^{\frac{1}{2}} G_{[j]} D_{[j]}^{-\frac{1}{2}} = I - D_{[j]}^{\frac{1}{2}} X_{[j]} (X_{[j]}^T D_{[j]} X_{[j]})^{-1} X_{[j]}^T D_{[j]}^{\frac{1}{2}}.$$

Let $\tilde{X}_{[j]} = D_{[j]}^{\frac{1}{2}} X_{[j]}$ and $H_j = \tilde{X}_{[j]} (\tilde{X}_{[j]}^T \tilde{X}_{[j]})^{-1} \tilde{X}_{[j]}^T$. Denote by $\tilde{X}_{(i),[j]}$ the matrix $\tilde{X}_{[j]}$ after removing i-th row, then by block matrix inversion formula (See Proposition A.5.1),

$$\begin{aligned} e_i^T H_j e_i &= \tilde{x}_{i,[j]}^T (\tilde{X}_{(i),[j]}^T \tilde{X}_{(i),[j]} + \tilde{x}_{i,[j]} \tilde{x}_{i,[j]}^T)^{-1} \tilde{x}_{i,[j]} \\ &= \tilde{x}_{i,[j]}^T \left((\tilde{X}_{(i),[j]}^T \tilde{X}_{(i),[j]})^{-1} - \frac{(\tilde{X}_{(i),[j]}^T \tilde{X}_{(i),[j]})^{-1} \tilde{x}_{i,[j]} \tilde{x}_{i,[j]}^T (\tilde{X}_{(i),[j]}^T \tilde{X}_{(i),[j]})^{-1}}{1 + \tilde{x}_{i,[j]}^T (\tilde{X}_{(i),[j]}^T \tilde{X}_{(i),[j]})^{-1} \tilde{x}_{i,[j]}} \right) \tilde{x}_{i,[j]} \end{aligned}$$

$$= \frac{\tilde{x}_{i,[j]}^T (\tilde{X}_{(i),[j]}^T \tilde{X}_{(i),[j]})^{-1} \tilde{x}_{i,[j]}}{1 + \tilde{x}_{i,[j]}^T (\tilde{X}_{(i),[j]}^T \tilde{X}_{(i),[j]})^{-1} \tilde{x}_{i,[j]}}.$$

This implies that

$$\begin{aligned}
e_i^T D_{[j]}^{\frac{1}{2}} G_{[j]} D_{[j]}^{-\frac{1}{2}} e_i = e_i^T (I - H_j) e_i &= \frac{1}{1 + \tilde{x}_{i,[j]}^T (\tilde{X}_{(i),[j]}^T \tilde{X}_{(i),[j]})^{-1} \tilde{x}_{i,[j]}} \\
&= \frac{1}{1 + e_i^T D_{[j]}^{\frac{1}{2}} X_{[j]} (X_{(i),[j]}^T D_{(i),[j]} X_{(i),[j]})^{-1} X_{[j]}^T D_{[j]}^{\frac{1}{2}} e_i} \\
&\geq \frac{1}{1 + K_0^{-1} e_i^T D_{[j]}^{\frac{1}{2}} X_{[j]} (X_{(i),[j]}^T X_{(i),[j]})^{-1} X_{[j]}^T D_{[j]}^{\frac{1}{2}} e_i} \\
&= \frac{1}{1 + K_0^{-1} (D_{[j]})_{i,i} \cdot e_i^T X_{[j]} (X_{(i),[j]}^T X_{(i),[j]})^{-1} X_{[j]}^T e_i} \\
&\geq \frac{1}{1 + K_0^{-1} K_1 e_i^T X_{[j]} (X_{(i),[j]}^T X_{(i),[j]})^{-1} X_{[j]}^T e_i} \\
&\geq \frac{K_0}{K_1} \cdot \frac{1}{1 + e_i^T X_{[j]} (X_{(i),[j]}^T X_{(i),[j]})^{-1} X_{[j]}^T e_i}.
\end{aligned} \tag{A.60}$$

Apply the above argument that replaces H_j by $X_{[j]}(X_{[j]}^T X_{[j]})^{-1} X_{[j]}^T$, we have

$$\frac{1}{1 + e_i^T X_{[j]}^T (X_{(i),[j]}^T X_{(i),[j]})^{-1} X_{[j]} e_i} = e_i^T (I - X_{[j]} (X_{[j]}^T X_{[j]})^{-1} X_{[j]}^T) e_i.$$

Thus, by (A.56) and (A.57),

$$\mathrm{Var}(\psi(r_{i,[j]})) \geq \frac{K_0^4}{K_1^2} \cdot [e_i^T (I - X_{[j]} (X_{[j]}^T X_{[j]})^{-1} X_{[j]}^T) e_i]^2.$$

Summing i over $1, \ldots, n$, we obtain that

$$\begin{aligned}
\frac{\mathrm{tr}(\mathrm{Cov}(h_{j,0}))}{n} &\geq \frac{K_0^4}{K_1^2} \cdot \frac{1}{n} \sum_{i=1}^n [e_i^T (I - X_{[j]} (X_{[j]}^T X_{[j]})^{-1} X_{[j]}^T) e_i]^2 \cdot \min_i \mathrm{Var}(\epsilon_i) \\
&\geq \frac{K_0^4}{K_1^2} \cdot \left(\frac{1}{n} \mathrm{tr}(I - X_{[j]} (X_{[j]}^T X_{[j]})^{-1} X_{[j]}^T) \right)^2 \cdot \min_i \mathrm{Var}(\epsilon_i) \\
&= \frac{K_0^4}{K_1^2} \cdot \left(\frac{n - p + 1}{n} \right)^2 \cdot \min_i \mathrm{Var}(\epsilon_i)
\end{aligned}$$

Since $\min_i \mathrm{Var}(\epsilon_i) = \Omega\left(\frac{1}{\mathrm{polyLog}(n)}\right)$ by assumption **A**2, we conclude that

$$\frac{\mathrm{tr}(\mathrm{Cov}(h_{j,0}))}{n} = \Omega\left(\frac{1}{\mathrm{polyLog}(n)}\right).$$

□

In summary,

$$\mathrm{Var}(N_j) = \Omega\left(\frac{1}{\mathrm{polyLog}(n)}\right).$$

Recall that

$$\xi_j = \frac{1}{n} X_j^T (D_{[j]} - D_{[j]} X_{[j]} (X_{[j]}^T D_{[j]} X_{[j]})^{-1} X_{[j]}^T D_{[j]}) X_j \leq \frac{1}{n} X_j^T D_{[j]} X_j \leq K_1 T^2,$$

we conclude that

$$I_1 \geq \frac{\mathrm{Var}(N_j)}{(K_1 T^2)^2} = \Omega\left(\frac{1}{\mathrm{polyLog}(n)}\right). \tag{A.61}$$

Controlling I_2

By definition,

$$\begin{aligned}
I_2 &= \mathbb{E}\left(\mathbb{E}N_j\left(\frac{1}{\xi_j} - \mathbb{E}\frac{1}{\xi_j}\right) + \mathbb{E}N_j \mathbb{E}\frac{1}{\xi_j} - \mathbb{E}\frac{N_j}{\xi_j}\right)^2 \\
&= \mathrm{Var}\left(\frac{\mathbb{E}N_j}{\xi_j}\right) + \left(\mathbb{E}N_j \mathbb{E}\frac{1}{\xi_j} - \mathbb{E}\frac{N_j}{\xi_j}\right)^2 \\
&= (\mathbb{E}N_j)^2 \cdot \mathrm{Var}\left(\frac{1}{\xi_j}\right) + \mathrm{Cov}\left(N_j, \frac{1}{\xi_j}\right)^2 \\
&\leq (\mathbb{E}N_j)^2 \cdot \mathrm{Var}\left(\frac{1}{\xi_j}\right) + \mathrm{Var}(N_j)\,\mathrm{Var}\left(\frac{1}{\xi_j}\right) \\
&= \mathbb{E}N_j^2 \cdot \mathrm{Var}\left(\frac{1}{\xi_j}\right).
\end{aligned} \tag{A.62}$$

By (A.27) in the proof of Theorem A.2.4,

$$\mathbb{E}N_j^2 \leq 2K_1 \mathbb{E}(\mathcal{E} \cdot \Delta_C^2) \leq 2K_1 \sqrt{\mathbb{E}\mathcal{E}^2 \cdot \mathbb{E}\Delta_C^4} = O\left(\mathrm{polyLog}(n)\right),$$

where the last equality uses the fact that $\mathcal{E} = O_{L^2}\left(\mathrm{polyLog}(n)\right)$ as proved in (A.40). On the other hand, let $\tilde{\xi}_j$ be an independent copy of ξ_j, then

$$\mathrm{Var}\left(\frac{1}{\xi_j}\right) = \frac{1}{2}\mathbb{E}\left(\frac{1}{\xi_j} - \frac{1}{\tilde{\xi}_j}\right)^2 = \frac{1}{2}\mathbb{E}\frac{(\xi_j - \tilde{\xi}_j)^2}{\xi_j^2 \tilde{\xi}_j^2}.$$

Since $\xi_j \geq K_0 \lambda_-$ as shown in (A.25), we have

$$\mathrm{Var}\left(\frac{1}{\xi_j}\right) \leq \frac{1}{2(K_0\lambda_-)^4}\mathbb{E}(\xi_j - \tilde{\xi}_j)^2 = \frac{1}{(K_0\lambda_-)^4} \cdot \mathrm{Var}(\xi_j). \tag{A.63}$$

To bound $\mathrm{Var}(\xi_j)$, we propose to using the standard Poincaré inequality (Chernoff 1981), which is stated as follows.

Proposition A.2.8. *Let $W = (W_1, \ldots, W_n) \sim N(0, I_{n\times n})$ and f be a twice differentiable function, then*

$$\mathrm{Var}(f(W)) \le \mathbb{E}\left\|\frac{\partial f(W)}{\partial W}\right\|_2^2.$$

In our case, $\epsilon_i = u_i(W_i)$, and hence for any twice differentiable function g,

$$\mathrm{Var}(g(\epsilon)) \le \mathbb{E}\left\|\frac{\partial g(\epsilon)}{\partial W}\right\|_2^2 = \mathbb{E}\left\|\frac{\partial g(\epsilon)}{\partial \epsilon}\cdot\frac{\partial \epsilon}{\partial W^T}\right\|_2^2 \le \max_i \|u_i'\|_\infty^2 \cdot \mathbb{E}\left\|\frac{\partial g(\epsilon)}{\partial \epsilon}\right\|_2^2.$$

Applying it to ξ_j, we have

$$\mathrm{Var}(\xi_j) \le c_1^2 \cdot \mathbb{E}\left\|\frac{\partial \xi_j}{\partial \epsilon}\right\|_2^2. \tag{A.64}$$

For given $k \in \{1, \ldots, n\}$, using the chain rule and the fact that $dB^{-1} = -B^{-1}dBB^{-1}$ for any square matrix B, we obtain that

$$\begin{aligned}
&\frac{\partial}{\partial \epsilon_k}\left(D_{[j]} - D_{[j]}X_{[j]}(X_{[j]}^T D_{[j]} X_{[j]})^{-1}X_{[j]}^T D_{[j]}\right)\\
=&\frac{\partial D_{[j]}}{\partial \epsilon_k} - \frac{\partial D_{[j]}}{\partial \epsilon_k}X_{[j]}(X_{[j]}^T D_{[j]} X_{[j]})^{-1}X_{[j]}^T D_{[j]} - D_{[j]}X_{[j]}(X_{[j]}^T D_{[j]} X_{[j]})^{-1}X_{[j]}^T \frac{\partial D_{[j]}}{\partial \epsilon_k}\\
&+ D_{[j]}X_{[j]}(X_{[j]}^T D_{[j]} X_{[j]})^{-1}X_{[j]}^T \frac{\partial D_{[j]}}{\partial \epsilon_k}X_{[j]}(X_{[j]}^T D_{[j]} X_{[j]})^{-1}X_{[j]}^T D_{[j]}\\
=&G_{[j]}^T \frac{\partial D_{[j]}}{\partial \epsilon_k} G_{[j]}
\end{aligned}$$

where $G_{[j]} = I - X_{[j]}(X_{[j]}^T D_{[j]} X_{[j]})^{-1}X_{[j]}^T D_{[j]}$ as defined in last subsection. This implies that

$$\frac{\partial \xi_j}{\partial \epsilon_k} = \frac{1}{n}X_j^T G_{[j]}^T \frac{\partial D_{[j]}}{\partial \epsilon_k} G_{[j]} X_j.$$

Then (A.64) entails that

$$\mathrm{Var}(\xi_j) \le \frac{1}{n^2}\sum_{k=1}^n \mathbb{E}\left(X_j^T G_{[j]}^T \frac{\partial D_{[j]}}{\partial \epsilon_k} G_{[j]} X_j\right)^2 \tag{A.65}$$

First we compute $\frac{\partial D_{[j]}}{\partial \epsilon_k}$. Similar to (A.44) in p.132 and recalling the definition of $D_{[j]}$ in (A.17) and that of $G_{[j]}$ in (A.18) in p.119, we have

$$\frac{\partial D_{[j]}}{\partial \epsilon_k} = \tilde{D}_{[j]}\,\mathrm{diag}(G_{[j]}e_k)\,\mathrm{diag}(\tilde{D}_{[j]}G_{[j]}e_k),$$

Let $\mathcal{X}_j = G_{[j]}X_j$ and $\tilde{\mathcal{X}}_j = \mathcal{X}_j \circ \mathcal{X}_j$ where $\circ$ denotes Hadamard product. Then

$$X_j^T G_{[j]}^T \frac{\partial D_{[j]}}{\partial \epsilon_k} G_{[j]} X_j = \mathcal{X}_j^T \frac{\partial D_{[j]}}{\partial \epsilon_k}\mathcal{X}_j = \mathcal{X}_j^T\,\mathrm{diag}(\tilde{D}_{[j]}G_{[j]}e_k)\mathcal{X}_j = \tilde{\mathcal{X}}_j^T \tilde{D}_{[j]}G_{[j]}e_k.$$

Here we use the fact that for any vectors $x, a \in \mathbb{R}^n$,

$$x^T \operatorname{diag}(a) x = \sum_{i=1}^{n} a_i x_i^2 = (x \circ x)^T a.$$

This together with (A.65) imply that

$$\operatorname{Var}(\xi_j) \le \frac{1}{n^2} \sum_{k=1}^{n} \mathbb{E}(\tilde{\mathcal{X}}_j^T \tilde{D}_{[j]} G_{[j]} e_k)^2 = \frac{1}{n^2} \mathbb{E} \left\| \tilde{\mathcal{X}}_j^T \tilde{D}_{[j]} G_{[j]} \right\|_2^2 = \frac{1}{n^2} \mathbb{E}(\tilde{\mathcal{X}}_j^T \tilde{D}_{[j]} G_{[j]} G_{[j]}^T \tilde{D}_{[j]} \tilde{\mathcal{X}}_j)$$

Note that $G_{[j]} G_{[j]}^T \preceq \|G_{[j]}\|_{\text{op}}^2 I$, and $\tilde{D}_{[j]} \preceq K_3 I$ by Lemma A.2.2 in p.121. Therefore we obtain that

$$\begin{aligned} \operatorname{Var}(\xi_j) &\le \frac{1}{n^2} \mathbb{E} \left(\left\| G_{[j]} \right\|_{op}^2 \cdot \tilde{\mathcal{X}}_j^T \tilde{D}_{[j]}^2 \tilde{\mathcal{X}}_j \right) \le \frac{K_3^2}{n^2} \cdot \mathbb{E} \left(\left\| G_{[j]} \right\|_{op}^2 \cdot \|\tilde{\mathcal{X}}_j\|_2^2 \right) \\ &= \frac{K_3^2}{n^2} \mathbb{E} \left(\left\| G_{[j]} \right\|_{op}^2 \cdot \|\mathcal{X}_j\|_4^4 \right) \le \frac{K_3^2}{n} \mathbb{E} \left(\left\| G_{[j]} \right\|_{op}^2 \cdot \|\mathcal{X}_j\|_\infty^4 \right) \end{aligned}$$

As shown in (A.34),

$$\|G_{[j]}\|_{\text{op}} \le \left(\frac{K_1}{K_0} \right)^{\frac{1}{2}}.$$

On the other hand, notice that the i-th row of $G_{[j]}$ is $h_{j,1,i}$ (see (A.20) for definition), by definition of Δ_C we have

$$\|\mathcal{X}_j\|_\infty = \|G_{[j]} X_j\|_\infty = \max_i |h_{j,1,i}^T X_j| \le \Delta_C \cdot \max \|h_{j,1,i}\|_2.$$

By (A.35) and assumption **A**5,

$$\|\mathcal{X}_j\|_\infty \le \Delta_C \cdot \left(\frac{K_1}{K_0} \right)^{\frac{1}{2}} = O_{L^4} \left(\text{polyLog}(n) \right).$$

This entails that

$$\operatorname{Var}(\xi_j) = O\left(\frac{\text{polyLog}(n)}{n} \right).$$

Combining with (A.62) and (A.63), we obtain that

$$I_2 = O\left(\frac{\text{polyLog}(n)}{n} \right).$$

Summary

Putting (A.55), (A.61) and (A.62) together, we conclude that

$$n\,\mathrm{Var}(b_j) = \Omega\left(\frac{1}{\mathrm{polyLog}(n)}\right) - O\left(\frac{1}{n\cdot\mathrm{polyLog}(n)}\right) = \Omega\left(\frac{1}{\mathrm{polyLog}(n)}\right)$$
$$\Longrightarrow \mathrm{Var}(b_j) = \Omega\left(\frac{1}{n\cdot\mathrm{polyLog}(n)}\right).$$

Combining with (A.54),

$$\mathrm{Var}(\hat{\beta}_j) = \Omega\left(\frac{1}{n\cdot\mathrm{polyLog}(n)}\right).$$

A.3 Proof of Other Results

A.3.1 Proofs of Propositions in Section 2.2.3

Proof of Proposition 2.2.1. Let $H_i(\alpha) = \mathbb{E}\rho(\epsilon_i - \alpha)$. First we prove that the conditions imply that 0 is the unique minimizer of $H_i(\alpha)$ for all i. In fact, since $\epsilon_i \overset{d}{=} -\epsilon_i$,

$$H_i(\alpha) = \mathbb{E}\rho(\epsilon_i - \alpha) = \frac{1}{2}\left(\mathbb{E}\rho(\epsilon_i - \alpha) + \rho(-\epsilon_i - \alpha)\right).$$

Using the fact that ρ is even, we have

$$H_i(\alpha) = \mathbb{E}\rho(\epsilon_i - \alpha) = \frac{1}{2}\left(\mathbb{E}\rho(\epsilon_i - \alpha) + \rho(\epsilon_i + \alpha)\right).$$

By (2.4), for any $\alpha \neq 0$, $H_i(\alpha) > H_i(0)$. As a result, 0 is the unique minimizer of H_i. Then for any $\beta \in \mathbb{R}^p$

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\rho(y_i - x_i^T\beta) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\rho(\epsilon_i - x_i^T(\beta-\beta^*)) = \frac{1}{n}\sum_{i=1}^{n}H_i(x_i^T(\beta-\beta^*)) \geq \frac{1}{n}\sum_{i=1}^{n}H_i(0).$$

The equality holds iff $x_i^T(\beta - \beta^*) = 0$ for all i since 0 is the unique minimizer of H_i. This implies that

$$X(\beta^*(\rho) - \beta^*) = 0.$$

Since X has full column rank, we conclude that

$$\beta^*(\rho) = \beta^*.$$

□

Proof of Proposition 2.2.2. For any $\alpha \in \mathbb{R}$ and $\beta \in \mathbb{R}^p$, let

$$G(\alpha; \beta) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\rho(y_i - \alpha - x_i^T \beta).$$

Since α_ρ minimizes $\mathbb{E}\rho(\epsilon_i - \alpha)$, it holds that

$$G(\alpha; \beta) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\rho(\epsilon_i - \alpha - x_i^T(\beta - \beta^*)) \geq \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\rho(\epsilon_i - \alpha_\rho) = G(\alpha_\rho, \beta^*).$$

Note that α_ρ is the unique minimizer of $\mathbb{E}\rho(\epsilon_i - \alpha)$, the above equality holds if and only if

$$\alpha + x_i^T(\beta - \beta^*) \equiv \alpha_\rho \Longrightarrow (\mathbf{1}\ X) \begin{pmatrix} \alpha - \alpha_\rho \\ \beta - \beta^* \end{pmatrix} = 0.$$

Since $(\mathbf{1}\ X)$ has full column rank, it must hold that $\alpha = \alpha_\rho$ and $\beta = \beta^*$. □

A.3.2 Proofs of Corollary 2.3.3

Proposition A.3.1. *Suppose that ϵ_i are i.i.d. such that $\mathbb{E}\rho(\epsilon_1 - \alpha)$ as a function of α has a unique minimizer α_ρ. Further assume that $X_{J_n^c}$ contains an intercept term, X_{J_n} has full column rank and*

$$\text{span}(\{X_j : j \in J_n\}) \cap \text{span}(\{X_j : j \in J_n^c\}) = \{0\} \tag{A.66}$$

Let

$$\beta_{J_n}(\rho) = \arg\min_{\beta_{J_n}} \left\{ \min_{\beta_{J_n^c}} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\rho(y_i - x_i^T \beta) \right\}.$$

Then $\beta_{J_n}(\rho) = \beta^*_{J_n}$.

Proof. let

$$G(\beta) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\rho(y_i - x_i^T \beta).$$

For any minimizer $\beta(\rho)$ of G, which might not be unique, we prove that $\beta_{J_n}(\rho) = \beta^*_{J_n}$. It follows by the same argument as in Proposition 2.2.2 that

$$x_i^T(\beta(\rho) - \beta^*) \equiv \alpha_0 \Longrightarrow X(\beta(\rho) - \beta^*) = \alpha_0 \mathbf{1} \Longrightarrow X_{J_n}(\beta_{J_n}(\rho)) = -X_{J_n^c}(\beta(\rho)_{J_n^c} - \beta^*_{J_n^c}) + \alpha_0 \mathbf{1}.$$

Since $X_{J_n^c}$ contains the intercept term, we have

$$X_{J_n}(\beta_{J_n}(\rho) - \beta^*_{J_n}) \in \text{span}(\{X_j : j \in J_n^c\}).$$

It then follows from (A.66) that

$$X_{J_n}(\beta_{J_n}(\rho) - \beta^*_{J_n}) = 0.$$

Since X_{J_n} has full column rank, we conclude that

$$\beta_{J_n}(\rho) = \beta^*_{J_n}.$$

□

The Proposition A.3.1 implies that $\beta^*_{J_n}$ is identifiable even when X is not of full column rank. A similar conclusion holds for the estimator $\hat{\beta}_{J_n}$ and the residuals R_i. The following two propositions show that under certain assumptions, $\hat{\beta}_{J_n}$ and R_i are invariant to the choice of $\hat{\beta}$ in the presense of multiple minimizers.

Proposition A.3.2. *Suppose that ρ is convex and twice differentiable with $\rho''(x) > c > 0$ for all $x \in \mathbb{R}$. Let $\hat{\beta}$ be any minimizer, which might not be unique, of*

$$F(\beta) \triangleq \frac{1}{n}\sum_{i=1}^{n} \rho(y_i - x_i^T\beta)$$

Then $R_i = y_i - x_i\hat{\beta}$ is independent of the choice of $\hat{\beta}$ for any i.

Proof. The conclusion is obvious if $F(\beta)$ has a unique minimizer. Otherwise, let $\hat{\beta}^{(1)}$ and $\hat{\beta}^{(2)}$ be two different minimizers of F denote by η their difference, i.e. $\eta = \hat{\beta}^{(2)} - \hat{\beta}^{(1)}$. Since F is convex, $\hat{\beta}^{(1)} + v\eta$ is a minimizer of F for all $v \in [0, 1]$. By Taylor expansion,

$$F(\hat{\beta}^{(1)} + v\eta) = F(\hat{\beta}^{(1)}) + v\nabla F(\hat{\beta}^{(1)})\eta + \frac{v^2}{2}\eta^T\nabla^2 F(\hat{\beta}^{(1)})\eta + o(v^2).$$

Since both $\hat{\beta}^{(1)} + v\eta$ and $\hat{\beta}^{(1)}$ are minimizers of F, we have $F(\hat{\beta}^{(1)} + v\eta) = F(\hat{\beta}^{(1)})$ and $\nabla F(\hat{\beta}^{(1)}) = 0$. By letting v tend to 0, we conclude that

$$\eta^T\nabla^2 F(\hat{\beta}^{(1)})\eta = 0.$$

The hessian of F can be written as

$$\nabla^2 F(\hat{\beta}^{(1)}) = \frac{1}{n}X^T \operatorname{diag}(\rho''(y_i - x_i^T\hat{\beta}^{(1)}))X \succeq \frac{cX^TX}{n}.$$

Thus, η satisfies that

$$\eta^T\frac{cX^TX}{n}\eta = 0 \Longrightarrow X\eta = 0. \tag{A.67}$$

This implies that

$$y - X\hat{\beta}^{(1)} = y - X\hat{\beta}^{(2)}$$

and hence R_i is the same for all i in both cases. □

Proposition A.3.3. *Suppose that ρ is convex and twice differentiable with $\rho''(x) > c > 0$ for all $x \in \mathbb{R}$. Further assume that X_{J_n} has full column rank and*

$$\text{span}(\{X_j : j \in J_n\}) \cap \text{span}(\{X_j : j \in J_n^c\}) = \{0\} \tag{A.68}$$

Let $\hat{\beta}$ be any minimizer, which might not be unique, of

$$F(\beta) \triangleq \frac{1}{n} \sum_{i=1}^{n} \rho(y_i - x_i^T \beta)$$

Then $\hat{\beta}_{J_n}$ is independent of the choice of $\hat{\beta}$.

Proof. As in the proof of Proposition A.3.2, we conclude that for any minimizers $\hat{\beta}^{(1)}$ and $\hat{\beta}^{(2)}$, $X\eta = 0$ where $\eta = \hat{\beta}^{(2)} - \hat{\beta}^{(1)}$. Decompose the term into two parts, we have

$$X_{J_n}\eta_{J_n} = -X^c_{J_n}\eta_{J_n^c} \in \text{span}(\{X_j : j \in J_n^c\}).$$

It then follows from (A.68) that $X_{J_n}\eta_{J_n} = 0$. Since X_{J_n} has full column rank, we conclude that $\eta_{J_n} = 0$ and hence $\hat{\beta}^{(1)}_{J_n} = \hat{\beta}^{(2)}_{J_n}$. □

Proof of Corollary 2.3.3. Under assumption **A**3*, X_{J_n} must have full column rank. Otherwise there exists $\alpha \in \mathbb{R}^{|J_n|}$ such that $X_{J_n}\alpha$, in which case $\alpha^T X_{J_n}^T (I - H_{J_n^c}) X_{J_n}\alpha = 0$. This violates the assumption that $\tilde{\lambda}_- > 0$. On the other hand, it also guarantees that

$$\text{span}(\{X_j : j \in J_n\}) \cap \text{span}(\{X_j : j \in J_n^c\}) = \{0\}.$$

This together with assumption **A**1 and Proposition A.3.3 implies that $\hat{\beta}_{J_n}$ is independent of the choice of $\hat{\beta}$.

Let $B_1 \in \mathbb{R}^{|J_n^c| \times |J_n|}$, $B_2 \in \mathbb{R}^{|J_n^c| \times |J_n^c|}$ and assume that B_2 is invertible. Let $\tilde{X} \in \mathbb{R}^{n \times p}$ such that

$$\tilde{X}_{J_n} = X_{J_n} - X_{J_n^c} B_1, \quad \tilde{X}_{J_n^c} = X_{J_n^c} B_2.$$

Then $\text{rank}(X) = \text{rank}(\tilde{X})$ and model (2.1) can be rewritten as

$$y = \tilde{X}\tilde{\beta}^* + \epsilon$$

where

$$\tilde{\beta}^*_{J_n} = \beta^*_{J_n}, \quad \tilde{\beta}^*_{J_n^c} = B_2^{-1}\beta^*_{J_n^c} + B_1\beta^*_{J_n}.$$

Let $\tilde{\hat{\beta}}$ be an M-estimator, which might not be unique, based on $\tilde{X}$. Then Proposition A.3.3 shows that $\tilde{\hat{\beta}}_{J_n}$ is independent of the choice of $\tilde{\hat{\beta}}$, and an invariance argument shows that

$$\tilde{\hat{\beta}}_{J_n} = \hat{\beta}_{J_n}.$$

In the rest of proof, we use $\tilde{\cdot}$ to denote the quantity obtained based on $\tilde{X}$. First we show that the assumption **A**4 is not affected by this transformation. In fact, for any $j \in J_n$, by definition we have

$$\mathrm{span}(\tilde{X}_{[j]}) = \mathrm{span}(X_{[j]})$$

and hence the leave-j-th-predictor-out residuals are not changed by Proposition A.3.2. This implies that $\tilde{h_{j,0}} = h_{j,0}$ and $\tilde{Q}_j = Q_j$. Recall the definition of $h_{j,0}$, the first-order condition of $\hat{\beta}$ entails that $X^T h_{j,0} = 0$. In particular, $X_{J_n^c}^T h_{j,0} = 0$ and this implies that for any $\alpha \in \mathbb{R}^n$,

$$0 = \mathrm{Cov}(X_{J_n^c}^T h_{j,0}, \alpha^T h_{j,0}) = X_{J_n^c} Q_j \alpha.$$

Thus,

$$\frac{\tilde{X}_j^T \tilde{Q}_j \tilde{X}_j}{\mathrm{tr}(\tilde{Q}_j)} = \frac{(X_j - X_{J_n}^c (B_1)_j)^T Q_j (X_j - X_{J_n^c}(B_1)_j)}{\mathrm{tr}(Q_j)} = \frac{X_j^T Q_j X_j}{\mathrm{tr}(Q_j)}.$$

Then we prove that the assumption **A**5 is also not affected by the transformation. The above argument has shown that

$$\frac{\tilde{h}_{j,0}^T \tilde{X}_j}{\|\tilde{h}_{j,0}\|_2} = \frac{h_{j,0}^T X_j}{\|h_{j,0}\|_2}.$$

On the other hand, let $B = \begin{pmatrix} I_{|J_n|} & 0 \\ -B_1 & B_2 \end{pmatrix}$, then B is non-singular and $\tilde{X} = XB$. Let $B_{(j),[j]}$ denote the matrix B after removing j-th row and j-th column. Then $B_{(j),[j]}$ is also non-singular and $\tilde{X}_{[j]} = X_{[j]} B_{(j),[j]}$. Recall the definition of $h_{j,1,i}$, we have

$$\begin{aligned}
\tilde{h}_{j,1,i} &= (I - \tilde{D}_{[j]} \tilde{X}_{[j]} (\tilde{X}_{[j]}^T \tilde{D}_{[j]} \tilde{X}_j)^{-1} \tilde{X}_{[j]}^T) e_i \\
&= (I - D_{[j]} X_{[j]} B_{(j),[j]} (B_{(j),[j]}^T X_{[j]}^T D_{[j]} X_j B_{(j),[j]})^{-1} B_{(j),[j]}^T X_{[j]}) e_i \\
&= (I - D_{[j]} X_{[j]} (X_{[j]}^T D_{[j]} X_j)^{-1} X_{[j]}) e_i \\
&= h_{j,1,i}.
\end{aligned}$$

On the other hand, by definition,

$$X_{[j]}^T h_{j,1,i} = X_{[j]}^T (I - D_{[j]} X_{[j]} (X_{[j]}^T D_{[j]} X_{[j]})^{-1} X_{[j]}^T) e_i = 0.$$

Thus,

$$h_{j,1,i}^T \tilde{X}_j = h_{j,1,i}^T (X_j - X_{J_n}^c (B_1)_j) = h_{j,1,i}^T X_j.$$

In summary, for any $j \in J_n$ and $i \le n$,

$$\frac{\tilde{h}_{j,1,i}^T \tilde{X}_j}{\|\tilde{h}_{j,1,i}\|_2} = \frac{h_{j,1,i}^T X_j}{\|h_{j,1,i}\|_2}.$$

Putting the pieces together we have

$$\tilde{\Delta}_C = \Delta_C.$$

By Theorem 2.3.1,

$$\max_{j\in J_n} d_{\mathrm{TV}}\left(\mathcal{L}\left(\frac{\hat{\beta}_j - \mathbb{E}\hat{\beta}_j}{\sqrt{\mathrm{Var}(\hat{\beta}_j)}}\right), N(0,1)\right) = o(1).$$

provided that $\tilde{X}$ satisfies the assumption **A**3.

Now let $U\Lambda V$ be the singular value decomposition of $X_{J_n^c}$, where $U \in \mathbb{R}^{n\times p}, \Lambda \in \mathbb{R}^{p\times p}, V \in \mathbb{R}^{p\times p}$ with $U^TU = V^TV = I_p$ and $\Lambda = \mathrm{diag}(\nu_1, \ldots, \nu_p)$ being the diagonal matrix formed by singular values of $X_{J_n^c}$. First we consider the case where $X_{J_n^c}$ has full column rank, then $\nu_j > 0$ for all $j \le p$. Let $B_1 = (X_{J_n}^T X_{J_n})^- X_{J_n}^T X_{J_n}$ and $B_2 = \sqrt{n/|J_n^c|}V^T\Lambda^{-1}$. Then

$$\frac{\tilde{X}^T\tilde{X}}{n} = \frac{1}{n}\begin{pmatrix} X_{J_n}^T(I - X_{J_n^c}(X_{J_n^c}^T X_{J_n^c})^{-1}X_{J_n^c})X_{J_n} & 0 \\ 0 & nI \end{pmatrix}.$$

This implies that

$$\lambda_{\max}\left(\frac{\tilde{X}^T\tilde{X}}{n}\right) = \max\left\{\tilde{\lambda}_{\max}, 1\right\}, \quad \lambda_{\min}\left(\frac{\tilde{X}^T\tilde{X}}{n}\right) = \min\left\{\tilde{\lambda}_{\min}, 1\right\}.$$

The assumption **A**3* implies that

$$\lambda_{\max}\left(\frac{\tilde{X}^T\tilde{X}}{n}\right) = O(\mathrm{polyLog}(n)), \quad \lambda_{\min}\left(\frac{\tilde{X}^T\tilde{X}}{n}\right) = \Omega\left(\frac{1}{\mathrm{polyLog}(n)}\right).$$

By Theorem 2.3.1, we conclude that

Next we consider the case where $X_{J_n}^c$ does not have full column rank. We first remove the redundant columns from $X_{J_n}^c$, i.e. replace $X_{J_n^c}$ by the matrix formed by its maximum linear independent subset. Denote by $\mathbf{X}$ this matrix. Then $\mathrm{span}(X) = \mathrm{span}(\mathbf{X})$ and $\mathrm{span}(\{X_j : j \notin J_n\}) = \mathrm{span}(\{\mathbf{X}_j : j \notin J_n\})$. As a consequence of Proposition A.3.1 and A.3.3, neither $\beta^*_{J_n}$ nor $\hat{\beta}_{J_n}$ is affected. Thus, the same reasoning as above applies to this case. □

A.3.3 Proofs of Results in Section 2.3.3

First we prove two lemmas regarding the behavior of Q_j. These lemmas are needed for justifying Assumption **A**4 in the examples.

Lemma A.3.4. *Under assumptions **A**1 and **A**2,*

$$\|Q_j\|_{\mathrm{op}} \le c_1^2\frac{K_3^2K_1}{K_0}, \quad \|Q_j\|_{\mathrm{F}} \le \sqrt{n}c_1^2\frac{K_3^2K_1}{K_0}$$

where $Q_j = \mathrm{Cov}(h_{j,0})$ as defined in section A.2.1.

Proof of Lemma A.3.4. By definition,

$$\|Q_j\|_{\mathrm{op}} = \sup_{\alpha \in \mathbb{S}^{n-1}} \alpha^T Q_j \alpha$$

where $\mathbb{S}^{n-1}$ is the n-dimensional unit sphere. For given $\alpha \in \mathbb{S}^{n-1}$,

$$\alpha^T Q_j \alpha = \alpha^T \operatorname{Cov}(h_{j,0})\alpha = \operatorname{Var}(\alpha^T h_{j,0})$$

It has been shown in (A.59) in Appendix A.2.6 that

$$\frac{\partial r_{i,[j]}}{\partial \epsilon_k} = e_i^T G_{[j]} e_k,$$

where $G_{[j]} = I - X_{[j]}(X_{[j]}^T D_{[j]} X_{[j]})^{-1} X_{[j]}^T D_{[j]}$. This yields that

$$\frac{\partial}{\partial \epsilon^T}\left(\sum_{i=1}^n \alpha_i \psi(r_{i,[j]})\right) = \sum_{i=1}^n \alpha_i \psi'(r_{i,[j]}) \cdot \frac{\partial r_{i,[j]}}{\partial \epsilon} = \sum_{i=1}^n \alpha_i \psi'(r_{i,[j]}) \cdot e_i^T G_{[j]} = \alpha^T \tilde{D}_{[j]} G_{[j]}.$$

By standard Poincaré inequality (see Proposition A.2.8), since $\epsilon_i = u_i(W_i)$,

$$\begin{aligned}
&\operatorname{Var}\left(\sum_{i=1}^n \alpha_i \psi(r_{i,[j]})\right) \le \max_k \|u_k'\|_\infty^2 \cdot \mathbb{E}\left\| \frac{\partial}{\partial \epsilon^T}\left(\sum_{i=1}^n \alpha_i \psi(r_{i,[j]})\right)\right\|^2 \\
\le & c_1^2 \cdot \mathbb{E}\left(\alpha^T \tilde{D}_{[j]} G_{[j]} G_{[j]}^T \tilde{D}_{[j]} \alpha\right) \le c_1^2 \mathbb{E}\|\tilde{D}_{[j]} G_{[j]} G_{[j]}^T \tilde{D}_{[j]}\|_2^2 \le c_1^2 \mathbb{E}\|\tilde{D}_j\|_{\mathrm{op}}^2 \|G_{[j]}\|_{\mathrm{op}}^2.
\end{aligned}$$

We conclude from Lemma A.2.2 and (A.34) in Appendix A.2.2 that

$$\|\tilde{D}_{[j]}\|_{\mathrm{op}} \le K_3, \quad \|G_{[j]}\|_{\mathrm{op}}^2 \le \frac{K_1}{K_0}.$$

Therefore,

$$\|Q_j\|_{\mathrm{op}} = \sup_{\alpha \in \mathbb{S}^{n-1}} \operatorname{Var}\left(\sum_{i=1}^n \alpha_i \psi(R_i)\right) \le c_1^2 \frac{K_3^2 K_1}{K_0}$$

and hence

$$\|Q_j\|_{\mathrm{F}} \le \sqrt{n}\|Q_j\|_{\mathrm{op}} \le \sqrt{n} \cdot c_1^2 \frac{K_3^2 K_1}{K_0}.$$

□

Lemma A.3.5. *Under assumptions **A**1 - **A**3,*

$$\operatorname{tr}(Q_j) \ge K^* n = \Omega(n \cdot \operatorname{polyLog}(n)),$$

where $K^* = \frac{K_0^4}{K_1^2} \cdot \left(\frac{n-p+1}{n}\right)^2 \cdot \min_i \operatorname{Var}(\epsilon_i)$.

Proof. This is a direct consequence of Lemma A.2.7 in p.142. □

Throughout the following proofs, we will use several results from the random matrix theory to bound the largest and smallest singular values of Z. The results are shown in Appendix A.5. Furthermore, in contrast to other sections, the notation $P(\cdot), \mathbb{E}(\cdot), \mathrm{Var}(\cdot)$ *denotes the probability, the expectation and the variance with respect to both* ϵ *and* Z *in this section.*

Proof of Proposition 2.3.4. By Proposition A.5.3,

$$\lambda_+ = (1+\sqrt{\kappa})^2 + o_p(1) = O_p(1), \quad \lambda_- = (1-\sqrt{\kappa})^2 - o_p(1) = \Omega_p(1)$$

and thus the assumption **A**3 holds with high probability. By Hanson-Wright inequality (Hanson and Wright 1971; Rudelson and Vershynin 2013); see Proposition A.5.2), for any given deterministic matrix A,

$$P(|Z_j^T A Z_j - \mathbb{E} Z_j^T A Z_j| \geq t) \leq 2\exp\left[-c\min\left\{\frac{t^2}{\sigma^4\|A\|_{\mathrm{F}}^2}, \frac{t}{\sigma^2\|A\|_{\mathrm{op}}}\right\}\right]$$

for some universal constant c. Let $A = Q_j$ and conditioning on $Z_{[j]}$, then by Lemma A.3.4, we know that

$$\|Q_j\|_{\mathrm{op}} \leq c_1^2 \frac{K_3^2 K_1}{K_0}, \quad \|Q_j\|_{\mathrm{F}} \leq \sqrt{n} c_1^2 \frac{K_3^2 K_1}{K_0}$$

and hence

$$\begin{aligned} &P\left(Z_j^T Q_j Z_j - \mathbb{E}(Z_j^T Q_j Z_j | Z_{[j]}) \leq -t \middle| Z_{[j]}\right) \\ &\leq 2\exp\left[-c\min\left\{\frac{t^2}{\sigma^4 \cdot n c_1^4 K_3^4 K_1^2/K_0^2}, \frac{t}{\sigma^2 c_1^2 K_3^2 K_1/K_0}\right\}\right]. \end{aligned} \tag{A.69}$$

Note that

$$\mathbb{E}(Z_j^T Q_j Z_j | Z_{[j]}) = \mathrm{tr}(\mathbb{E}[Z_j Z_j^T | Z_{[j]}] Q_j) = \mathbb{E} Z_{1j}^2 \,\mathrm{tr}(Q_j) = \tau^2 \,\mathrm{tr}(Q_j).$$

By Lemma A.3.5, we conclude that

$$\begin{aligned} &P\left(\frac{Z_j^T Q_j Z_j}{\mathrm{tr}(Q_j)} \leq \tau^2 - \frac{t}{nK^*} \middle| Z_{[j]}\right) \leq P\left(\frac{Z_j^T Q_j Z_j}{\mathrm{tr}(Q_j)} \leq \tau^2 - \frac{t}{\mathrm{tr}(Q_j)} \middle| Z_{[j]}\right) \\ \leq &2\exp\left[-c\min\left\{\frac{t^2}{\sigma^4 \cdot n c_1^4 K_3^4 K_1^2/K_0^2}, \frac{t}{2\sigma^2 c_1^2 K_3^2 K_1/K_0}\right\}\right]. \end{aligned} \tag{A.70}$$

Let $t = \frac{1}{2}\tau^2 n K^*$ and take expectation of both sides over $Z_{[j]}$, we obtain that

$$P\left(\frac{Z_j^T Q_j Z_j}{\mathrm{tr}(Q_j)} \leq \frac{\tau^2}{2}\right) \leq 2\exp\left[-cn\min\left\{\frac{K^{*2}\tau^4}{4\sigma^4 c_1^4 K_3^4 K_1^2/K_0^2}, \frac{K^*\tau^2}{2\sigma^2 c_1^2 K_3^2 K_1/K_0}\right\}\right]$$

and hence

$$P\left(\min_{j\in J_n}\frac{Z_j^TQ_jZ_j}{\mathrm{tr}(Q_j)}\le\frac{\tau^2}{2}\right)\le 2n\exp\left[-cn\min\left\{\frac{K^{*2}\tau^4}{4\sigma^4c_1^4K_3^4K_1^2/K_0^2},\frac{K^*\tau^2}{2\sigma^2c_1^2K_3^2K_1/K_0}\right\}\right]=o(1). \tag{A.71}$$

This entails that

$$\min_{j\in J_n}\frac{Z_j^TQ_jZ_j}{\mathrm{tr}(Q_j)}=\Omega_p\left(\frac{1}{\mathrm{polyLog}(n)}\right).$$

Thus, assumption **A**4 is also satisfied with high probability. On the other hand, since Z_j has i.i.d. mean-zero σ^2-sub-gaussian entries, for any deterministic unit vector $\alpha\in\mathbb{R}^n$, α^TZ_j is σ^2-sub-gaussian and mean-zero, and hence

$$P(|\alpha^TZ_j|\ge t)\le 2e^{-\frac{t^2}{2\sigma^2}}.$$

Let $\alpha_{j,i}=h_{j,1,i}/\|h_{j,1,i}\|_2$ and $\alpha_{j,0}=h_{j,0}/\|h_{j,0}\|_2$. Since $h_{j,1,i}$ and $h_{j,0}$ are independent of Z_j, a union bound then gives

$$P\left(\Delta_C\ge t+2\sigma\sqrt{\log n}\right)\le 2n^2e^{-\frac{t^2+4\sigma^2\log n}{2\sigma^2}}=2e^{-\frac{t^2}{2\sigma^2}}.$$

By Fubini's formula ((Durrett 2010), Lemma 2.2.8.),

$$\begin{aligned}\mathbb{E}\Delta_C^8&=\int_0^\infty 8t^7P(\Delta_C\ge t)dt\le\int_0^{2\sigma\sqrt{\log n}}8t^7dt+\int_{2\sigma\sqrt{\log n}}^\infty 8t^7P(\Delta_C\ge t)dt\\&=(2\sigma\sqrt{\log n})^8+\int_0^\infty 8(t+2\sigma\sqrt{\log n})^7P(\Delta_C\ge t+2\sigma\sqrt{\log n})dt\\&\le(2\sigma\sqrt{\log n})^8+\int_0^\infty 64(8t^7+128\sigma^7(\log n)^{\frac{7}{2}})\cdot 2e^{-\frac{t^2}{2\sigma^2}}dt\\&=O(\sigma^8\cdot\mathrm{polyLog}(n))=O\left(\mathrm{polyLog}(n)\right).\end{aligned} \tag{A.72}$$

This, together with Markov inequality, guarantees that assumption **A**5 is also satisfied with high probability. □

Proof of Proposition 2.3.5. It is left to prove that assumption **A**3 holds with high probability. The proof of assumption **A**4 and **A**5 is exactly the same as the proof of Proposition 2.3.5. By Proposition A.5.4,

$$\lambda_+=O_p(1).$$

On the other hand, by Proposition A.5.7 (Litvak et al. 2005),

$$P\left(\lambda_{\min}\left(\frac{Z^TZ}{n}\right)<c_1\right)\le e^{-c_2n}.$$

and thus

$$\lambda_-=\Omega_p(1).$$

□

Proof of Proposition 2.3.6. Since J_n excludes the intercept term, the proof of assumption **A**4 and **A**5 is still the same as Proposition 2.3.5. It is left to prove assumption **A**3. Let $R_1, \ldots, R_n$ be i.i.d. Rademacher random variables, i.e. $P(R_i = 1) = P(R_i = -1) = \frac{1}{2}$, and

$$Z^* = \text{diag}(B_1, \ldots, B_n)Z.$$

Then $(Z^*)^T Z^* = Z^T Z$. It is left to show that the assumption **A**3 holds for Z^* with high probability. Note that

$$(Z_i^*)^T = (B_i, B_i \tilde{x}_i^T).$$

For any $r \in \{1, -1\}$ and borel sets $B_1, \ldots, B_p \subset \mathbb{R}$,

$$\begin{aligned}
&P(B_i = r, B_i\tilde{Z}_{i1} \in B_1, \ldots, B_i\tilde{Z}_{i(p-1)} \in B_{p-1}) \\
&= P(B_i = r, \tilde{Z}_{i1} \in rB_1, \ldots, \tilde{Z}_{i(p-1)} \in rB_{p-1}) \\
&= P(B_i = r)P(\tilde{Z}_{i1} \in rB_1) \ldots P(\tilde{Z}_{i(p-1)} \in rB_{p-1}) \\
&= P(B_i = r)P(\tilde{Z}_{i1} \in B_1) \ldots P(\tilde{Z}_{i(p-1)} \in B_{p-1}) \\
&= P(B_i = r)P(B_i\tilde{Z}_{i1} \in B_1) \ldots P(B_i\tilde{Z}_{i(p-1)} \in B_{p-1})
\end{aligned}$$

where the last two lines uses the symmetry of $\tilde{Z}_{ij}$. Then we conclude that Z_i^* has independent entries. Since the rows of Z^* are independent, Z^* has independent entries. Since B_i are symmetric and sub-gaussian with unit variance and $B_i\tilde{Z}_{ij} \overset{d}{=} \tilde{Z}_{ij}$, which is also symmetric and sub-gaussian with variance bounded from below, Z^* satisfies the conditions of Propsition 2.3.5 and hence the assumption **A**3 is satisfied with high probability. □

Proof of Proposition 2.3.8 (with Proposition 2.3.7 being a special case). Let $Z_* = \Lambda^{-\frac{1}{2}} Z \Sigma^{-\frac{1}{2}}$, then Z_* has i.i.d. standard gaussian entries. By Proposition 2.3.6, Z_* satisfies assumption **A**3 with high probability. Thus,

$$\lambda_+ = \lambda_{\max}\left(\frac{\Sigma^{\frac{1}{2}} Z_*^T \Lambda Z_* \Sigma^{\frac{1}{2}}}{n}\right) \leq \lambda_{\max}(\Sigma) \cdot \lambda_{\max}(\Lambda) \cdot \lambda_{\max}\left(\frac{Z_*^T Z_*}{n}\right) = O_p(\text{polyLog}(n)),$$

and

$$\lambda_- = \lambda_{\min}\left(\frac{\Sigma^{\frac{1}{2}} Z_*^T \Lambda Z_* \Sigma^{\frac{1}{2}}}{n}\right) \geq \lambda_{\min}(\Sigma) \cdot \lambda_{\min}(\Lambda) \cdot \lambda_{\min}\left(\frac{Z_*^T Z_*}{n}\right) = \Omega_p\left(\frac{1}{\text{polyLog}(n)}\right).$$

As for assumption **A**4, the first step is to calculate $\mathbb{E}(Z_j^T Q_j Z_j | Z_{[j]})$. Let $\tilde{Z} = \Lambda^{-\frac{1}{2}} Z$, then $\text{vec}(\tilde{Z}) \sim N(0, I \otimes \Sigma)$. As a consequence,

$$\tilde{Z}_j | \tilde{Z}_{[j]} \sim N(\tilde{\mu}_j, \sigma_j^2 I)$$

where

$$\tilde{\mu}_j = \tilde{Z}_{[j]} \Sigma_{[j],[j]}^{-1} \Sigma_{[j],j} = \Lambda^{-\frac{1}{2}} Z_{[j]} \Sigma_{[j],[j]}^{-1} \Sigma_{[j],j}.$$

Thus,

$$Z_j|Z_{[j]} \sim N(\mu_j, \sigma_j^2\Lambda)$$

where $\mu_j = Z_{[j]}\Sigma_{[j],[j]}^{-1}\Sigma_{[j],j}$. It is easy to see that

$$\lambda_- \leq \min_j \sigma_j^2 \leq \max_j \sigma_j^2 \leq \lambda_+. \tag{A.73}$$

It has been shown that $Q_j\mu_j = 0$ and hence

$$Z_j^T Q_j Z_j = (Z_j - \mu_j)^T Q_j (Z_j - \mu_j).$$

Let $\mathscr{Z}_j = \Lambda^{-\frac{1}{2}}(Z_j - \mu_j)$ and $\tilde{Q}_j = \Lambda^{\frac{1}{2}} Q_j \Lambda^{\frac{1}{2}}$, then $\mathscr{Z}_j \sim N(0, \sigma_j^2 I)$ and

$$Z_j^T Q_j Z_j = \mathscr{Z}_j^T \tilde{Q}_j \mathscr{Z}_j.$$

By Lemma A.3.4,

$$\|\tilde{Q}_j\|_{\text{op}} \leq \|\Lambda\|_{\text{op}} \cdot \|Q_j\|_{\text{op}} \leq \lambda_{\max}(\Lambda) \cdot c_1^2 \frac{K_3^2 K_1}{K_0},$$

and hence

$$\|\tilde{Q}_j\|_{\text{F}} \leq \sqrt{n}\lambda_{\max}(\Lambda) \cdot c_1^2 \frac{K_3^2 K_1}{K_0}.$$

By Hanson-Wright inequality ((Hanson and Wright 1971; Rudelson and Vershynin 2013); see Proposition A.5.2), we obtain a similar inequality to (A.69) as follows:

$$\begin{aligned} &P\left(|Z_j^T Q_j Z_j - \mathbb{E}(Z_j^T Q_j Z_j | Z_{[j]})| \geq t \middle| Z_{[j]}\right) \\ \leq &2\exp\left[-c\min\left\{\frac{t^2}{\sigma_j^4 \cdot n\lambda_{\max}(\Lambda)^2 c_1^4 K_3^4 K_1^2/K_0^2}, \frac{t}{\sigma_j^2 \lambda_{\max}(\Lambda) c_1^2 K_3^2 K_1/K_0}\right\}\right]. \end{aligned}$$

On the other hand,

$$\mathbb{E}(Z_j^T Q_j Z_j | Z_{[j]}) = \mathbb{E}(\mathscr{Z}_j^T \tilde{Q}_j \mathscr{Z}_j | Z_{[j]}) = \sigma_j^2 \operatorname{tr}(\tilde{Q}_j).$$

By definition,

$$\operatorname{tr}(\tilde{Q}_j) = \operatorname{tr}(\Lambda^{\frac{1}{2}} Q_j \Lambda^{\frac{1}{2}}) = \operatorname{tr}(\Sigma Q_j) = \operatorname{tr}(Q_j^{\frac{1}{2}} \Lambda Q_j^{\frac{1}{2}}) \geq \lambda_{\min}(\Lambda) \operatorname{tr}(Q_j).$$

By Lemma A.3.5,

$$\operatorname{tr}(\tilde{Q}_j) \geq \lambda_{\min}(\Lambda) \cdot nK^*.$$

Similar to (A.70), we obtain that

$$P\left(\frac{Z_j^T Q_j Z_j}{\operatorname{tr}(Q_j)} \geq \sigma_j^2 - \frac{t}{nK^*} \middle| Z_{[j]}\right)$$

$$\leq 2\exp\left[-c\min\left\{\frac{t^2}{\sigma_j^4\cdot n\lambda_{\max}(\Lambda)^2c_1^4K_3^4K_1^2/K_0^2},\frac{t}{\sigma_j^2\lambda_{\max}(\Lambda)c_1^2K_3^2K_1/K_0}\right\}\right].$$

Let $t=\frac{1}{2}\sigma_j^2nK^*$, we have

$$P\left(\frac{Z_j^TQ_jZ_j}{\mathrm{tr}(Q_j)}\geq\frac{\sigma_j^2}{2}\right)$$
$$\leq 2\exp\left[-cn\min\left\{\frac{K^{*2}}{4\lambda_{\max}(\Lambda)^2c_1^4K_3^4K_1^2/K_0^2},\frac{K^*}{2\lambda_{\max}(\Lambda)c_1^2K_3^2K_1/K_0}\right\}\right]=o\left(\frac{1}{n}\right)$$

and a union bound together with (A.73) yields that

$$\min_{j\in J_n}\frac{Z_j^TQ_jZ_j}{\mathrm{tr}(Q_j)}=\Omega_p\left(\min_j\sigma_j^2\cdot\frac{1}{\mathrm{polyLog}(n)}\right)=\Omega_p\left(\frac{1}{\mathrm{polyLog}(n)}\right).$$

As for assumption **A**5, let

$$\alpha_{j,0}=\frac{\Lambda^{\frac{1}{2}}h_{j,0}}{\|h_{j,0}\|_2},\quad \alpha_{j,i}=\frac{\Lambda^{\frac{1}{2}}h_{j,1,i}}{\|h_{j,1,i}\|_2}$$

then for $i=0,1,\ldots,p$,

$$\|\alpha_{j,i}\|_2\leq\sqrt{\lambda_{\max}(\Lambda)}.$$

Note that

$$\frac{h_{j,0}^TZ_j}{\|h_{j,0}\|_2}=\alpha_{j,0}^TZ_j,\quad \frac{h_{j,1,i}^TZ_j}{\|h_{j,1,i}\|_2}=\alpha_{j,i}^TZ_j$$

using the same argument as in (A.72), we obtain that

$$\mathbb{E}\Delta_C^8=O\left(\lambda_{\max}(\Lambda)^4\cdot\max_j\sigma_j^8\cdot\mathrm{polyLog}(n)\right)=O\left(\mathrm{polyLog}(n)\right),$$

and by Markov inequality and (A.73),

$$\mathbb{E}(\Delta_C^8|Z)=O_p\left(\mathbb{E}\Delta_C^8\right)=O_p(\mathrm{polyLog}(n)).$$

□

Proof of Proposition 2.3.9. The proof that assumptions **A**4 and **A**5 hold with high probability is exactly the same as the proof of Proposition 2.3.8. It is left to prove assumption **A**3*; see Corollary 2.3.3. Let $c=(\min_i|(\Lambda^{-\frac{1}{2}}\mathbf{1})_i|)^{-1}$ and $\mathbf{Z}=(c\mathbf{1}\ \tilde{Z})$. Recall the the definition of $\tilde{\lambda}_+$ and $\tilde{\lambda}_-$, we have

$$\tilde{\lambda}_+=\lambda_{\max}(\Sigma_{\{1\}}),\quad \tilde{\lambda}_-=\lambda_{\min}(\Sigma_{\{1\}}),$$

where

$$\Sigma_{\{1\}} = \frac{1}{n}\tilde{Z}^T\left(I - \frac{\mathbf{1}\mathbf{1}^T}{n}\right)\tilde{Z}.$$

Rewrite $\Sigma_{\{1\}}$ as

$$\Sigma_{\{1\}} = \frac{1}{n}\left(\left(I - \frac{\mathbf{1}\mathbf{1}^T}{n}\right)\tilde{Z}\right)^T\left(\left(I - \frac{\mathbf{1}\mathbf{1}^T}{n}\right)\tilde{Z}\right).$$

It is obvious that

$$\operatorname{span}\left(\left(I - \frac{\mathbf{1}\mathbf{1}^T}{n}\right)\tilde{Z}\right) \subset \operatorname{span}(\mathbf{Z}).$$

As a consequence

$$\tilde{\lambda}_+ \leq \lambda_{\max}\left(\frac{\mathbf{Z}^T\mathbf{Z}}{n}\right), \quad \tilde{\lambda}_- \geq \lambda_{\min}\left(\frac{\mathbf{Z}^T\mathbf{Z}}{n}\right).$$

It remains to prove that

$$\lambda_{\max}\left(\frac{\mathbf{Z}^T\mathbf{Z}}{n}\right) = O_p\left(\operatorname{polyLog}(n)\right), \quad \lambda_{\min}\left(\frac{\mathbf{Z}^T\mathbf{Z}}{n}\right) = \Omega_p\left(\frac{1}{\operatorname{polyLog}(n)}\right).$$

To prove this, we let

$$Z_* = \Lambda^{-\frac{1}{2}}\mathbf{Z}\begin{pmatrix} 1 & 0 \\ 0 & \Sigma^{-\frac{1}{2}} \end{pmatrix} \triangleq (\nu\ \tilde{Z}_*),$$

where $\nu = c\Lambda^{-\frac{1}{2}}\mathbf{1}$ and $\tilde{Z}_* = \Lambda^{-\frac{1}{2}}\tilde{Z}\Sigma^{-\frac{1}{2}}$. Then

$$\lambda_{\max}\left(\frac{\mathbf{Z}^T\mathbf{Z}}{n}\right) = \lambda_{\max}\left(\frac{\Sigma^{\frac{1}{2}}Z_*^T\Lambda Z_*\Sigma^{\frac{1}{2}}}{n}\right) \leq \lambda_{\max}(\Sigma)\cdot\lambda_{\max}(\Lambda)\cdot\lambda_{\max}\left(\frac{Z_*^TZ_*}{n}\right),$$

and

$$\lambda_{\min}\left(\frac{\mathbf{Z}^T\mathbf{Z}}{n}\right) = \lambda_{\min}\left(\frac{\Sigma^{\frac{1}{2}}Z_*^T\Lambda Z_*\Sigma^{\frac{1}{2}}}{n}\right) \geq \lambda_{\min}(\Sigma)\cdot\lambda_{\min}(\Lambda)\cdot\lambda_{\min}\left(\frac{Z_*^TZ_*}{n}\right).$$

It is left to show that

$$\lambda_{\max}\left(\frac{Z_*^TZ_*}{n}\right) = O_p(\operatorname{polyLog}(n)), \quad \lambda_{\min}\left(\frac{Z_*^TZ_*}{n}\right) = \Omega_p\left(\frac{1}{\operatorname{polyLog}(n)}\right).$$

By definition, $\min_i|\nu_i| = 1$ and $\max_i|\nu_i| = O\left(\operatorname{polyLog}(n)\right)$, then

$$\lambda_{\max}\left(\frac{Z_*^TZ_*}{n}\right) = \lambda_{\max}\left(\frac{\tilde{Z}_*^T\tilde{Z}_*}{n} + \frac{\nu\nu^T}{n}\right) \leq \lambda_{\max}\left(\frac{\tilde{Z}_*^T\tilde{Z}_*}{n}\right) + \frac{\|\nu\|_2^2}{n}.$$

Since $\tilde{Z}_*$ has i.i.d. standard gaussian entries, by Proposition A.5.3,

$$\lambda_{\max}\left(\frac{\tilde{Z}_*^T\tilde{Z}_*}{n}\right) = O_p(1).$$

Moreover, $\|\nu\|_2^2 \le n \max_i |\nu_i|^2 = O(n \cdot \text{polyLog}(n))$ and thus,

$$\lambda_{\max}\left(\frac{Z_*^T Z_*}{n}\right) = O_p(\text{polyLog}(n)).$$

On the other hand, similar to Proposition 2.3.6,

$$\mathbf{Z}_* = \text{diag}(B_1, \ldots, B_n) Z_*$$

where $B_1, \ldots, B_n$ are i.i.d. Rademacher random variables. The same argument in the proof of Proposition 2.3.6 implies that $\mathbf{Z}_*$ has independent entries with sub-gaussian norm bounded by $\|\nu\|_\infty^2 \vee 1$ and variance lower bounded by 1. By Proposition A.5.7, Z_* satisfies assumption **A**3 with high probability. Therefore, **A**3* holds with high probability. □

Proof of Proposition 2.3.10. Let $\Lambda = (\lambda_1, \ldots, \lambda_n)$ and $\mathcal{Z}$ be the matrix with entries $\mathcal{Z}_{ij}$, then by Proposition 2.3.4 or Proposition 2.3.5, $\mathcal{Z}_{ij}$ satisfies assumption **A**3 with high probability. Notice that

$$\lambda_+ = \lambda_{\max}\left(\frac{\mathcal{Z}^T \Lambda^2 \mathcal{Z}}{n}\right) \le \lambda_{\max}(\Lambda)^2 \cdot \lambda_{\max}\left(\frac{\mathcal{Z}^T \mathcal{Z}}{n}\right) = O_p(\text{polyLog}(n)),$$

and

$$\lambda_- = \lambda_{\min}\left(\frac{\mathcal{Z}^T \Lambda^2 \mathcal{Z}}{n}\right) \ge \lambda_{\min}(\Lambda)^2 \cdot \lambda_{\min}\left(\frac{\mathcal{Z}^T \mathcal{Z}}{n}\right) = \Omega_p\left(\frac{1}{\text{polyLog}(n)}\right).$$

Thus Z satisfies assumption **A**3 with high probability.

Conditioning on any realization of Λ, the law of $\mathcal{Z}_{ij}$ does not change due to the independence between Λ and $\mathcal{Z}$. Repeating the arguments in the proof of Proposition 2.3.4 and Proposition 2.3.5, we can show that

$$\frac{\mathcal{Z}_j^T \tilde{Q}_j \mathcal{Z}_j}{\text{tr}(\tilde{Q}_j)} = \Omega_p\left(\frac{1}{\text{polyLog}(n)}\right), \quad \text{and} \quad \mathbb{E} \max_{i=0,\ldots,n; j=1,\ldots,p} |\tilde{\alpha}_{j,i}^T \mathcal{Z}_j|^8 = O_p(\text{polyLog}(n)), \quad \text{(A.74)}$$

where

$$\tilde{Q}_j = \Lambda Q_j \Lambda, \quad \tilde{\alpha}_{j,0} = \frac{\Lambda h_{j,0}}{\|\Lambda h_{j,0}\|_2}, \quad \tilde{\alpha}_{j,1,i} = \frac{\Lambda h_{j,1,i}}{\|\Lambda h_{j,1,i}\|_2}.$$

Then

$$\frac{Z_j^T Q_j Z_j}{\text{tr}(Q_j)} = \frac{\mathcal{Z}_j^T \tilde{Q}_j \mathcal{Z}_j}{\text{tr}(\tilde{Q}_j)} \cdot \frac{\text{tr}(\Lambda Q_j \Lambda)}{\text{tr}(Q_j)} \ge a^2 \cdot \frac{\mathcal{Z}_j^T \tilde{Q}_j \mathcal{Z}_j}{\text{tr}(\tilde{Q}_j)} = \Omega_p\left(\frac{1}{\text{polyLog}(n)}\right), \quad \text{(A.75)}$$

and

$$\mathbb{E}\Delta_C^8 = \mathbb{E}\left[\max_{i=0,\ldots,n; j=1,\ldots,p} |\tilde{\alpha}_{j,i}^T \mathcal{Z}_j|^8 \cdot \max\left\{\max_j \frac{\|\Lambda h_{j,0}\|_2}{\|h_{j,0}\|_2}, \max_{i,j} \frac{\|\Lambda h_{j,1,i}\|_2}{\|h_{j,1,i}\|_2}\right\}^8\right] \quad \text{(A.76)}$$

$$\leq b^8 \mathbb{E}\left[\max_{i=0,\ldots,n;j=1,\ldots,p} |\tilde{\alpha}_{j,i}^T \mathcal{Z}_j|^8\right]$$
$$= O_p(\text{polyLog}(n)).$$

By Markov inequality, the assumption **A**5 is satisfied with high probability. □

Proof of Proposition 2.3.11. The concentration inequality of ζ_i plus a union bound imply that

$$P\left(\max_i \zeta_i > (\log n)^{\frac{2}{\alpha}}\right) \leq nc_1 e^{-c_2(\log n)^2} = o(1).$$

Thus, with high probability,

$$\lambda_{\max} = \lambda_{\max}\left(\frac{\mathcal{Z}^T \Lambda^2 \mathcal{Z}}{n}\right) \leq (\log n)^{\frac{4}{\alpha}} \cdot \lambda_{\max}\left(\frac{\mathcal{Z}^T \mathcal{Z}}{n}\right) = O_p(\text{polyLog}(n)).$$

Let $n' = \lfloor (1-\delta)n \rfloor$ for some $\delta \in (0, 1-\kappa)$. Then for any subset I of $\{1, \ldots, n\}$ with size n', by Proposition A.5.6 (Proposition A.5.7), under the conditions of Proposition 2.3.4 (Proposition 2.3.5), there exists constants c_3 and c_4, which only depend on κ, such that

$$P\left(\lambda_{\min}\left(\frac{\mathcal{Z}_I^T \mathcal{Z}_I}{n}\right) < c_3\right) \leq e^{-c_4 n}$$

where $\mathcal{Z}_I$ represents the sub-matrix of $\mathcal{Z}$ formed by $\{\mathcal{Z}_i : i \in I\}$, where $\mathcal{Z}_i$ is the i-th row of $\mathcal{Z}$. Then by a union bound,

$$P\left(\min_{|I|=n'} \lambda_{\min}\left(\frac{\mathcal{Z}_I^T \mathcal{Z}_I}{n}\right) < c_3\right) \leq \binom{n}{n'} e^{-c_4 n}.$$

By Stirling's formula, there exists a constant $c_5 > 0$ such that

$$\binom{n}{n'} = \frac{n!}{n'!(n-n')!} \leq c_5 \exp\left\{(-\tilde{\delta}\log\tilde{\delta} - (1-\tilde{\delta})\log(1-\tilde{\delta}))n\right\}$$

where $\tilde{\delta} = n'/n$. For sufficiently small δ and sufficiently large n,

$$-\tilde{\delta}\log\tilde{\delta} - (1-\tilde{\delta})\log(1-\tilde{\delta}) < c_4$$

and hence

$$P\left(\min_{|I|=n'} \lambda_{\min}\left(\frac{\mathcal{Z}_I^T \mathcal{Z}_I}{n}\right) < c_3\right) < c_5 e^{-c_6 n} \tag{A.77}$$

for some $c_6 > 0$. By Borel-Cantelli Lemma,

$$\liminf_{n\to\infty} \min_{|I|=\lfloor(1-\delta)n\rfloor} \lambda_{\min}\left(\frac{\mathcal{Z}_I^T \mathcal{Z}_I}{n}\right) \geq c_3 \quad a.s..$$

On the other hand, since F^{-1} is continuous at δ, then

$$\zeta_{(\lfloor(1-\delta)n\rfloor)} \xrightarrow{a.s.} F^{-1}(\delta) > 0.$$

where $\zeta_{(k)}$ is the k-th largest of $\{\zeta_i : i = 1, \ldots, n\}$. Let I^* be the set of indices corresponding to the largest $\lfloor(1-\delta)n\rfloor$ ζ_i's. Then with probability 1,

$$\begin{aligned}
&\liminf_{n\to\infty} \lambda_{\min}\left(\frac{Z^TZ}{n}\right) = \liminf_{n\to\infty} \lambda_{\min}\left(\frac{\mathcal{Z}^T\Lambda^2\mathcal{Z}}{n}\right) \\
&\geq \liminf_{n\to\infty} \zeta_{(\lfloor(1-\delta)n\rfloor)} \cdot \liminf_{n\to\infty} \lambda_{\min}\left(\frac{\mathcal{Z}_{I^*}^T\Lambda_{I^*}^2\mathcal{Z}_{I^*}}{n}\right) \\
&\geq \liminf_{n\to\infty} \zeta_{(\lfloor(1-\delta)n\rfloor)} \cdot \liminf_{n\to\infty} \min_{|I|=\lfloor(1-\delta)n\rfloor} \lambda_{\min}\left(\frac{\mathcal{Z}_I^T\mathcal{Z}_I}{n}\right) \\
&\geq c_3 F^{-1}(\delta)^2 > 0.
\end{aligned}$$

To prove assumption **A**4, similar to (A.75) in the proof of Proposition 2.3.10, it is left to show that

$$\min_j \frac{\mathrm{tr}(\Lambda Q_j \Lambda)}{\mathrm{tr}(Q_j)} = \Omega_p\left(\frac{1}{\mathrm{polyLog}(n)}\right).$$

Furthermore, by Lemma A.3.5, it remains to prove that

$$\min_j \mathrm{tr}(\Lambda Q_j \Lambda) = \Omega_p\left(\frac{n}{\mathrm{polyLog}(n)}\right).$$

Recalling the equation (A.60) in the proof of Lemma A.2.7, we have

$$e_i^T Q_j e_i \geq \frac{K_0}{K_1} \cdot \frac{1}{1 + e_i^T Z_{[j]}^T (Z_{(i),[j]}^T Z_{(i),[j]})^{-1} Z_{[j]} e_i}. \tag{A.78}$$

By Proposition A.5.5,

$$P\left(\sqrt{\lambda_{\max}\left(\frac{\mathcal{Z}_j^T\mathcal{Z}_j}{n}\right)} > 3C_1\right) \leq 2e^{-C_2 n}.$$

On the other hand, apply (A.77) to $\mathcal{Z}_{(i),[j]}$, we have

$$P\left(\min_{|I|=\lfloor(1-\delta)n\rfloor} \lambda_{\min}\left(\frac{(\mathcal{Z}_{(i),[j]})_I^T(\mathcal{Z}_{(i),[j]})_I}{n}\right) < c_3\right) < c_5 e^{-c_6 n}.$$

A union bound indicates that with probability $(c_5 np + 2p)e^{-\min\{C_2,c_6\}n} = o(1)$,

$$\max_j \lambda_{\max}\left(\frac{\mathcal{Z}_{[j]}^T\mathcal{Z}_{[j]}}{n}\right) \leq 9C_1^2, \quad \min_{i,j} \min_{|I|=\lfloor(1-\delta)n\rfloor} \lambda_{\min}\left(\frac{(\mathcal{Z}_{(i),[j]})_I^T(\mathcal{Z}_{(i),[j]})_I}{n}\right) \geq c_3.$$

This implies that for any j,

$$\lambda_{\max}\left(\frac{Z_{[j]}^T Z_{[j]}}{n}\right) = \lambda_{\max}\left(\frac{\mathcal{Z}_{[j]}^T \Lambda^2 \mathcal{Z}_{[j]}}{n}\right) \le \zeta_{(1)}^2 \cdot 9C_1^2$$

and for any i and j,

$$\begin{aligned}
&\lambda_{\min}\left(\frac{Z_{(i),[j]}^T Z_{(i),[j]}}{n}\right) = \lambda_{\min}\left(\frac{\mathcal{Z}_{(i),[j]}^T \zeta_{(i)}^2 \mathcal{Z}_{(i),[j]}}{n}\right) \\
\ge &\min\{\zeta_{(\lfloor(1-\delta)n\rfloor)}, \zeta_{(\lfloor(1-\delta)n\rfloor)}+1\}^2 \cdot \min_{|I|=\lfloor(1-\delta)n\rfloor} \lambda_{\min}\left(\frac{(\mathcal{Z}_{(i),[j]})_I^T \zeta_{(i)}^2 (\mathcal{Z}_{(i),[j]})_I}{n}\right) \\
\ge &c_3 \min\{\zeta_{(\lfloor(1-\delta)n\rfloor)}, \zeta_{(\lfloor(1-\delta)n\rfloor)}+1\}^2 > 0.
\end{aligned}$$

Moreover, as discussed above,

$$\zeta_{(1)} \le (\log n)^{\frac{2}{\alpha}}, \min\{\zeta_{(\lfloor(1-\delta)n\rfloor)}, \zeta_{(\lfloor(1-\delta)n\rfloor)}+1\} \to F^{-1}(\delta)$$

almost surely. Thus, it follows from (A.78) that with high probability,

$$\begin{aligned}
e_i^T Q_j e_i &\ge \frac{K_0}{K_1} \cdot \frac{1}{1 + e_i^T Z_{[j]}^T (Z_{(i),[j]}^T Z_{(i),[j]})^{-1} Z_{[j]} e_i} \\
&\ge \frac{K_0}{K_1} \cdot \frac{1}{1 + e_i^T \frac{Z_{[j]}^T Z_{[j]}}{n} e_i \cdot c_3 (F^{-1}(\delta))^2} \\
&\ge \frac{K_0}{K_1} \cdot \frac{1}{1 + (\log n)^{\frac{4}{\alpha}} \cdot 9C_1^2 \cdot c_3 (F^{-1}(\delta))^2}.
\end{aligned}$$

The above bound holds for all diagonal elements of Q_j uniformly with high probability. Therefore,

$$\mathrm{tr}(\Lambda Q_j \Lambda) \ge \zeta_{(\lfloor(1-\delta)n\rfloor)}^2 \cdot \lfloor(1-\delta)n\rfloor \cdot \frac{K_0}{K_1} \cdot \frac{1}{1 + (\log n)^{\frac{4}{\alpha}} \cdot 9C_1^2 \cdot c_3 (F^{-1}(\delta))^2} = \Omega_p\left(\frac{n}{\mathrm{polyLog}(n)}\right).$$

As a result, the assumption **A**4 is satisfied with high probability. Finally, by (A.76), we obtain that

$$\mathbb{E}\Delta_C^8 \le \mathbb{E}\left[\max_{i=0,\ldots,n; j=1,\ldots,p} |\tilde{\alpha}_{j,i}^T \mathcal{Z}_j|^8 \cdot \|\Lambda\|_{\mathrm{op}}^8\right].$$

By Cauchy's inequality,

$$\mathbb{E}\Delta_C^8 \le \sqrt{\mathbb{E}\max_{i=0,\ldots,n; j=1,\ldots,p} |\tilde{\alpha}_{j,i}^T \mathcal{Z}_j|^{16}} \cdot \sqrt{\mathbb{E}\max_i \zeta_i^{16}}.$$

Similar to (A.72), we conclude that

$$\mathbb{E}\Delta_C^8 = O\left(\mathrm{polyLog}(n)\right)$$

and by Markov inequality, the assumption **A**5 is satisfied with high probability. □

A.3.4 More Results of Least-Squares (Section 2.5)

The Relation Between $S_j(X)$ and Δ_C

In Section 2.5, we give a sufficient and almost necessary condition for the coordinate-wise asymptotic normality of the least-square estimator $\hat{\beta}^{LS}$; see Theorem 2.5.1. In this subsubsection, we show that Δ_C is a generalization of $\max_{j\in J_n} S_j(X)$ for general M-estimators.

Consider the matrix $(X^TDX)^{-1}X^T$, where D is obtain by using general loss functions, then by block matrix inversion formula (see Proposition A.5.1),

$$\begin{aligned} e_1^T(X^TDX)^{-1}X^T &= e_1^T \begin{pmatrix} X_1^TDX_1 & X_1^TDX_{[1]} \\ X_{[1]}^TDX_1 & X_{[1]}^TDX_{[1]} \end{pmatrix}^{-1} \begin{pmatrix} X_1^T \\ X_{[1]}^T \end{pmatrix} \\ &= \frac{X_1^T(I - DX_{[1]}(X_{[1]}^TDX_{[1]})^{-1}X_{[1]}^T)}{X_1^T(D - DX_{[1]}(X_{[1]}^TDX_{[1]})^{-1}X_{[1]}^TD)X_1} \\ &\approx \frac{X_1^T(I - D_{[1]}X_{[1]}(X_{[1]}^TD_{[1]}X_{[1]})^{-1}X_{[1]}^T)}{X_1^T(D - DX_{[1]}(X_{[1]}^TDX_{[1]})^{-1}X_{[1]}^TD)X_1} \end{aligned}$$

where we use the approximation $D \approx D_{[1]}$. The same result holds for all $j \in J_n$, then

$$\frac{\|e_j^T(X^TDX)^{-1}X^T\|_\infty}{\|e_j^T(X^TDX)^{-1}X^T\|_2} \approx \frac{\|X_1^T(I - D_{[1]}X_{[1]}(X_{[1]}^TD_{[1]}X_{[1]})^{-1}X_{[1]}^T)\|_\infty}{\|X_1^T(I - D_{[1]}X_{[1]}(X_{[1]}^TD_{[1]}X_{[1]})^{-1}X_{[1]}^T)\|_2}.$$

Recall that $h_{j,1,i}^T$ is i-th row of $I - D_{[1]}X_{[1]}(X_{[1]}^TD_{[1]}X_{[1]})^{-1}X_{[1]}^T$, we have

$$\max_i \frac{|h_{j,1,i}^TX_1|}{\|h_{j,1,i}\|_2} \approx \frac{\|e_j^T(X^TDX)^{-1}X^T\|_\infty}{\|e_j^T(X^TDX)^{-1}X^T\|_2}.$$

The right-handed side equals to $S_j(X)$ in the least-square case. Therefore, although of complicated form, assumption **A**5 is not an artifact of the proof but is essential for the asymptotic normality.

Additional Examples

Benefit from the analytical form of the least-square estimator, we can depart from sub-gaussinity of the entries. The following proposition shows that a random design matrix Z with i.i.d. entries under appropriate moment conditions satisfies $\max_{j\in J_n} S_j(Z) = o(1)$ with high probability. This implies that, when X is one realization of Z, the conditions Theorem 2.5.1 are satisfied for X with high probability over Z.

Proposition A.3.6. *If $\{Z_{ij} : i \le n, j \in J_n\}$ are independent random variables with*

1. $\max_{i\le n, j\in J_n}(\mathbb{E}|Z_{ij}|^{8+\delta})^{\frac{1}{8+\delta}} \le M$ *for some* $\delta, M > 0$;

2. $\min_{i\le n, j\in J_n} \mathrm{Var}(Z_{ij}) > \tau^2$ *for some* $\tau > 0$

3. $P(Z$ *has full column rank*$) = 1 - o(1)$*;*

4. $\mathbb{E}Z_j \in \mathrm{span}\{Z_j : j \in J_n^c\}$ *almost surely for all* $j \in J_n$*;*

where Z_j *is the* j*-th column of* Z*. Then*

$$\max_{j\in J_n} S_j(Z) = O_p\left(\frac{1}{n^{\frac{1}{4}}}\right) = o_p(1).$$

A typical practically interesting example is that Z contains an intercept term, which is not in J_n, and Z_j has i.i.d. entries for $j \in J_n$ with continuous distribution and sufficiently many moments, in which case the first three conditions are easily checked and $\mathbb{E}Z_j$ is a multiple of $(1, \ldots, 1)$, which belongs to $\mathrm{span}\{Z_j : j \in J_n^c\}$.

In fact, the condition 4 allows Proposition A.3.6 to cover more general cases than the above one. For example, in a census study, a state-specific fix effect might be added into the model, i.e.

$$y_i = \alpha_{s_i} + z_i^T\beta^* + \epsilon_i$$

where s_i represents the state of subject i. In this case, Z contains a sub-block formed by z_i and a sub-block with ANOVA forms as mentioned in Example 1. The latter is usually incorporated only for adjusting group bias and not the target of inference. Then condition 4 is satisfied if only Z_{ij} has same mean in each group for each j, i.e. $\mathbb{E}Z_{ij} = \mu_{s_i,j}$.

Proof of Proposition A.3.6. By Sherman-Morison-Woodbury formula,

$$e_j^T(Z^TZ)^{-1}Z^T = \frac{Z_j^T(I - H_j)}{Z_j^T(I - H_j)Z_j}$$

where $H_j = Z_{[j]}(Z_{[j]}^TZ_{[j]})^{-1}Z_{[j]}^T$ is the projection matrix generated by $Z_{[j]}$. Then

$$S_j(Z) = \frac{\|e_j^T(Z^TZ)^{-1}Z^T\|_\infty}{\|e_j^T(Z^TZ)^{-1}Z^T\|_2} = \frac{\|Z_j^T(I - H_j)\|_\infty}{\sqrt{Z_j^T(I - H_j)Z_j}}. \tag{A.79}$$

Similar to the proofs of other examples, the strategy is to show that the numerator, as a linear contrast of Z_j, and the denominator, as a quadratic form of Z_j, are both concentrated around their means. Specifically, we will show that there exists some constants C_1 and C_2 such that

$$\max_{j\in J_n} \sup_{\substack{A\in\mathbb{R}^{n\times n}, A^2=A,\\ \mathrm{tr}(A)=n-p+1}} \left\{P\left(\|AZ_j\|_\infty > C_1n^{\frac{1}{4}}\right) + P\left(Z_j^TAZ_j < C_2n\right)\right\} = o\left(\frac{1}{n}\right). \tag{A.80}$$

If (A.80) holds, since H_j is independent of Z_j by assumptions, we have

$$\begin{aligned}
&P\left(S_j(Z) \geq \frac{C_1}{\sqrt{C_2}} \cdot n^{-\frac{1}{4}}\right) = P\left(\frac{\|Z_j^T(I-H_j)\|_\infty}{\sqrt{Z_j^T(I-H_j)Z_j}} \geq \frac{C_1}{\sqrt{C_2}} \cdot n^{-\frac{1}{4}}\right) \\
\leq & P\left(\|(I-H_j)Z_j\|_\infty > C_1 n^{\frac{1}{4}}\right) + P\left(Z_j^T(I-H_j)Z_j < C_2 n\right) \\
= & \mathbb{E}\left[P\left(\|(I-H_j)Z_j\|_\infty > C_1 n^{\frac{1}{4}}\right) \middle| Z_{[j]}\right] + \mathbb{E}\left[P\left(Z_j^T(I-H_j)Z_j < C_2 n\right) \middle| Z_{[j]}\right] \qquad \text{(A.81)} \\
\leq & \sup_{A\in\mathbb{R}^{n\times n}, A^2=A, \mathrm{tr}(A)=n-p+1} P\left(\|AZ_j\|_\infty > C_1 n^{\frac{1}{4}}\right) + P\left(Z_j^T A Z_j < C_2 n\right) \\
\leq & \max_{j\in J_n}\left\{\sup_{A\in\mathbb{R}^{n\times n}, A^2=A, \mathrm{tr}(A)=n-p+1} P\left(\|AZ_j\|_\infty > C_1 n^{\frac{1}{4}}\right) + P\left(Z_j^T A Z_j < C_2 n\right)\right\} = o\left(\frac{1}{n}\right). \qquad \text{(A.82)}
\end{aligned}$$

Thus with probability $1 - o(|J_n|/n) = 1 - o(1)$,

$$\max_{j\in J_n} S_j(Z) \leq \frac{C_1}{\sqrt{C_2}} \cdot n^{-\frac{1}{4}}$$

and hence

$$\max_{j\in J_n} S_j(Z) = O_p\left(\frac{1}{n^{\frac{1}{4}}}\right).$$

Now we prove (A.80). The proof, although looks messy, is essentially the same as the proof for other examples. Instead of relying on the exponential concentration given by the sub-gaussianity, we show the concentration in terms of higher-order moments.

In fact, for any idempotent A, the sum square of each row is bounded by 1 since

$$\sum_i A_{ij}^2 = (A^2)_{j,j} \leq \lambda_{\max}(A^2) = 1.$$

By Jensen's inequality,

$$\mathbb{E}Z_{ij}^2 \leq (\mathbb{E}|Z_{ij}|^{8+\delta})^{\frac{2}{8+\delta}}.$$

For any j, by Rosenthal's inequality (Rosenthal 1970), there exists some universal constant C such that

$$\begin{aligned}
\mathbb{E}\left|\sum_{i=1}^n A_{ij}Z_{ij}\right|^{8+\delta} &\leq C\left\{\sum_{i=1}^n |A_{ij}|^{8+\delta}\mathbb{E}|Z_{ij}|^{8+\delta} + \left(\sum_{i=1}^n A_{ij}^2\mathbb{E}Z_{ij}^2\right)^{4+\delta/2}\right\} \\
&\leq C\left\{\sum_{i=1}^n |A_{ij}|^{2}\mathbb{E}|Z_{ij}|^{8+\delta} + \left(\sum_{i=1}^n A_{ij}^2\mathbb{E}Z_{ij}^2\right)^{4+\delta/2}\right\}
\end{aligned}$$

$$\leq CM^{8+\delta}\left\{\sum_{i=1}^{n}A_{ij}^2+\left(\sum_{i=1}^{n}A_{ij}^2\right)^{4+\delta/2}\right\}\leq 2CM^{8+\delta}.$$

Let $C_1=(2CM^{8+\delta})^{\frac{1}{8+\delta}}$, then for given i, by Markov inequality,

$$P\left(\left|\sum_{i=1}^{n}A_{ij}Z_{ij}\right|>C_1n^{\frac{1}{4}}\right)\leq\frac{1}{n^{2+\delta/4}}$$

and a union bound implies that

$$P\left(\|AZ_j\|_\infty>C_1n^{\frac{1}{4}}\right)\leq\frac{1}{n^{1+\delta/4}}=o\left(\frac{1}{n}\right). \tag{A.83}$$

Now we derive a bound for $Z_j^TAZ_j$. Since $p/n\to\kappa\in(0,1)$, there exists $\tilde{\kappa}\in(0,1-\kappa)$ such that $n-p>\tilde{\kappa}n$. Then

$$\mathbb{E}Z_j^TAZ_j=\sum_{i=1}^{n}A_{ii}\mathbb{E}Z_{ij}^2>\tau^2\operatorname{tr}(A)=\tau^2(n-p+1)>\tilde{\kappa}\tau^2n. \tag{A.84}$$

To bound the tail probability, we need the following result:

Lemma A.3.7 (Bai and Silverstein (2010), Lemma 6.2)**.** *Let B be an $n\times n$ nonrandom matrix and $W=(W_1,\ldots,W_n)^T$ be a random vector of independent entries. Assume that $\mathbb{E}W_i=0$, $\mathbb{E}W_i^2=1$ and $\mathbb{E}|W_i|^k\leq\nu_k$. Then, for any $q\geq1$,*

$$E|W^TBW-\operatorname{tr}(B)|^q\leq C_q\left((\nu_4\operatorname{tr}(BB^T))^{\frac{q}{2}}+\nu_{2q}\operatorname{tr}(BB^T)^{\frac{q}{2}}\right),$$

where C_q is a constant depending on q only.

It is easy to extend Lemma A.3.7 to non-isotropic case by rescaling. In fact, denote σ_i^2 by the variance of W_i, and let $\Sigma=\operatorname{diag}(\sigma_1,\ldots,\sigma_n)$, $Y=(W_1/\sigma_1,\ldots,W_n/\sigma_n)$. Then

$$W^TBW=Y^T\Sigma^{\frac{1}{2}}B\Sigma^{\frac{1}{2}}Y,$$

with $\operatorname{Cov}(Y)=I$. Let $\tilde{B}=\Sigma^{\frac{1}{2}}B\Sigma^{\frac{1}{2}}$, then

$$\tilde{B}\tilde{B}^T=\Sigma^{\frac{1}{2}}B\Sigma B^T\Sigma^{\frac{1}{2}}\preceq\nu_2\Sigma^{\frac{1}{2}}BB^T\Sigma^{\frac{1}{2}}.$$

This entails that

$$\operatorname{tr}(\tilde{B}\tilde{B}^T)\leq nu_2\operatorname{tr}(\Sigma^{\frac{1}{2}}BB^T\Sigma^{\frac{1}{2}})=\nu_2\operatorname{tr}(\Sigma BB^T)\leq\nu_2^2\operatorname{tr}(BB^T).$$

On the other hand,

$$\operatorname{tr}(\tilde{B}\tilde{B}^T)^{\frac{q}{2}}\leq n\lambda_{\max}(\tilde{B}\tilde{B}^T)^{\frac{q}{2}}=n\nu_2^{\frac{q}{2}}\lambda_{\max}\left(\Sigma^{\frac{1}{2}}BB^T\Sigma^{\frac{1}{2}}\right)^{\frac{q}{2}}\leq n\nu_2^q\lambda_{\max}(BB^T)^{\frac{q}{2}}.$$

Thus we obtain the following result

Lemma A.3.8. *Let B be an $n \times n$ nonrandom matrix and $W = (W_1, \ldots, W_n)^T$ be a random vector of independent mean-zero entries. Suppose $\mathbb{E}|W_i|^k \le \nu_k$, then for any $q \ge 1$,*

$$E|W^T BW - \mathbb{E}W^T BW|^q \le C_q \nu_2^q \left((\nu_4 \operatorname{tr}(BB^T))^{\frac{q}{2}} + \nu_{2q} \operatorname{tr}(BB^T)^{\frac{q}{2}} \right),$$

where C_q is a constant depending on q only.

Apply Lemma A.3.8 with $W = Z_j$, $B = A$ and $q = 4 + \delta/2$, we obtain that

$$E|Z_j^T AZ_j - \mathbb{E}Z_j^T AZ_j|^{4+\delta/2} \le CM^{16+2\delta} \left((\operatorname{tr}(AA^T))^{2+\delta/4} + \operatorname{tr}(AA^T)^{2+\delta/4} \right)$$

for some constant C. Since A is idempotent, all eigenvalues of A is either 1 or 0 and thus $AA^T \preceq I$. This implies that

$$\operatorname{tr}(AA^T) \le n, \quad \operatorname{tr}(AA^T)^{2+\delta/4} \le n$$

and hence

$$E|Z_j^T AZ_j - \mathbb{E}Z_j^T AZ_j|^{4+\delta/2} \le 2CM^{16+2\delta} n^{2+\delta/4}$$

for some constant C_1, which only depends on M. By Markov inequality,

$$P\left(|Z_j^T AZ_j - \mathbb{E}Z_j^T AZ_j| \ge \frac{\tilde{\kappa}\tau^2 n}{2} \right) \le 2CM^{16+2\delta} \left(\frac{2}{\tilde{\kappa}\tau^2} \right)^{4+\delta/2} \cdot \frac{1}{n^{2+\delta/4}}.$$

Combining with (A.84), we conclude that

$$P\left(Z_j^T AZ_j < C_2 n \right) = O\left(\frac{1}{n^{2+\delta/4}} \right) = o\left(\frac{1}{n} \right) \tag{A.85}$$

where $C_2 = \frac{\tilde{\kappa}\tau^2}{2}$. Notice that both (A.83) and (A.85) do not depend on j and A. Therefore, (A.80) is proved and hence the Proposition. □

A.4 Additional Numerical Experiments

In this section, we repeat the experiments in section 2.6 by using L_1 loss, i.e. $\rho(x) = |x|$. L_1-loss is not smooth and does not satisfy our technical conditions. The results are displayed below. It is seen that the performance is quite similar to that with the huber loss.

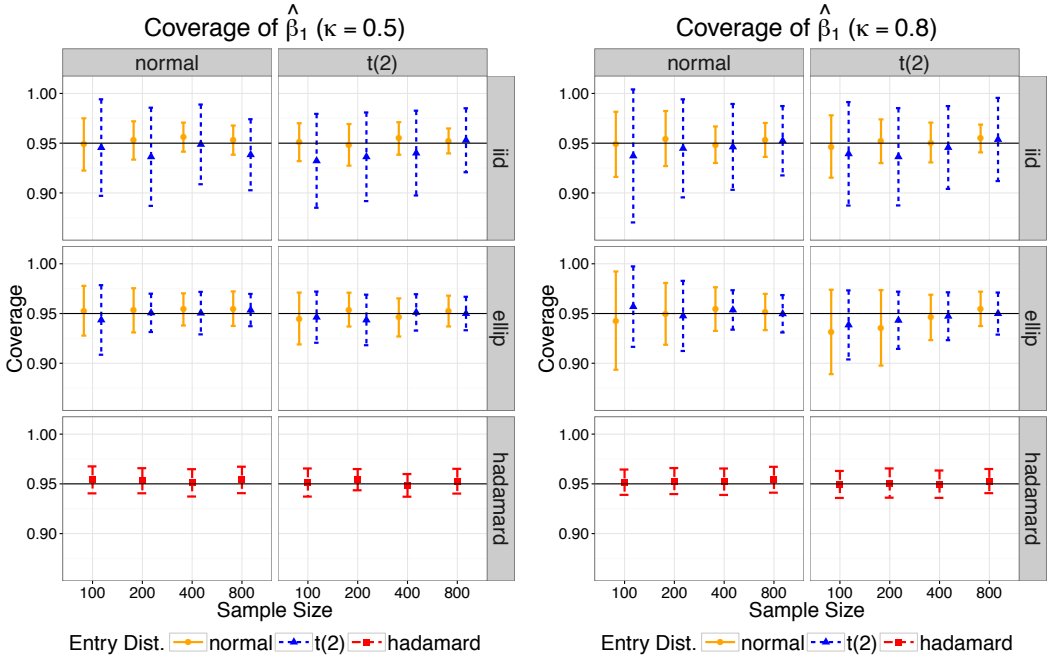


Figure A.1: Empirical 95% coverage of $\hat{\beta}_1$ with $\kappa = 0.5$ (left) and $\kappa = 0.8$ (right) using L_1 loss. The x-axis corresponds to the sample size, ranging from 100 to 800; the y-axis corresponds to the empirical 95% coverage. Each column represents an error distribution and each row represents a type of design. The orange solid bar corresponds to the case F = Normal; the blue dotted bar corresponds to the case $F = \mathrm{t}_2$; the red dashed bar represents the Hadamard design.

A.5 Miscellaneous

In this appendix we state several technical results for the sake of completeness.

Proposition A.5.1 ((Horn and Johnson 2012), formula (0.8.5.6)). *Let $A \in \mathbb{R}^{p\times p}$ be an invertible matrix and write A as a block matrix*

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$$

with $A_{11} \in \mathbb{R}^{p_1\times p_1}, A_{22} \in \mathbb{R}^{(p-p_1)\times(p-p_1)}$ being invertible matrices. Then

$$A^{-1} = \begin{pmatrix} A_{11} + A_{11}^{-1}A_{12}S^{-1}A_{21}A_{11}^{-1} & -A_{11}^{-1}A_{12}S^{-1} \\ -S^{-1}A_{21}A_{11}^{-1} & S^{-1} \end{pmatrix}$$

where $S = A_{22} - A_{21}A_{11}^{-1}A_{12}$ is the Schur's complement.

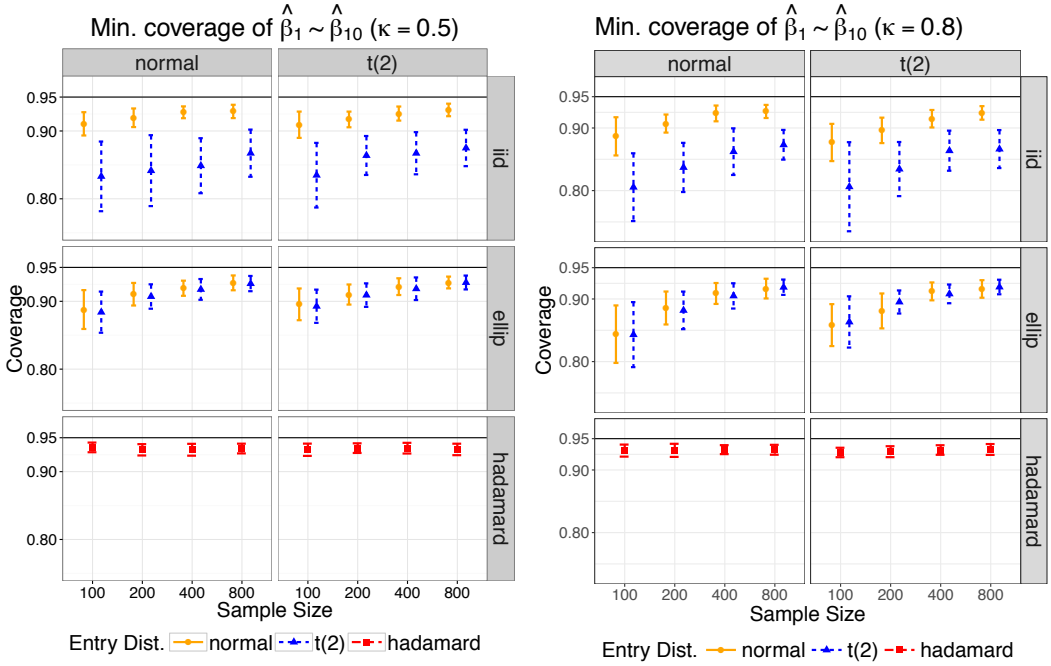


Figure A.2: Mininum empirical 95% coverage of $\hat{\beta}_1 \sim \hat{\beta}_{10}$ with $\kappa = 0.5$ (left) and $\kappa = 0.8$ (right) using L_1 loss. The x-axis corresponds to the sample size, ranging from 100 to 800; the y-axis corresponds to the minimum empirical 95% coverage. Each column represents an error distribution and each row represents a type of design. The orange solid bar corresponds to the case $F = \text{Normal}$; the blue dotted bar corresponds to the case $F = \text{t}_2$; the red dashed bar represents the Hadamard design.

Proposition A.5.2 ((Rudelson and Vershynin 2013); improved version of the original form by (Hanson and Wright 1971))**.** *Let $X = (X_1, \ldots, X_n) \in \mathbb{R}^n$ be a random vector with independent mean-zero σ^2-sub-gaussian components X_i. Then, for every t,*

$$P\left(|X^T A X - \mathbb{E}X^T A X| > t\right) \leq 2\exp\left\{-c\min\left(\frac{t^2}{\sigma^4\|A\|_F^2}, \frac{t}{\sigma^2\|A\|_{\text{op}}}\right)\right\}$$

Proposition A.5.3 ((Bai and Yin 1993))**.** *If $\{Z_{ij} : i = 1, \ldots, n, j = 1, \ldots, p\}$ are i.i.d. random variables with zero mean, unit variance and finite fourth moment and $p/n \to \kappa$, then*

$$\lambda_{\max}\left(\frac{Z^T Z}{n}\right) \overset{a.s.}{\to} (1 + \sqrt{\kappa})^2, \quad \lambda_{\min}\left(\frac{Z^T Z}{n}\right) \overset{a.s.}{\to} (1 - \sqrt{\kappa})^2.$$

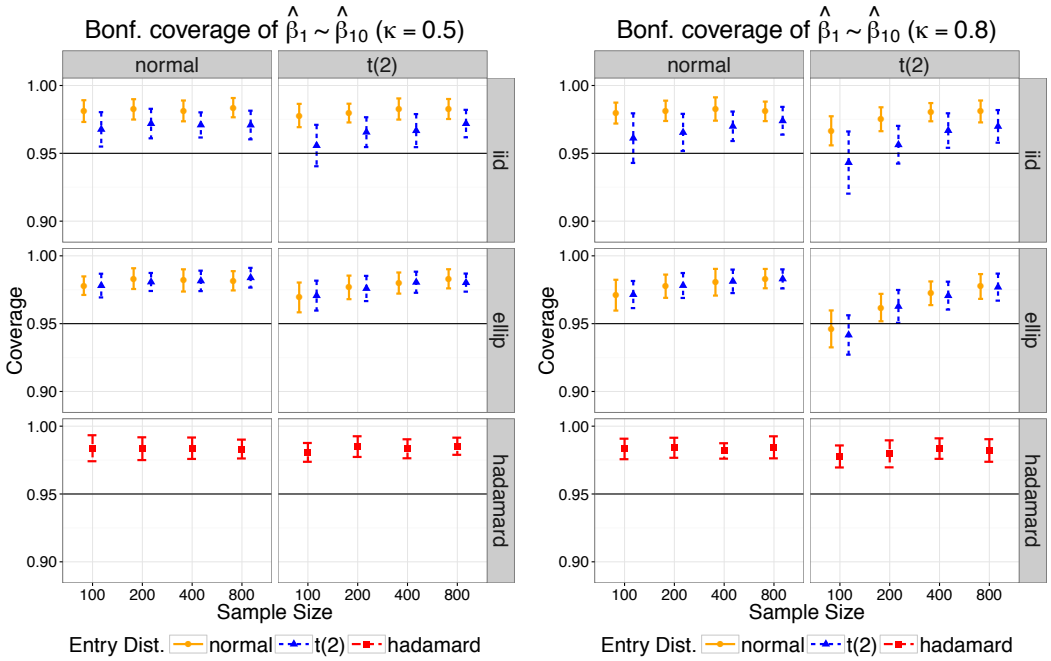


Figure A.3: Empirical 95% coverage of $\hat{\beta}_1 \sim \hat{\beta}_{10}$ after Bonferroni correction with $\kappa = 0.5$ (left) and $\kappa = 0.8$ (right) using L_1 loss. The x-axis corresponds to the sample size, ranging from 100 to 800; the y-axis corresponds to the empirical uniform 95% coverage after Bonferroni correction. Each column represents an error distribution and each row represents a type of design. The orange solid bar corresponds to the case $F = \text{Normal}$; the blue dotted bar corresponds to the case $F = \text{t}_2$; the red dashed bar represents the Hadamard design.

Proposition A.5.4 ((Latała 2005)). *Suppose* $\{Z_{ij} : i = 1, \ldots, n, j = 1, \ldots, p\}$ *are independent mean-zero random variables with finite fourth moment, then*

$$\mathbb{E}\sqrt{\lambda_{\max}(Z^T Z)} \leq C \left(\max_i \sqrt{\sum_j \mathbb{E} Z_{ij}^2} + \max_j \sqrt{\sum_i \mathbb{E} Z_{ij}^2} + \sqrt[4]{\sum_{i,j} \mathbb{E} Z_{ij}^4} \right)$$

for some universal constant C. *In particular, if* $\mathbb{E} Z_{ij}^4$ *are uniformly bounded, then*

$$\lambda_{\max}\left(\frac{Z^T Z}{n}\right) = O_p\left(1 + \sqrt{\frac{p}{n}}\right).$$

Proposition A.5.5 ((Rudelson and Vershynin 2010)). *Suppose* $\{Z_{ij} : i = 1, \ldots, n, j = 1, \ldots, p\}$ *are independent mean-zero* σ^2*-sub-gaussian random variables. Then there exists a*

universal constant $C_1, C_2 > 0$ *such that*

$$P\left(\sqrt{\lambda_{\max}\left(\frac{Z^TZ}{n}\right)} > C\sigma\left(1+\sqrt{\frac{p}{n}}+t\right)\right) \leq 2e^{-C_2nt^2}.$$

Proposition A.5.6 ((Rudelson and Vershynin 2009))**.** *Suppose* $\{Z_{ij} : i = 1,\ldots,n, j = 1,\ldots,p\}$ *are i.i.d.* σ^2*-sub-gaussian random variables with zero mean and unit variance, then for* $\epsilon \geq 0$

$$P\left(\sqrt{\lambda_{\min}\left(\frac{Z^TZ}{n}\right)} \leq \epsilon(1-\sqrt{\frac{p-1}{n}})\right) \leq (C\epsilon)^{n-p+1} + e^{-cn}$$

for some universal constants C *and* c*.*

Proposition A.5.7 ((Litvak et al. 2005))**.** *Suppose* $\{Z_{ij} : i = 1,\ldots,n, j = 1,\ldots,p\}$ *are independent* σ^2*-sub-gaussian random variables such that*

$$Z_{ij} \stackrel{d}{=} -Z_{ij}, \quad \mathrm{Var}(Z_{ij}) > \tau^2$$

for some $\sigma, \tau > 0$*, and* $p/n \to \kappa \in (0,1)$*, then there exists constants* $c_1, c_2 > 0$*, which only depends on* σ *and* τ*, such that*

$$P\left(\lambda_{\min}\left(\frac{Z^TZ}{n}\right) < c_1\right) \leq e^{-c_2n}.$$

Appendix B

Appendix for Chapter 3

B.1 Complementary Experimental Results

In this appendix we present experimental results that complement Section 3.3. Figure B.1 - B.4 display the power comparison for testing a single coordinate under the same setting as subsection 3.3.2 for four extra scenarios: realizations of Gaussian matrices + Cauchy errors, realizations of Cauchy matrices + Gaussian errors and realizations of random one-way ANOVA matrices + Gaussian/Cauchy errors.

Figure B.5 - B.10 display the power results under the same setting as subsection 3.3.3 for six scenarios realizations of Gaussian matrices + Gaussian/Cauchy errors, realizations of Cauchy matrices + Gaussian/Cauchy errors and realizations of random one-way ANOVA matrices + Gaussian/Cauchy errors.

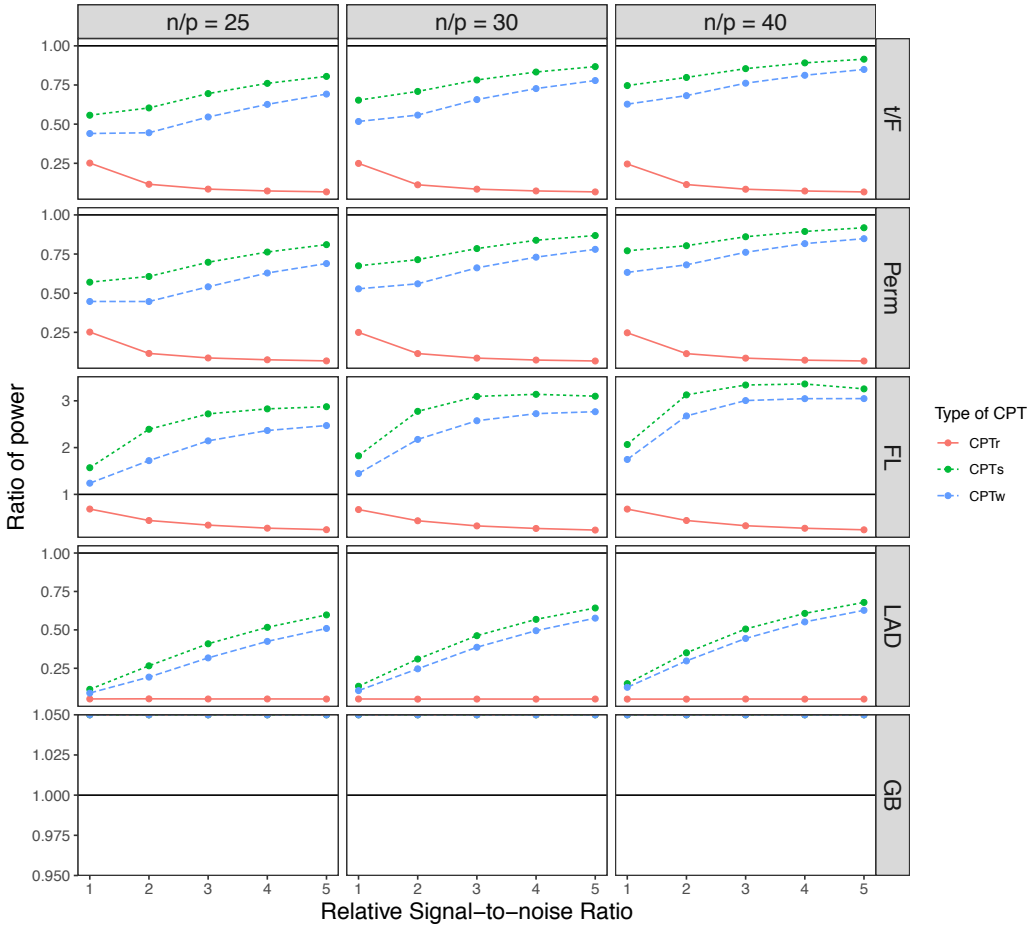


Figure B.1: Median power ratio between each variant of CPT and each competing test for testing a single coordinate with realizations of Gaussian matrices and Cauchy errors. The black solid line marks the equal power. The missing values in the last row correspond to infinite ratios.

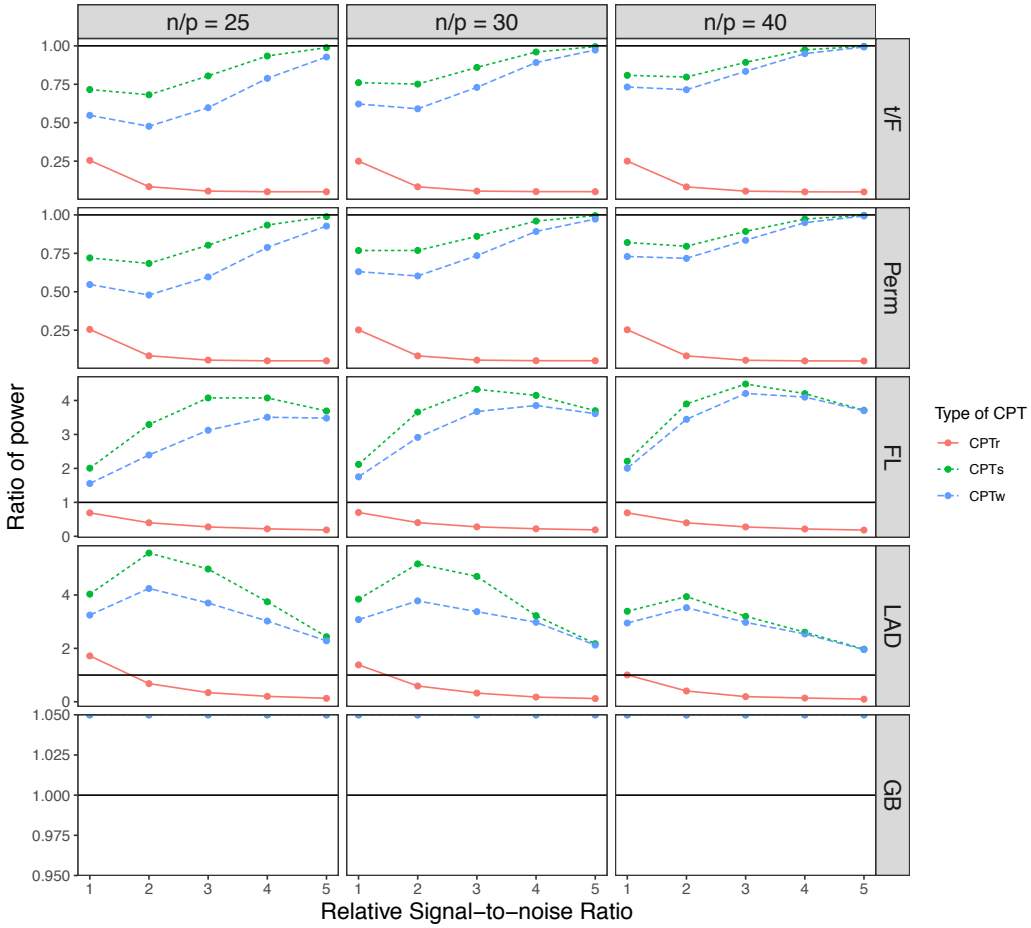


Figure B.2: Median power ratio between each variant of CPT and each competing test for testing a single coordinate with realizations of Cauchy matrices and Gaussian errors. The black solid line marks the equal power. The missing values in the last row correspond to infinite ratios.

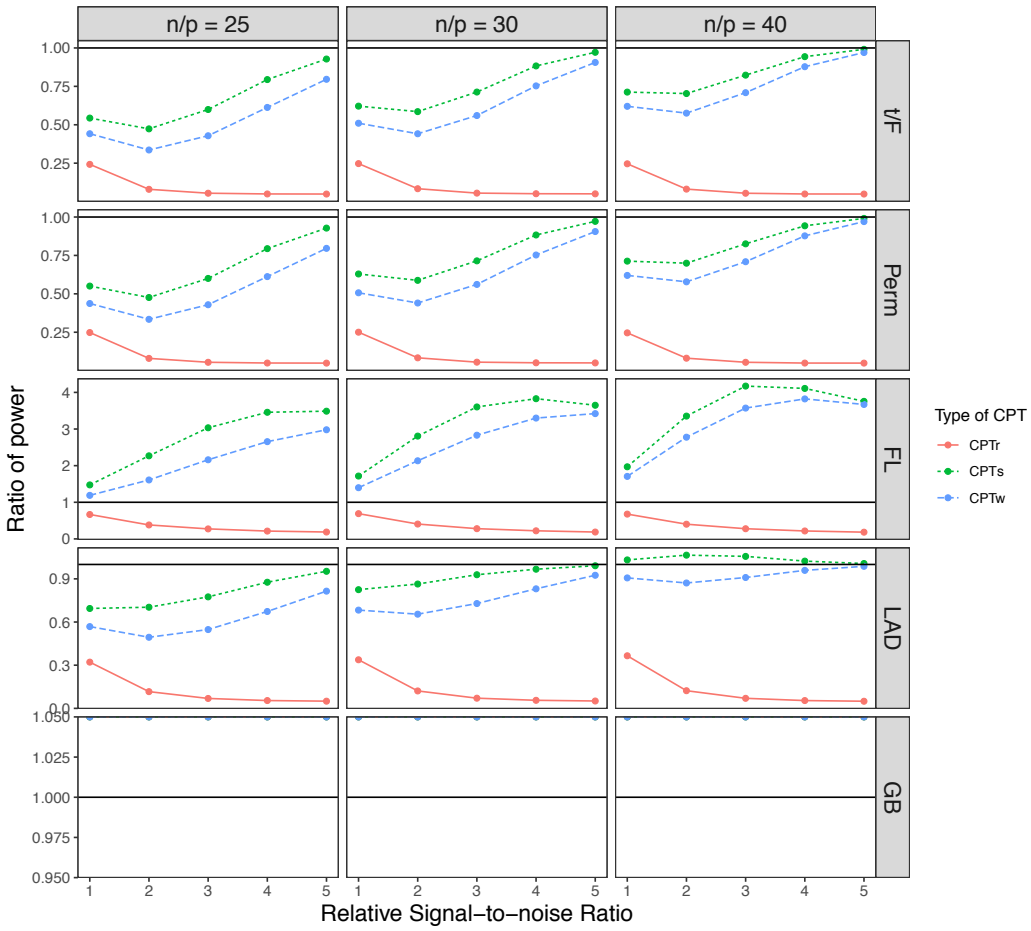


Figure B.3: Median power ratio between each variant of CPT and each competing test for testing a single coordinate with realizations of random one-way ANOVA matrices and Gaussian errors. The black solid line marks the equal power. The missing values in the last row correspond to infinite ratios.

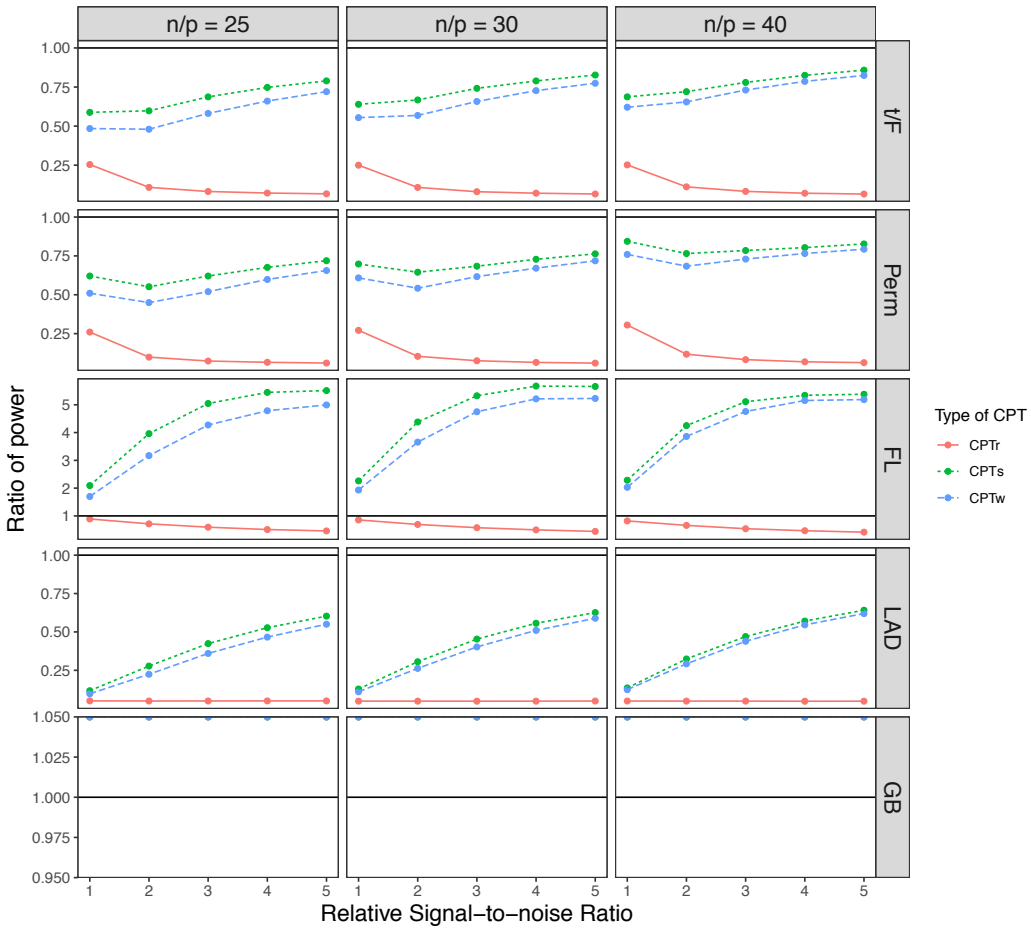


Figure B.4: Median power ratio between each variant of CPT and each competing test for testing a single coordinate with realizations of random one-way ANOVA matrices and Cauchy errors. The black solid line marks the equal power. The missing values in the last row correspond to infinite ratios.

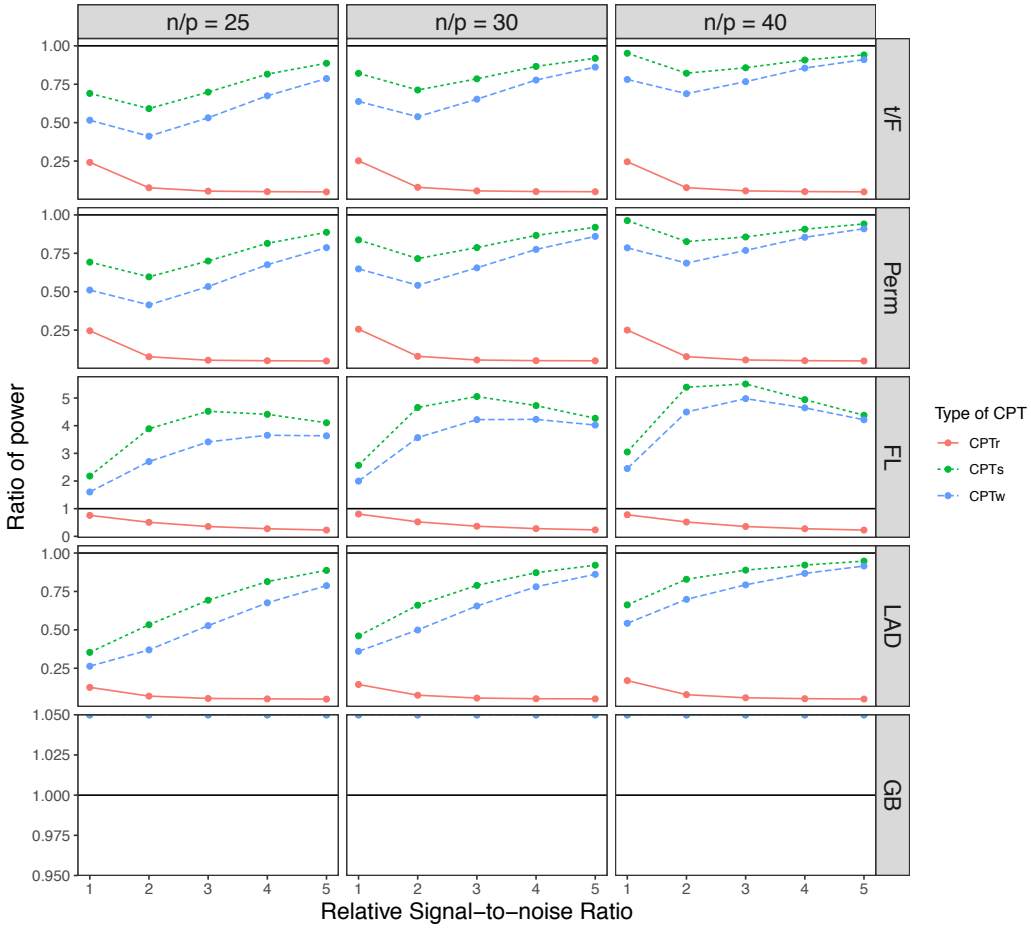


Figure B.5: Median power ratio between each variant of CPT and each competing test for testing five coordinates with realizations of Gaussian matrices and Gaussian errors. The black solid line marks the equal power. The missing values in the last row correspond to infinite ratios.

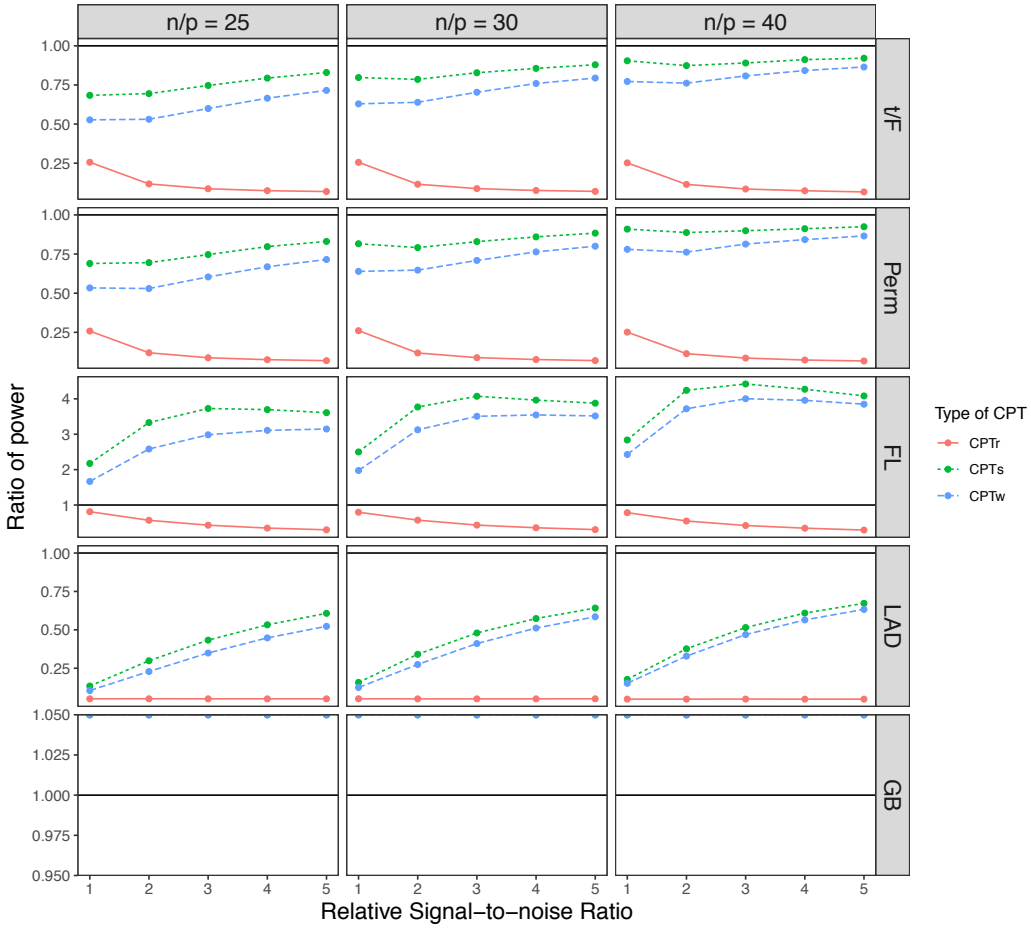


Figure B.6: Median power ratio between each variant of CPT and each competing test for testing five coordinates with realizations of Gaussian matrices and Cauchy errors. The black solid line marks the equal power. The missing values in the last row correspond to infinite ratios.

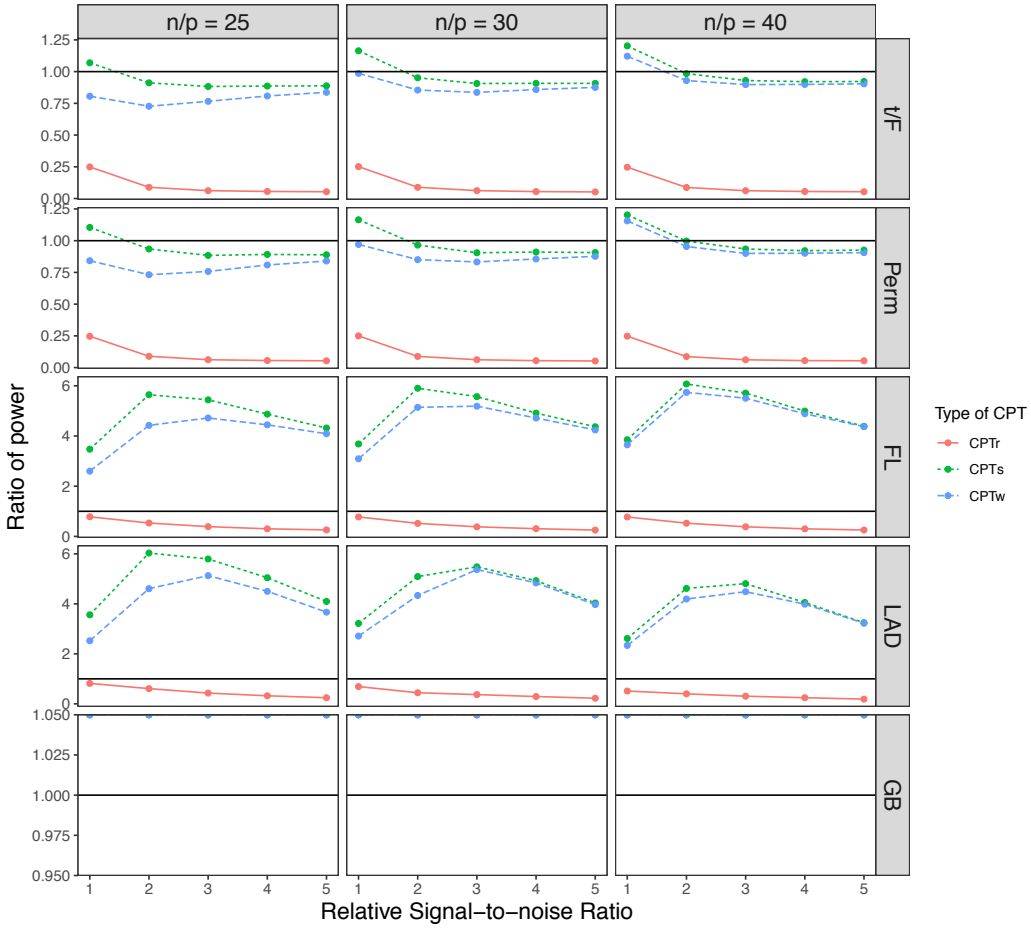


Figure B.7: Median power ratio between each variant of CPT and each competing test for testing five coordinates with realizations of Cauchy matrices and Gaussian errors. The black solid line marks the equal power. The missing values in the last row correspond to infinite ratios.

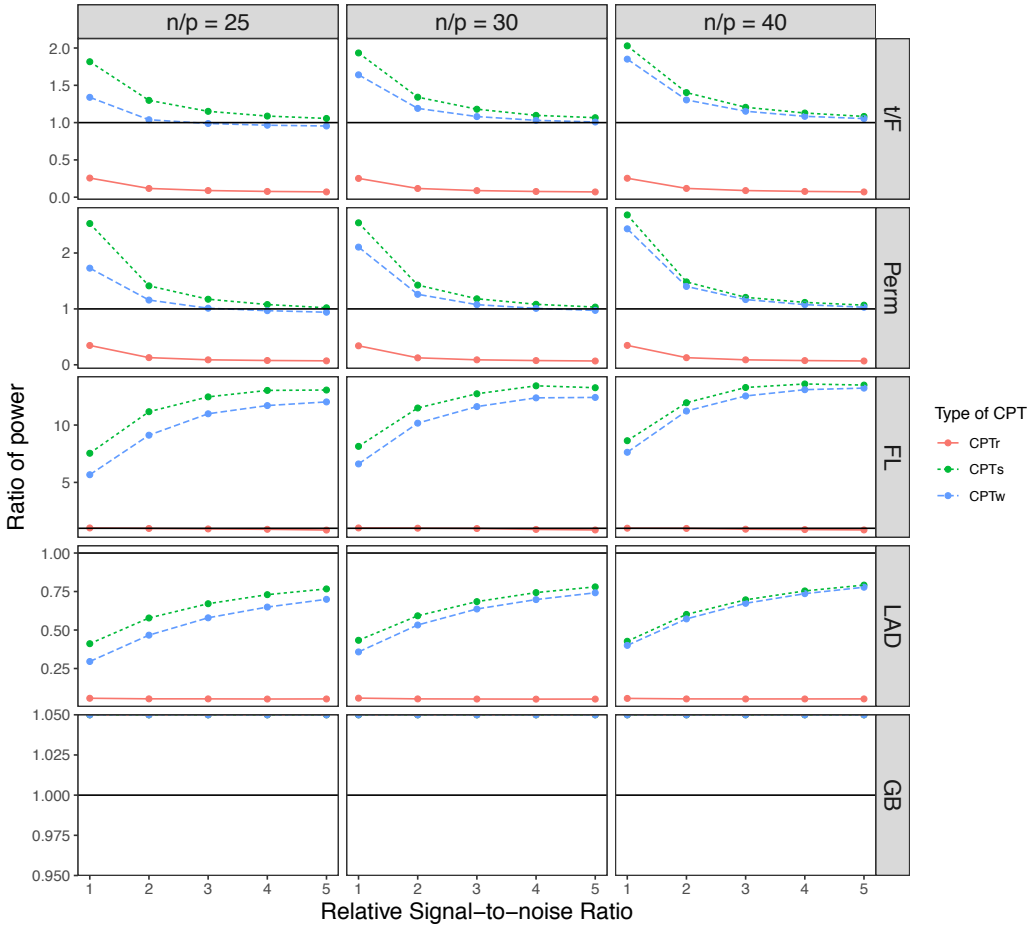


Figure B.8: Median power ratio between each variant of CPT and each competing test for testing five coordinates with realizations of Cauchy matrices and Cauchy errors. The black solid line marks the equal power. The missing values in the last row correspond to infinite ratios.

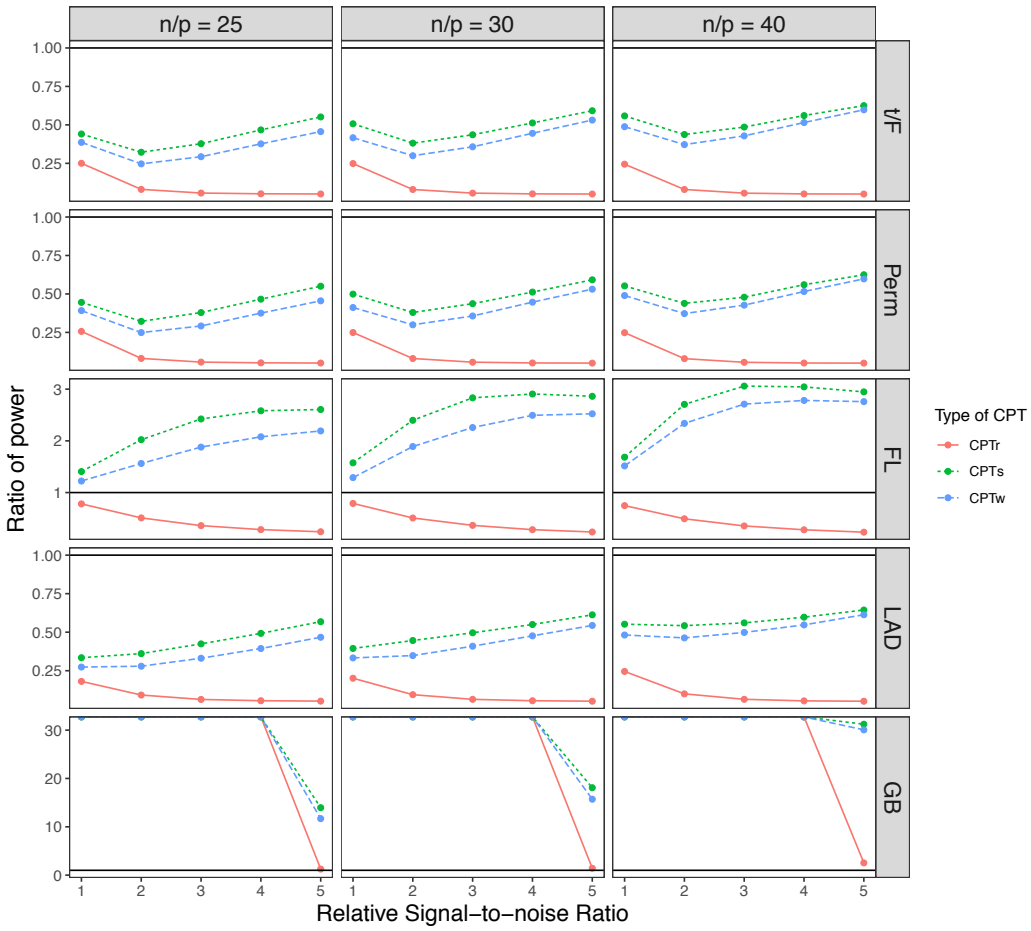


Figure B.9: Median power ratio between each variant of CPT and each competing test for testing five coordinates with realizations of random one-way ANOVA matrices and Gaussian errors. The black solid line marks the equal power. The missing values in the last row correspond to infinite ratios.

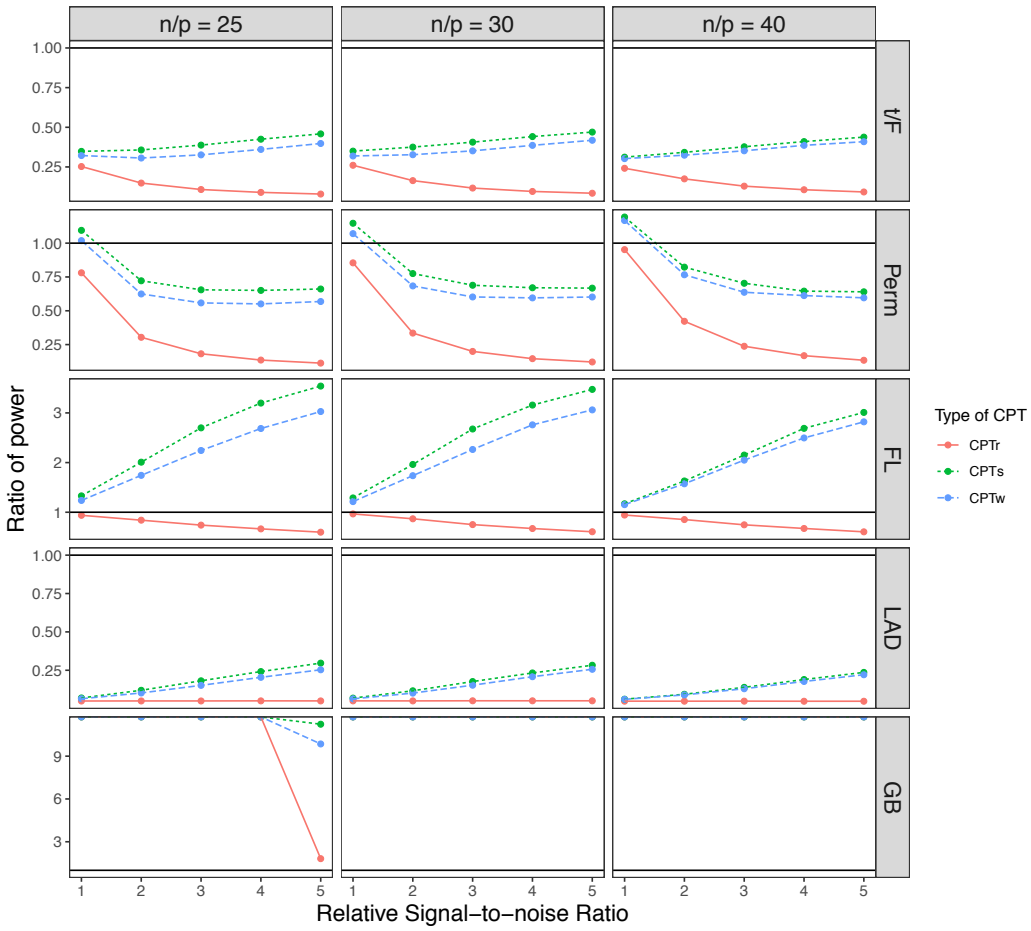


Figure B.10: Median power ratio between each variant of CPT and each competing test for testing five coordinates with realizations of random one-way ANOVA matrices and Cauchy errors. The black solid line marks the equal power. The missing values in the last row correspond to infinite ratios.

Appendix C

Appendix for Chapter 4

C.1 Concentration Inequalities for Sampling Without Replacement

C.1.1 Some existing tools

The proofs rely on concentration inequalities for sampling without replacement. Hoeffding (1963, Theorem 4) proved the following result that sampling without replacement is more concentrated in convex ordering than i.i.d. sampling.

Proposition C.1.1. *Let $C = (c_1, \ldots, c_n)$ be a finite population of fixed elements. Let $Z_1, \ldots, Z_m$ be a random sample with replacement from C and $W_1, \ldots, W_m$ be a random sample without replacement from C. If the function $f(x)$ is continuous and convex, then*

$$\mathbb{E}f\left(\sum_{i=1}^{m} Z_i\right) \geq \mathbb{E}f\left(\sum_{i=1}^{m} W_i\right).$$

From Proposition C.1.1, most concentration inequalities for independent sampling carry over to sampling without replacement. Later a line of works, in different contexts, showed an even more surprising phenomenon that sampling without replacement can have strictly better concentration than independent sampling (e.g., Serfling 1974; Diaconis and Shahshahani 1987; Lee and Yau 1998; Bobkov 2004; Cortes et al. 2009; El-Yaniv and Pechyony 2009; Bardenet and Maillard 2015; Tolstikhin 2017). In particular, Tolstikhin (2017, Theorem 9) proved a useful concentration inequality for the empirical processes for sampling without replacement.

Proposition C.1.2. *Let $C = (c_1, \ldots, c_n)$ be a finite population of fixed elements, and $W_1, \ldots, W_m$ be a random sample without replacement from C. Let $\mathcal{F}$ be a class of functions on C, and*

$$S(\mathcal{F}) = \sup_{f \in \mathcal{F}} \sum_{i=1}^{m} f(W_i), \quad \nu(\mathcal{F})^2 = \sup_{f \in \mathcal{F}} \mathrm{Var}(f(W_1)).$$

Then

$$\mathbb{P}(S(\mathcal{F}) - \mathbb{E}[S(\mathcal{F})] \geq t) \leq \exp\left\{-\frac{(n+2)t^2}{8n^2\nu(\mathcal{F})^2}\right\}.$$

Proposition C.1.2 gives a sub-gaussian tail of $S(\mathcal{F})$ with the sub-gaussian parameter depending solely on the variance. In contrast, the concentration inequalities in the standard empirical process theory for independent sampling usually requires the functions in $\mathcal{F}$ to be uniformly bounded and the tail is either sub-gaussian with the sub-gaussian parameter being the uniform bound on $\mathcal{F}$ or sub-exponential with Bernstein-style behaviors; see Boucheron et al. (2013) for instance. Therefore, Proposition C.1.2 provides a more precise statement that sampling without replacement is more concentrated than independent sampling for a large class of statistics.

We need the following result from Tropp (2016, Theorem 5.1.(2)) to prove the matrix concentration inequality.

Proposition C.1.3. *Let $\tilde{V}_1, \ldots, \tilde{V}_m$ be independent $p \times p$ random matrices with $\mathbb{E}\tilde{V}_i = 0$ for all i. Let $C(p) = 4(1 + \lceil 2\log p \rceil)$. Then*

$$\left(\mathbb{E}\left\|\sum_{i=1}^n \tilde{V}_i\right\|_{\text{op}}^2\right)^{\frac{1}{2}} \leq \sqrt{C(p)}\left\|\sum_{i=1}^n \mathbb{E}\tilde{V}_i^2\right\|_{\text{op}}^{\frac{1}{2}} + C(p)\left(\mathbb{E}\max_{1\leq i\leq n}\|\tilde{V}_i\|_{\text{op}}^2\right)^{\frac{1}{2}}.$$

We will also use the facts that for any $u \in \mathbb{R}^p$ and Hermitian $V \in \mathbb{R}^{p\times p}$,

$$\|u\|_2 = \sup_{\omega\in\mathcal{S}^{p-1}} u^\mathsf{T}\omega, \quad \|V\|_{\text{op}} = \sup_{\omega\in\mathcal{S}^{p-1}} \omega^\mathsf{T} V\omega.$$

C.1.2 Proofs of Lemmas 4.6.3 and 4.6.4

Proof of Lemma 4.6.3. Let

$$C = (u_1, \ldots, u_n), \quad \text{and} \quad \mathcal{F} = \{f_\omega(u) = u^\mathsf{T}\omega : \omega \in \mathcal{S}^{p-1}\}.$$

Let u be a vector that is randomly sampled from C. Then

$$\begin{aligned}
\nu^2(\mathcal{F}) &= \sup_{\omega\in\mathcal{S}^{p-1}} \text{Var}(u^\mathsf{T}\omega) \leq \sup_{\omega\in\mathcal{S}^{p-1}} \mathbb{E}(u^\mathsf{T}\omega)^2 \\
&= \sup_{\omega\in\mathcal{S}^{p-1}} \frac{1}{n}\sum_{i=1}^n (u_i^\mathsf{T}\omega)^2 = \sup_{\omega\in\mathcal{S}^{p-1}} \omega^\mathsf{T}\left(\frac{1}{n}\sum_{i=1}^n u_i u_i^\mathsf{T}\right)\omega \\
&= \sup_{\omega\in\mathcal{S}^{p-1}} \omega^\mathsf{T}\left(\frac{U^\mathsf{T}U}{n}\right)\omega \leq \frac{\|U\|_{\text{op}}^2}{n}.
\end{aligned}$$

By Proposition C.1.2,

$$\mathbb{P}\left(\left\|\sum_{i\in\mathcal{T}} u_i\right\|_2 \geq \mathbb{E}\left\|\sum_{i\in\mathcal{T}} u_i\right\|_2 + t\right) \leq \exp\left\{-\frac{(n+2)t^2}{8n\|U\|_{\text{op}}^2}\right\} \leq \exp\left\{-\frac{t^2}{8\|U\|_{\text{op}}^2}\right\},$$

or, equivalently, with probability $1-\delta$,

$$\left\|\sum_{i\in\mathcal{T}} u_i\right\|_2 \le \mathbb{E}\left\|\sum_{i\in\mathcal{T}} u_i\right\|_2 + \|U\|_{\mathrm{op}}\sqrt{8\log\frac{1}{\delta}}. \tag{S1}$$

By the Cauchy–Schwarz inequality,

$$\left(\mathbb{E}\left\|\sum_{i\in\mathcal{T}} u_i\right\|_2\right)^2 \le \mathbb{E}\left\|\sum_{i\in\mathcal{T}} u_i\right\|_2^2 = \sum_{j=1}^{p}\mathbb{E}\left(\sum_{i\in\mathcal{T}} u_{ij}\right)^2.$$

Lemma 4.6.1 implies

$$\mathbb{E}\left(\sum_{i\in\mathcal{T}} u_{ij}\right)^2 = \frac{m(n-m)}{n(n-1)}\sum_{i=1}^{n} u_{ij}^2.$$

As a result,

$$\left(\mathbb{E}\left\|\sum_{i\in\mathcal{T}} u_i\right\|_2\right)^2 \le \frac{m(n-m)}{n(n-1)}\sum_{i=1}^{n}\|u_i\|_2^2 = \|U\|_F^2\,\frac{m(n-m)}{n(n-1)}. \tag{S2}$$

We complete the proof by using (S1) and (S2). □

Proof of Lemma 4.6.4. Let

$$C = (V_1,\ldots,V_n), \quad \text{and} \quad \mathcal{F} = \{f_\omega(V) = \omega^{\mathsf{T}}V\omega : \omega\in\mathcal{S}^{p-1}\}.$$

Let V be a vector that is randomly sampled from C. Then

$$\nu^2(\mathcal{F}) = \sup_{\omega\in\mathcal{S}^{p-1}}\mathrm{Var}(\omega^{\mathsf{T}}V\omega) \le \sup_{\omega\in\mathcal{S}^{p-1}}\mathbb{E}(\omega^{\mathsf{T}}V\omega)^2 = \sup_{\omega\in\mathcal{S}^{p-1}}\frac{1}{n}\sum_{i=1}^{n}(\omega^{\mathsf{T}}V_i\omega)^2 = \nu_-^2.$$

By Proposition C.1.2,

$$\mathbb{P}\left(\left\|\sum_{i\in\mathcal{T}} V_i\right\|_{\mathrm{op}} \ge \mathbb{E}\left\|\sum_{i\in\mathcal{T}} V_i\right\|_{\mathrm{op}} + t\right) \le \exp\left\{-\frac{(n+2)t^2}{8n^2\nu_-^2}\right\} \le \exp\left\{-\frac{t^2}{8n\nu_-^2}\right\},$$

or, equivalently, with probability $1-\delta$,

$$\left\|\sum_{i\in\mathcal{T}} V_i\right\|_{\mathrm{op}} \le \mathbb{E}\left\|\sum_{i\in\mathcal{T}} V_i\right\|_{\mathrm{op}} + \sqrt{8n\log\frac{1}{\delta}}\,\nu_-. \tag{S3}$$

We then bound $\mathbb{E}\left\|\sum_{i\in\mathcal{T}} V_i\right\|_{\mathrm{op}}$. Let $\tilde{V}_1,\ldots,\tilde{V}_m$ be an i.i.d. random sample with replacement from C. We have

$$\mathbb{E}\left\|\sum_{i\in\mathcal{T}} V_i\right\|_{\mathrm{op}} \le \mathbb{E}\left\|\sum_{i=1}^{m}\tilde{V}_i\right\|_{\mathrm{op}} \le \left(\mathbb{E}\left\|\sum_{i=1}^{m}\tilde{V}_i\right\|_{\mathrm{op}}^2\right)^{\frac{1}{2}} \le \sqrt{nC(p)}\nu + C(p)\nu_+, \tag{S4}$$

where the first inequality follows from Proposition C.1.1 due to the convexity of $\|\cdot\|_{\mathrm{op}}$, the second inequality follows from the Cauchy–Schwarz inequality, and the third inequality follows from Proposition C.1.3.

Combining (S3) and (S4), we complete the proof. □

C.2 Mean and Variance of the Sum of Random Rows and Columns of a Matrix

We give a full proof of Lemma 4.6.5. When $m = 0$ or $m = n$, $Q_{\mathcal{T}}$ is deterministic with zero variance and the inequality holds automatically. Thus we assume $1 \le m \le n-1$.

Let $\sum_{[i_1,\dots,i_k]}$ denote the sum over all $(i_1,\dots,i_k)$ with mutually distinct elements in $\{1,\dots,n\}$. For instance, $\sum_{[i,j]}$ denotes the sum over all pairs (i,j) with $i \neq j$. We first state a basic result for sampling without replacement.

Lemma C.2.1. *Let $i_1,\dots,i_k$ be distinct indices in $\{1,\dots,n\}$ and $\mathcal{T}$ be a uniformly random subset of $\{1,\dots,n\}$ with size m. Then*

$$\mathbb{P}\left(i_1,\dots,i_k \in \mathcal{T}\right) = \frac{m\cdots(m-k+1)}{n\cdots(n-k+1)}.$$

By definition,

$$Q_{\mathcal{T}} = \sum_{i=1}^{n} Q_{ii} I(i \in \mathcal{T}) + \sum_{[i,j]} Q_{ij} I(i,j \in \mathcal{T}). \tag{S5}$$

The mean of $Q_{\mathcal{T}}$ follows directly from Lemma C.2.1:

$$\begin{aligned}\mathbb{E}Q_{\mathcal{T}} &= \sum_{i=1}^{n} Q_{ii}\cdot\frac{m}{n} + \sum_{[i,j]} Q_{ij}\cdot\frac{m(m-1)}{n(n-1)} \\ &= \frac{m(n-m)}{n(n-1)}\operatorname{tr}(Q) + \frac{m(m-1)}{n(n-1)}(\mathbf{1}^{\mathsf{T}}Q\mathbf{1}).\end{aligned}$$

The rest of this section proves the result of the variance. Let

$$\begin{aligned}
c_1 &= \frac{m(n-m)}{n(n-1)}, \quad c_2 = \operatorname{Var}\left(I(1,2\in\mathcal{T})\right) = c_1\frac{(m-1)(n+m-1)}{n(n-1)}, \\
c_3 &= \operatorname{Cov}\left(I(1,2\in\mathcal{T}), I(1,3\in\mathcal{T})\right) = c_1\frac{(m-1)(mn-2m-2n+2)}{n(n-1)(n-2)}, \\
c_4 &= \operatorname{Cov}\left(I(1,2\in\mathcal{T}), I(3,4\in\mathcal{T})\right) = c_1\frac{(m-1)(-4mn+6n+6m-6)}{n(n-1)(n-2)(n-3)}, \\
c_5 &= \operatorname{Cov}\left(I(1\in\mathcal{T}), I(1,2\in\mathcal{T})\right) = c_1\frac{m-1}{n},
\end{aligned}$$

$$c_6 = \mathrm{Cov}\left(I(1 \in \mathcal{T}), I(2, 3 \in \mathcal{T})\right) = -c_1 \frac{2(m-1)}{n(n-2)}.$$

Using (S5), we have

$$\mathrm{Var}(Q_{\mathcal{T}}) = \underbrace{\mathrm{Var}\left(\sum_{i=1}^{n} Q_{ii} I(i \in \mathcal{T})\right)}_{V_{\mathrm{I}}} + \underbrace{\mathrm{Var}\left(\sum_{[i,j]} Q_{ij} I(i, j \in \mathcal{T})\right)}_{V_{\mathrm{II}}} + \underbrace{2\,\mathrm{Cov}\left(\sum_{i=1}^{n} Q_{ii} I(i \in \mathcal{T}), \sum_{[i,j]} Q_{ij} I(i, j \in \mathcal{T})\right)}_{V_{\mathrm{III}}}. \tag{S6}$$

The next subsection deals with the three terms in (S6), separately.

C.2.1 Simplifying (S6)

Term V_{I} Lemma 4.6.1 implies

$$\begin{aligned} V_{\mathrm{I}} = \mathrm{Var}\left(\sum_{i=1}^{n} Q_{ii} I(i \in \mathcal{T})\right) &= \frac{m(n-m)}{n(n-1)} \sum_{i=1}^{n} \left(Q_{ii} - \frac{1}{n}\sum_{i=1}^{n} Q_{ii}\right)^2 \\ &= c_1 \sum_{i=1}^{n} Q_{ii}^2 - \frac{c_1}{n}(\mathrm{tr}(Q))^2. \end{aligned} \tag{S7}$$

Term V_{II} We expand V_{II} as

$$\begin{aligned} V_{\mathrm{II}} &= \mathrm{Var}\left(\sum_{[i,j]} Q_{ij} I(i, j \in \mathcal{T})\right) = \mathrm{Cov}\left(\sum_{[i,j]} Q_{ij} I(i, j \in \mathcal{T}), \sum_{[i',j']} Q_{i'j'} I(i', j' \in \mathcal{T})\right) \\ &= \sum_{[i,j]} \left(Q_{ij}^2 + Q_{ij}Q_{ji}\right) \mathrm{Var}(I(i, j \in \mathcal{T})) + \sum_{[i,j,k,\ell]} Q_{ij}Q_{k\ell}\,\mathrm{Cov}\left(I(i, j \in \mathcal{T}), I(k, \ell \in \mathcal{T})\right) \\ &\quad + \sum_{[i,j,k]} \left(Q_{ij}Q_{ik} + Q_{ij}Q_{ki}\right) \mathrm{Cov}\left(I(i, j \in \mathcal{T}), I(i, k \in \mathcal{T})\right) \\ &\quad + \sum_{[i,j,k]} \left(Q_{ij}Q_{jk} + Q_{ij}Q_{kj}\right) \mathrm{Cov}\left(I(i, j \in \mathcal{T}), I(j, k \in \mathcal{T})\right) \\ &= c_2 \sum_{[i,j]} \left(Q_{ij}^2 + Q_{ij}Q_{ji}\right) + c_4 \sum_{[i,j,k,\ell]} Q_{ij}Q_{k\ell} \\ &\quad + c_3 \sum_{[i,j,k]} \left(Q_{ij}Q_{ik} + Q_{ij}Q_{ki} + Q_{ij}Q_{jk} + Q_{ij}Q_{kj}\right). \end{aligned}$$

We then reduce the summation over $[i,j,k,l]$ to summations over fewer indices. First,

$$\left(\sum_{[i,j]} Q_{ij}\right)^2 = \sum_{[i,j]} \left(Q_{ij}^2 + Q_{ij}Q_{ji}\right) + \sum_{[i,j,k,\ell]} Q_{ij}Q_{k\ell} + \sum_{[i,j,k]} \left(Q_{ij}Q_{ik} + Q_{ij}Q_{ki} + Q_{ij}Q_{jk} + Q_{ij}Q_{kj}\right).$$

Second, $\mathbf{1}^\mathsf{T} Q \mathbf{1} = 0$ implies $\sum_{[i,j]} Q_{ij} = -\sum_{i=1}^n Q_{ii} = -\operatorname{tr}(Q)$, which further implies

$$\sum_{[i,j,k,\ell]} Q_{ij}Q_{k\ell} = (\operatorname{tr}(Q))^2 - \sum_{[i,j]} \left(Q_{ij}^2 + Q_{ij}Q_{ji}\right) - \sum_{[i,j,k]} \left(Q_{ij}Q_{ik} + Q_{ij}Q_{ki} + Q_{ij}Q_{jk} + Q_{ij}Q_{kj}\right).$$

The above two facts simplify V_{II} to

$$V_{\mathrm{II}} = c_4(\operatorname{tr}(Q))^2 + (c_2 - c_4)\sum_{[i,j]} \left(Q_{ij}^2 + Q_{ij}Q_{ji}\right) + (c_3 - c_4)\sum_{[i,j,k]} \left(Q_{ij}Q_{ik} + Q_{ij}Q_{ki} + Q_{ij}Q_{jk} + Q_{ij}Q_{kj}\right). \tag{S8}$$

We then reduce the summation over $[i,j,k]$ to summations over fewer indices. Note that $\mathbf{1}^\mathsf{T} Q = Q\mathbf{1} = 0$ implies $\sum_{j=1}^n Q_{ij} = \sum_{i=1}^n Q_{ij} = 0$, which further implies

$$\sum_{[i,j,k]} Q_{ij}Q_{ik} = \sum_{[i,j]} Q_{ij} \sum_{k\neq i,j} Q_{ik} = -\sum_{[i,j]} Q_{ij}(Q_{ii} + Q_{ij}) = -\sum_{i=1}^n Q_{ii}\sum_{j\neq i} Q_{ij} - \sum_{[i,j]} Q_{ij}^2 = \sum_{i=1}^n Q_{ii}^2 - \sum_{[i,j]} Q_{ij}^2.$$

Similarly,

$$\sum_{[i,j,k]} Q_{ij}Q_{kj} = \sum_{i=1}^n Q_{ii}^2 - \sum_{[i,j]} Q_{ij}^2,$$

$$\sum_{[i,j,k]} Q_{ij}Q_{ki} = \sum_{[i,j,k]} Q_{ij}Q_{jk} = \sum_{i=1}^n Q_{ii}^2 - \sum_{[i,j]} Q_{ij}Q_{ji}.$$

Using the above three identities to simplify the third term in (S8), we obtain

$$V_{\mathrm{II}} = c_4(\operatorname{tr}(Q))^2 + 4(c_3 - c_4)\sum_{i=1}^n Q_{ii}^2 + (c_2 - 2c_3 + c_4)\sum_{[i,j]} \left(Q_{ij}^2 + Q_{ij}Q_{ji}\right). \tag{S9}$$

Term $V_{\mathbf{III}}$ The covariance term is

$$\begin{aligned} V_{\text{III}} &= \text{Cov}\left(\sum_{i=1}^{n} Q_{ii} I(i \in \mathcal{T}), \sum_{[i,j]} Q_{ij} I(i,j \in \mathcal{T})\right) \\ &= \sum_{[i,j]} Q_{ii}(Q_{ij}+Q_{ji}) \,\text{Cov}\left(I(i \in \mathcal{T}), I(i,j \in \mathcal{T})\right) \\ &\qquad + \sum_{[i,j,k]} Q_{ii} Q_{jk} \,\text{Cov}\left(I(i \in \mathcal{T}), I(j,k \in \mathcal{T})\right) \\ &= c_5 \sum_{[i,j]} Q_{ii}(Q_{ij}+Q_{ji}) + c_6 \sum_{[i,j,k]} Q_{ii} Q_{jk}. \end{aligned}$$

Similar to previous arguments,

$$\begin{aligned} \sum_{[i,j]} Q_{ii}(Q_{ij}+Q_{ji}) &= \sum_{i=1}^{n} Q_{ii} \sum_{j \neq i} (Q_{ij}+Q_{ji}) = -2\sum_{i=1}^{n} Q_{ii}^2, \\ \sum_{[i,j,k]} Q_{ii} Q_{jk} &= \sum_{[i,j]} Q_{ii} \sum_{k \neq i,j} Q_{jk} = -\sum_{[i,j]} Q_{ii}(Q_{jj}+Q_{ji}) \\ &= -\sum_{i=1}^{n} Q_{ii} \sum_{j \neq i} (Q_{jj}+Q_{ji}) = -\sum_{i=1}^{n} Q_{ii}\left(\text{tr}(Q) - Q_{ii} - Q_{ii}\right) \\ &= -(\text{tr}(Q))^2 + 2\sum_{i=1}^{n} Q_{ii}^2. \end{aligned}$$

Using the above two identities, we can simplify V_{III} to

$$V_{\text{III}} = -c_6(\text{tr}(Q))^2 - 2(c_5 - c_6)\sum_{i=1}^{n} Q_{ii}^2. \tag{S10}$$

Putting (S7), (S9) and (S10) together, we obtain that

$$\begin{aligned} \text{Var}(Q_{\mathcal{T}}) = \underbrace{(c_1 + 4c_3 - 4c_4 - 4c_5 + 4c_6)}_{C_{\text{I}}} \sum_{i=1}^{n} Q_{ii}^2 + \underbrace{\left(c_4 - \frac{c_1}{n} - 2c_6\right)}_{C_{\text{II}}} (\text{tr}(Q))^2 \\ + \underbrace{(c_2 - 2c_3 + c_4)}_{C_{\text{III}}} \sum_{[i,j]} (Q_{ij}^2 + Q_{ij}Q_{ji}). \end{aligned} \tag{S11}$$

We simplify (S11) in the next subsection by deriving bounds for the coefficients.

C.2.2 Bounding the coefficients $C_{\mathbf{I}}, C_{\mathbf{II}}$ and $C_{\mathbf{III}}$ in (S11)

Bounding $C_{\mathbf{I}}$ We have

$$\begin{aligned} C_{\mathrm{I}} &= c_1 + 4c_3 - 4c_4 - 4c_5 + 4c_6 \\ &= c_1 + 4c_1 \frac{m-1}{n}\left(\frac{mn-2m-2n+2}{(n-1)(n-2)} + \frac{4mn-6m-6n+6}{(n-1)(n-2)(n-3)} - 1 - \frac{2}{n-2}\right). \end{aligned}$$

Through tedious calculation, we obtain that

$$\frac{mn-2m-2n+2}{(n-1)(n-2)} + \frac{4mn-6m-6n+6}{(n-1)(n-2)(n-3)} - 1 - \frac{2}{n-2} = -\frac{(n-m-1)n}{(n-2)(n-3)}.$$

Thus, $C_{\mathrm{I}} = c_1\left(1 - \frac{4(m-1)(n-m-1)}{(n-2)(n-3)}\right)$.

Bounding $C_{\mathbf{II}}$ We have

$$\begin{aligned} C_{\mathrm{II}} &= c_4 - \frac{c_1}{n} - 2c_6 = -\frac{c_1}{n} + c_1 \frac{m-1}{n(n-2)}\left(\frac{-4mn+6m+6n-6}{(n-1)(n-3)} + 4\right) \\ &= -\frac{c_1}{n} + c_1 \frac{(m-1)(4n^2-4mn+6m-10n+6)}{n(n-1)(n-2)(n-3)} \\ &= -\frac{c_1}{n}\left(1 - \frac{(m-1)(n-m-1)(4n-6)}{(n-1)(n-2)(n-3)}\right) \\ &\le c_1 \frac{(m-1)(n-m-1)(4n-6)}{n(n-1)(n-2)(n-3)} \le \frac{c_1}{n}\frac{4(m-1)(n-m-1)}{n(n-2)(n-3)}. \end{aligned}$$

Bounding $C_{\mathbf{III}}$ We consider four cases.

- If $m = 1$, then $c_2 = c_3 = c_4 = 0$ and $C_{\mathrm{III}} \le \frac{c_1}{2}$.
- If $m = 2$, then

$$\begin{aligned} C_{\mathrm{III}} &= c_1\left(\frac{n+1}{n(n-1)} - \frac{-4}{n(n-1)(n-2)} - \frac{2}{n(n-1)(n-2)}\right) \\ &= c_1\left(\frac{n+1}{n(n-1)} + \frac{2}{n(n-1)(n-2)}\right) \le \frac{c_1}{2}. \end{aligned}$$

- If $m = 3$, then

$$\begin{aligned} C_{\mathrm{III}} &= c_1\left(\frac{2(n+2)}{n(n-1)} - \frac{4(n-4)}{n(n-1)(n-2)} - \frac{12n-24}{n(n-1)(n-2)(n-3)}\right) \\ &= c_1\left(\frac{2(n+2)}{n(n-1)} - \frac{4(n-4)}{n(n-1)(n-2)} - \frac{12}{n(n-1)(n-3)}\right). \end{aligned}$$

If $n \geq 7$,

$$C_{\mathrm{III}} \leq c_1 \frac{2(n+2)}{n(n-1)} \leq \frac{c_1}{2}.$$

For $n = 4, 5, 6$, we can also verify that $C_{\mathrm{III}} \leq \frac{c_1}{2}$.
• If $m \geq 4$, then

$$4mn - 6m - 6n + 6 = (2m-6)(n-3) + 2(mn-6) \geq (2m+2)(n-3).$$

and thus

$$c_4 \leq c_1 \frac{(2m+2)(m-1)}{n(n-1)(n-2)}.$$

Then we have

$$\begin{aligned}
C_{\mathrm{III}} &\leq c_1 \frac{m-1}{n(n-1)} \left(n+m-1 - \frac{2(mn-2m-2n+2)}{n-2} - \frac{2m+2}{n-2} \right) \\
&= c_1 \frac{m-1}{n(n-1)} \left(n+m-1 - \frac{2mn-4n-2m+6}{n-2} \right) \\
&= c_1 \frac{m-1}{n(n-1)} \left(n-m+3 - \frac{2m-2}{n-2} \right) \\
&\leq c_1 \left(\frac{(m-1)(n-m+3)}{n(n-1)} - \frac{2(m-1)^2}{n(n-1)(n-2)} \right) \\
&\leq c_1 \left(\frac{(n+2)^2}{4n(n-1)} - \frac{2(m-1)^2}{n(n-1)(n-2)} \right) \\
&\leq c_1 \left(\frac{(n+2)^2}{4n(n-1)} - \frac{18}{n(n-1)(n-2)} \right). \qquad \text{(S12)}
\end{aligned}$$

If $n \geq 7$,

$$C_{\mathrm{III}} \leq c_1 \frac{(n+2)^2}{4n(n-1)} \leq \frac{81 c_1}{168} \leq \frac{c_1}{2}.$$

For $n = 4, 5, 6$, we can also verify that $C_{\mathrm{III}} \leq \frac{c_1}{2}$.

Therefore, we always have $C_{\mathrm{III}} \leq \frac{c_1}{2}$.

Using the above bounds for $(C_{\mathrm{I}}, C_{\mathrm{II}}, C_{\mathrm{III}})$ in (S11), we obtain that

$$\begin{aligned}
\mathrm{Var}\,(Q_{\mathcal{T}}) \leq\ & c_1 \left(1 - \frac{4(m-1)(n-m-1)}{(n-2)(n-3)} \right) \sum_{i=1}^{n} Q_{ii}^2 \\
& + c_1 \frac{4(m-1)(n-m-1)}{(n-2)(n-3)} \frac{(\mathrm{tr}(Q))^2}{n} + \frac{c_1}{2} \sum_{[i,j]} (Q_{ij}^2 + Q_{ij} Q_{ji}).
\end{aligned}$$

Because $(\mathrm{tr}(Q))^2 \leq n \sum_{i=1}^{n} Q_{ii}^2$ and $Q_{ij} Q_{ji} \leq (Q_{ij}^2 + Q_{ji}^2)/2$, we conclude that $\mathrm{Var}\,(Q_{\mathcal{T}}) \leq c_1 \|Q\|_F^2$.

C.3 Proofs of the Lemmas in Section 6.2

Proof of Lemma 4.6.6. Using the definitions of σ_n^2 and ρ_e, we have

$$\begin{aligned}\sigma_n^2 &= \left(\frac{1}{n_1} - \frac{1}{n}\right)\sum_{i=1}^{n} e_i^2(1) + \left(\frac{1}{n_0} - \frac{1}{n}\right)\sum_{i=1}^{n} e_i^2(0) + \frac{2}{n}\sum_{i=1}^{n} e_i(1)e_i(0) \\ &= \frac{n_0}{n_1 n}\sum_{i=1}^{n} e_i^2(1) + \frac{n_1}{n_0 n}\sum_{i=1}^{n} e_i^2(0) + \frac{2\rho_e}{n}\sqrt{\sum_{i=1}^{n} e_i^2(1)}\sqrt{\sum_{i=1}^{n} e_i^2(0)}.\end{aligned}$$

If $\rho_e \geq 0$, then

$$\sigma_n^2 \geq \frac{n_0}{n_1 n}\sum_{i=1}^{n} e_i^2(1) + \frac{n_1}{n_0 n}\sum_{i=1}^{n} e_i^2(0) \geq \min\left\{\frac{n_1}{n_0}, \frac{n_0}{n_1}\right\}\mathcal{E}_2.$$

If $\rho_e < 0$, then using the fact

$$\left(\sqrt{\frac{n_0}{n_1}}a - \sqrt{\frac{n_1}{n_0}}b\right)^2 \geq 0 \Longleftrightarrow 2ab \leq \frac{n_0}{n_1}a^2 + \frac{n_1}{n_0}b^2,$$

we obtain that

$$\sigma_n^2 \geq (1+\rho_e)\left(\frac{n_0}{n_1 n}\sum_{i=1}^{n} e_i^2(1) + \frac{n_1}{n_0 n}\sum_{i=1}^{n} e_i^2(0)\right) \geq \eta \min\left\{\frac{n_1}{n_0}, \frac{n_0}{n_1}\right\}\mathcal{E}_2.$$

Putting the pieces together, we complete the proof. □

Proof of Lemma 4.6.7. Recall that $\hat{\mu}_t$ is the intercept from the OLS fit of Y_t^{obs} on $\mathbf{1}$ and X_t. From the Frisch–Waugh Theorem, it is identical to the coefficient of the OLS fit of the residual $(\mathrm{I} - H_t)Y_t^{\text{obs}}$ on the residual $(\mathrm{I} - H_t)\mathbf{1}$, after projecting onto X_t:

$$\hat{\mu}_t = \frac{\mathbf{1}^\mathsf{T}(\mathrm{I} - H_t)^\mathsf{T}(\mathrm{I} - H_t)Y_t^{\text{obs}}}{\|(\mathrm{I} - H_t)\mathbf{1}\|_2^2} = \frac{\mathbf{1}^\mathsf{T}(\mathrm{I} - H_t)Y_t^{\text{obs}}}{\mathbf{1}^\mathsf{T}(\mathrm{I} - H_t)\mathbf{1}}.$$

Using the definition (4.6) and the fact that $(\mathrm{I} - H_t)X_t = 0$, we have

$$\begin{aligned}(\mathrm{I} - H_t)Y_t^{\text{obs}} = (\mathrm{I} - H_t)(\mu_t\mathbf{1} + X_t\beta_t + e_t(t)) = \mu_t(\mathrm{I} - H_t)\mathbf{1} + (\mathrm{I} - H_t)e_t(t), \\ \Longrightarrow \hat{\mu}_t = \mu_t + \frac{\mathbf{1}^\mathsf{T}(\mathrm{I} - H_t)e_t(t)}{\mathbf{1}^\mathsf{T}(\mathrm{I} - H_t)\mathbf{1}} = \mu_t + \frac{\mathbf{1}^\mathsf{T}e_t(t)/n_t - \mathbf{1}^\mathsf{T}H_t e_t(t)/n_t}{1 - \mathbf{1}^\mathsf{T}H_t\mathbf{1}/n_t}.\end{aligned}$$

Recalling that $\tau = \mu_1 - \mu_0$, we complete the proof. □

Proof of Lemma 4.6.8. Because $\|U\|_{\text{op}} \le \|U\|_F$, Lemma 4.6.3 implies that with probability $1-\delta$,

$$\left\|\sum_{i\in\mathcal{T}} u_i\right\|_2 \Big/ \|U\|_F \le \sqrt{\frac{m(n-m)}{n(n-1)}} + \sqrt{8\log\frac{1}{\delta}},$$

which further implies $\left\|\sum_{i\in\mathcal{T}} u_i\right\|_2 = O_{\mathbb{P}}\left(\|U\|_F\right)$. This immediately implies the three results in Lemma 4.6.8 by choosing appropriate U.

Let $u_i = e_i(t)$ with $\sum_{i=1}^n u_i = 0$, $U = (u_1, \ldots, u_n)^{\mathsf{T}} \in \mathbb{R}^{n\times 1}$, and $\|U\|_F^2 = \sum_{i=1}^n u_i^2 = \sum_{i=1}^n e_i^2(t)$. Therefore,

$$\mathbf{1}^{\mathsf{T}} e_t(t) = \left\|\sum_{i\in\mathcal{T}_t} u_i\right\|_2 = O_{\mathbb{P}}(\|U\|_F) = O_{\mathbb{P}}\left(\sqrt{\sum_{i=1}^n e_i^2(t)}\right) = O_{\mathbb{P}}\left(\sqrt{n\mathcal{E}_2}\right).$$

Let $u_i = x_i$ with $\sum_{i=1}^n u_i = 0$, $U = X$, and $\|U\|_F = \|X\|_F = \sqrt{\text{tr}(X^{\mathsf{T}}X)} = \text{tr}(n\text{I}) = np$. Therefore,

$$\|X_t^{\mathsf{T}}\mathbf{1}\|_2 = \left\|\sum_{i\in\mathcal{T}_t} u_i\right\|_2 = O_{\mathbb{P}}\left(\|U\|_F\right) = O_{\mathbb{P}}\left(\sqrt{np}\right).$$

Let $u_i = x_i e_i(t)$ with $\sum_{i=1}^n u_i = 0$ due to (4.7). Therefore,

$$\|X_t^{\mathsf{T}} e_t(t)\|_2 = \left\|\sum_{i\in\mathcal{T}_t} u_i\right\|_2 = O_{\mathbb{P}}\left(\sqrt{\sum_{i=1}^n \|x_i\|^2 e_i^2(t)}\right).$$

Recalling (4.30) that $\|x_i\|_2^2 = nH_{ii} \le n\kappa$, we have $\|X_t^{\mathsf{T}} e_t(t)\|_2 = O_{\mathbb{P}}\left(n\sqrt{\mathcal{E}_2\kappa}\right)$.

□

We need the following proposition to prove Lemma 4.6.9.

Proposition C.3.1. *A and B are two symmetric matrices. A is positive definite, and $A+B$ is invertible. Then*

$$\|(A+B)^{-1} - A^{-1}\|_{\text{op}} \le \frac{\|A^{-1}\|_{\text{op}}^2 \cdot \|B\|_{\text{op}}}{1 - \min\{1, \|A^{-1}\|_{\text{op}} \cdot \|B\|_{\text{op}}\}}.$$

Proof of Proposition C.3.1. Let $M = A^{-\frac{1}{2}} B A^{-\frac{1}{2}}$ and $\Lambda(M)$ be its spectrum. By definition, $\|M\|_{\text{op}} \le \|A^{-1}\|_{\text{op}} \cdot \|B\|_{\text{op}}$. If $\|A^{-1}\|_{\text{op}} \cdot \|B\|_{\text{op}} \ge 1$, the inequality is trivial because the right-hand side of it is ∞. Without loss of generality, we assume $\|A^{-1}\|_{\text{op}} \cdot \|B\|_{\text{op}} < 1$, which implies $\|M\|_{\text{op}} < 1$.

Proposition C.3.1 follows by combining

$$\|(A+B)^{-1} - A^{-1}\|_{\text{op}} = \|A^{-\frac{1}{2}}((\text{I}+M)^{-1} - \text{I})A^{-\frac{1}{2}}\|_{\text{op}}$$

$$\leq \|A^{-1}\|_{\mathrm{op}} \cdot \|\mathrm{I} - (\mathrm{I}+M)^{-1}\|_{\mathrm{op}}$$

and

$$\|\mathrm{I} - (\mathrm{I}+M)^{-1}\|_{\mathrm{op}} \leq \sup_{\lambda \in \Lambda(M)} \left|\frac{\lambda}{1+\lambda}\right| = \frac{\|M\|_{\mathrm{op}}}{1-\|M\|_{\mathrm{op}}} \leq \frac{\|A^{-1}\|_{\mathrm{op}} \cdot \|B\|_{\mathrm{op}}}{1-\|A^{-1}\|_{\mathrm{op}} \cdot \|B\|_{\mathrm{op}}}.$$

□

Proof of Lemma 4.6.9. Let $V_i = x_i x_i^\mathsf{T} - \mathrm{I}$, then $\sum_{i=1}^n V_i = 0$. In the following, we will repeatedly use the basic facts: $n^{-1}X^\mathsf{T}X = \mathrm{I}$, $\|x_i\|_2^2 = nH_{ii}$, and $\sum_{i=1}^n x_i x_i^\mathsf{T} = XX^\mathsf{T} = nH$. Recalling the definitions of ν, ν_+ and ν_- in Lemma 4.6.4, we have

$$\begin{aligned}
\nu^2 &= \left\|\frac{1}{n}\sum_{i=1}^n V_i^2\right\|_{\mathrm{op}} = \left\|\frac{1}{n}\sum_{i=1}^n \left(\|x_i\|_2^2 x_i x_i^\mathsf{T} - 2x_i x_i^\mathsf{T} + \mathrm{I}\right)\right\|_{\mathrm{op}} \\
&= \left\|\left(\frac{1}{n}\sum_{i=1}^n \|x_i\|_2^2 x_i x_i^\mathsf{T}\right) - \mathrm{I}\right\|_{\mathrm{op}} = \left\|\left(\sum_{i=1}^n H_{ii} x_i x_i^\mathsf{T}\right) - \mathrm{I}\right\|_{\mathrm{op}} \\
&\leq \left\|\sum_{i=1}^n H_{ii} x_i x_i^\mathsf{T}\right\|_{\mathrm{op}} + 1 \leq \kappa \left\|\sum_{i=1}^n x_i x_i^\mathsf{T}\right\|_{\mathrm{op}} + 1 = n\kappa\|H\|_{\mathrm{op}} + 1 = n\kappa + 1, \\
\nu_+ &= \max_{1\leq i\leq n} \|x_i x_i^\mathsf{T} - \mathrm{I}\|_{\mathrm{op}} \leq \max_{1\leq i\leq n} \|x_i\|_2^2 + 1 = n \max_{1\leq i\leq n} H_{ii} + 1 = n\kappa + 1, \\
\nu_-^2 &= \sup_{\omega\in\mathcal{S}^{p-1}} \frac{1}{n}\sum_{i=1}^n (\omega^\mathsf{T} V_i \omega)^2 = \sup_{\omega\in\mathcal{S}^{p-1}} \frac{1}{n}\sum_{i=1}^n ((x_i^\mathsf{T}\omega)^2 - 1)^2 \\
&= \sup_{\omega\in\mathcal{S}^{p-1}} \frac{1}{n}\sum_{i=1}^n \left[(x_i^\mathsf{T}\omega)^4 - 2(x_i^\mathsf{T}\omega)^2 + 1\right] \\
&= \sup_{\omega\in\mathcal{S}^{p-1}} \frac{1}{n}\sum_{i=1}^n (x_i^\mathsf{T}\omega)^4 - 2\omega^\mathsf{T}\left(\frac{X^\mathsf{T}X}{n}\right)\omega + 1 \\
&= \sup_{\omega\in\mathcal{S}^{p-1}} \frac{1}{n}\sum_{i=1}^n (x_i^\mathsf{T}\omega)^4 - 1 \leq \sup_{\omega\in\mathcal{S}^{p-1}} \frac{1}{n}\sum_{i=1}^n (x_i^\mathsf{T}\omega)^4 \\
&\leq \sup_{\omega\in\mathcal{S}^{p-1}} \frac{1}{n}\sum_{i=1}^n \|x_i\|_2^2 (x_i^\mathsf{T}\omega)^2 = \left\|\sum_{i=1}^n H_{ii} x_i x_i^\mathsf{T}\right\|_{\mathrm{op}} \leq n\kappa.
\end{aligned}$$

By Lemma 4.6.4,

$$\begin{aligned}
\|\Sigma_t - \mathrm{I}\|_{\mathrm{op}} = \frac{1}{n_t}\left\|\sum_{i\in\mathcal{T}_t} V_i\right\|_{\mathrm{op}} &= O_{\mathbb{P}}\left(\frac{1}{n_t}\left[n\sqrt{C(p)\kappa} + nC(p)\kappa + n\sqrt{\kappa}\right]\right) \\
&= O_{\mathbb{P}}\left(\sqrt{\kappa\log p} + \kappa\log p\right).
\end{aligned}$$

By Assumption 2, $\kappa \log p = o(1)$, and therefore the first result holds:

$$\|\Sigma_t - \mathrm{I}\|_{\mathrm{op}} = O_{\mathbb{P}}\left(\sqrt{\kappa \log p}\right) = o_{\mathbb{P}}(1). \tag{S13}$$

Thus with probability $1 - o(1)$,

$$\|\Sigma_t - \mathrm{I}\|_{\mathrm{op}} \leq \frac{1}{2} \Longrightarrow \|\Sigma_t\|_{\mathrm{op}} \geq \frac{1}{2}, \tag{S14}$$

where we use the convexity of $\|\cdot\|_{\mathrm{op}}$. Note that for any Hermitian matrix A, $\|A^{-1}\|_{\mathrm{op}} = \lambda_{\min}(A)^{-1}$ where $\lambda_{\min}$ denotes the minimum eigenvalue. Thus with probability $1 - o(1)$,

$$\left\|\Sigma_t^{-1}\right\|_{\mathrm{op}} \leq 2. \tag{S15}$$

Therefore, the second result holds: $\left\|\Sigma_t^{-1}\right\|_{\mathrm{op}} = O_{\mathbb{P}}(1)$.

To prove the third result, we apply Proposition C.3.1 with $A = \mathrm{I}$ and $B = \Sigma_t - \mathrm{I}$. By (S14) and (S15), with probability $1 - o(1)$, $A + B$ is invertible and $\|B\|_{\mathrm{op}} \leq 1/2$. Together with (S13), we have

$$\left\|\Sigma_t^{-1} - \mathrm{I}\right\|_{\mathrm{op}} = O_{\mathbb{P}}\left(\frac{\|B\|_{\mathrm{op}}}{1 - \|B\|_{\mathrm{op}}}\right) = O_{\mathbb{P}}(\|B\|_{\mathrm{op}}) = O_{\mathbb{P}}(\sqrt{\kappa \log p}).$$

□

Proof of Lemma 4.6.10. First, (4.7) implies

$$\mathbf{1}^{\mathsf{T}} Q(t) = \mathbf{1}^{\mathsf{T}} H \operatorname{diag}(e(t)) = \mathbf{1}^{\mathsf{T}} X (X^{\mathsf{T}} X)^{-1} X^{\mathsf{T}} \operatorname{diag}(e(t)) = 0,$$

$$Q(t)\mathbf{1} = H \operatorname{diag}(e(t))\mathbf{1} = He(t) = X(X^{\mathsf{T}} X)^{-1} X^{\mathsf{T}} e(t) = 0,$$

which further imply $\mathbf{1}^{\mathsf{T}} Q(t) \mathbf{1} = 0$. Second, (4.9) implies $\operatorname{tr}(Q(t)) = n\Delta_t$. Third,

$$\|Q(t)\|_F^2 = \sum_{i=1}^n \sum_{j=1}^n H_{ij}^2 e_j^2(t) = \sum_{j=1}^n e_j^2(t) \left(\sum_{i=1}^n H_{ij}^2\right).$$

Because H is idempotent, $H^{\mathsf{T}} H = H \Longrightarrow \sum_{i=1}^n H_{ij}^2 = H_{jj}$ for all j. Thus, $\|Q(t)\|_F^2 = \sum_{j=1}^n e_j^2(t) H_{jj} \leq n \mathcal{E}_2 \kappa$. □

C.4 Proof of Proposition 4.3.1

C.4.1 Preparatory lemmas

The proofs rely on the following results.

Proposition C.4.1. *[modified version of Corollary 3.1 of Yaskov (2014)] Let Z_i be i.i.d. random vectors in $\mathbb{R}^p$ with mean 0 and covariance* I. *Suppose*

$$L(\delta) \triangleq \sup_{\nu \in \mathcal{S}^{p-1}} \mathbb{E}|\nu^\mathsf{T} Z_i|^\delta < \infty$$

for some $\delta > 2$. For any constant $C > 0$, with probability $1 - e^{-Cp}$,

$$\lambda_{\min}\left(\frac{Z^\mathsf{T} Z}{n}\right) \geq 1 - 5\left(\frac{pC}{n}\right)^{\frac{\delta}{\delta+2}} L(\delta)^{\frac{2}{\delta+2}}\left(1 + \frac{1}{C}\right).$$

Proof of Proposition C.4.1. Write $y = p/n$ and $L = L(\delta)$. The proof of Corollary 3.1 of Yaskov (2014, page 6) showed that for any $a > 0$,

$$\mathbb{P}\left(\lambda_{\min}\left(\frac{Z^\mathsf{T} Z}{n}\right) < 1 - 4La^{-\delta/2} - 5ay\right) \leq \exp\left\{-La^{-1-\delta/2}n\right\}.$$

Let $a = (Cy/L)^{-2/(\delta+2)}$. Then the right-hand side is $1 - e^{-Cp}$. Thus with probability $1 - e^{-Cp}$,

$$\begin{aligned}\lambda_{\min}\left(\frac{Z^\mathsf{T} Z}{n}\right) &\geq 1 - y^{\frac{\delta}{\delta+2}} L^{\frac{2}{\delta+2}}\left(5C^{-\frac{2}{\delta+2}} + 4C^{\frac{\delta}{\delta+2}}\right) \\ &\geq 1 - 5\left(Cy\right)^{\frac{\delta}{\delta+2}} L^{\frac{2}{\delta+2}}\left(1 + \frac{1}{C}\right).\end{aligned}$$

□

Proposition C.4.2 (Theorem 1 of Tikhomirov (2017))**.** *Let Z_i be i.i.d. random vectors in $\mathbb{R}^p$ with mean 0 and covariance* I. *Suppose*

$$L(\delta) \triangleq \sup_{\nu \in \mathcal{S}^{p-1}} \mathbb{E}|\nu^\mathsf{T} Z_i|^\delta < \infty$$

for some $\delta > 2$. Then with probability at least $1 - 1/n$,

$$\nu(\delta)^{-1}\left\|\frac{Z^\mathsf{T} Z}{n} - \mathrm{I}\right\|_{\mathrm{op}} \leq \frac{\max_{1\leq i\leq n}\|Z_i\|_2^2}{n} + L(\delta)^{\frac{2}{\delta}}\left\{\left(\frac{p}{n}\right)^{\frac{\delta-2}{\delta}}\log^4\left(\frac{n}{p}\right) + \left(\frac{p}{n}\right)^{\frac{\min\{\delta-2,2\}}{\min\{\delta,4\}}}\right\},$$

for some constant $\nu(\delta)$ depending only on δ.

Proposition C.4.3 (Theorem 2 of Bahr and Esseen (1965))**.** *Let Z_i be independent mean-zero random variables. Then for any $r \in [1, 2)$,*

$$\mathbb{E}\left|\sum_{i=1}^n Z_i\right|^r \leq 2\sum_{i=1}^n \mathbb{E}|Z_i|^r.$$

C.4.2 A lemma

First we prove a more general result.

Lemma C.4.4. *Let Z_i be i.i.d. random vectors in $\mathbb{R}^p$ with mean $\mu \in \mathbb{R}^p$ and covariance matrix $\Sigma \in \mathbb{R}^{p\times p}$. Let $\tilde{Z}_i = \Sigma^{-1/2}(Z_i - \mu)$, and assume*

$$\sup_{\nu\in\mathcal{S}^{p-1}} \mathbb{E}|\nu^\mathsf{T}\tilde{Z}_i|^\delta = O(1), \quad \text{and} \quad \max_{1\le i\le n} \left|\|\tilde{Z}_i\|_2^2 - \mathbb{E}\|\tilde{Z}_i\|_2^2\right| = O_\mathbb{P}(\omega(n,p)),$$

for some $\delta > 2$ and some function $\omega(n,p)$ increasing in n and p. Let $Z = (Z_1^\mathsf{T},\dots,Z_n^\mathsf{T})^\mathsf{T}$ and $X = \mathrm{V}Z$ so that X has centered columns. If $p = O(n^\gamma)$ for some $\gamma < 1$, then over the randomness of Z,

$$\kappa = \frac{p}{n} + O_\mathbb{P}\left(\frac{\omega(n,p)}{n} + \left(\frac{p}{n}\right)^{\frac{2\delta-2}{\delta}} \log^4\left(\frac{n}{p}\right) + \left(\frac{p}{n}\right)^{\frac{\min\{2\delta-2,6\}}{\min\{\delta,4\}}}\right).$$

Proof of Lemma C.4.4. Let $\tilde{Z} = (\tilde{Z}_1^\mathsf{T},\dots,\tilde{Z}_n)^\mathsf{T}$ and $\tilde{X} = \mathrm{V}\tilde{Z}$. Then $\tilde{X} = \mathrm{V}\left(Z - \mathbf{1}\mu^\mathsf{T}\right)\Sigma^{-\frac{1}{2}} = \mathrm{V}Z\Sigma^{-\frac{1}{2}}$, and thus

$$\tilde{X}(\tilde{X}^\mathsf{T}\tilde{X})^{-1}\tilde{X}^\mathsf{T} = \mathrm{V}Z\left(Z^\mathsf{T}\mathrm{V}Z\right)^{-1}Z^\mathsf{T}\mathrm{V} = X(X^\mathsf{T}X)^{-1}X^\mathsf{T}.$$

Therefore, we can assume $\mu = 0$ and $\Sigma = \mathrm{I}$ without loss of generality, in which case $Z_i = \tilde{Z}_i$ has mean 0 and covariance matrix I.

By definition, $H_{ii} = x_i^\mathsf{T}(X^\mathsf{T}X)^{-1}x_i$, and therefore

$$\begin{aligned} H_{ii} &= \frac{1}{n}x_i^\mathsf{T}\left(\left(\frac{X^\mathsf{T}X}{n}\right)^{-1} - \mathrm{I}\right)x_i + \frac{\|x_i\|_2^2}{n} \\ &\le \frac{\|x_i\|_2^2}{n}\left(1 + \left\|\left(\frac{X^\mathsf{T}X}{n}\right)^{-1} - \mathrm{I}\right\|_{\mathrm{op}}\right). \end{aligned} \tag{S16}$$

To bound κ, we need to bound two key terms below.

Bounding $\left\|\left(n^{-1}X^\mathsf{T}X\right)^{-1} - \mathbf{I}\right\|_{\mathrm{op}}$ Let $\bar{Z} = n^{-1}\sum_{i=1}^n Z_i$. Note that

$$\mathbb{E}\|\bar{Z}\|_2^2 = \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}\|Z_i\|_2^2 = \frac{1}{n}\mathbb{E}\|Z_1\|_2^2 = \frac{p}{n}.$$

By Markov's inequality,

$$\|\bar{Z}\|_2^2 = O_\mathbb{P}\left(\frac{p}{n}\right), \tag{S17}$$

and more precisely,

$$\mathbb{P}\left(\|\bar{Z}\|_2^2 \le \sqrt{\frac{p}{n}}\right) = 1 - \mathbb{P}\left(\|\bar{Z}\|_2^2 > \sqrt{\frac{p}{n}}\right) \ge 1 - \sqrt{\frac{p}{n}}. \tag{S18}$$

Let $\mathcal{A}_1$ denote the above event that $\|\bar{Z}\|_2^2 \le \sqrt{p/n}$. Then $\mathbb{P}(\mathcal{A}_1) \ge 1 - \sqrt{p/n}$.

By Proposition C.4.2,

$$\left\|\frac{Z^\mathsf{T} Z}{n} - \mathrm{I}\right\|_{\mathrm{op}} = O_{\mathbb{P}}\left(\frac{\max_{1\le i\le n}\|Z_i\|_2^2}{n} + \left(\frac{p}{n}\right)^{\frac{\delta-2}{\delta}} \log^4\left(\frac{n}{p}\right) + \left(\frac{p}{n}\right)^{\frac{\min\{\delta-2,2\}}{\min\{\delta,4\}}}\right).$$

By the condition of Lemma C.4.4,

$$\frac{\max_{1\le i\le n}\|Z_i\|_2^2}{n} = \frac{p}{n} + \max_{1\le i\le n}\frac{\|Z_i\|_2^2 - \mathbb{E}\|Z_i\|_2^2}{n} = \frac{p}{n} + O_{\mathbb{P}}\left(\frac{\omega(n,p)}{n}\right). \tag{S19}$$

Combining the above three equations, we have

$$\begin{aligned}
\left\|\frac{X^\mathsf{T} X}{n} - \mathrm{I}\right\|_{\mathrm{op}} &= \left\|\frac{Z^\mathsf{T} Z}{n} - \mathrm{I} - \bar{Z}\bar{Z}^\mathsf{T}\right\|_{\mathrm{op}} \le \left\|\frac{Z^\mathsf{T} Z}{n} - \mathrm{I}\right\|_{\mathrm{op}} + \|\bar{Z}\|_2^2 \\
&= O_{\mathbb{P}}\left(\frac{p}{n} + \frac{\omega(n,p)}{n} + \left(\frac{p}{n}\right)^{\frac{\delta-2}{\delta}} \log^4\left(\frac{n}{p}\right) + \left(\frac{p}{n}\right)^{\frac{\min\{\delta-2,2\}}{\min\{\delta,4\}}}\right) \\
&= O_{\mathbb{P}}\left(\frac{\omega(n,p)}{n} + \left(\frac{p}{n}\right)^{\frac{\delta-2}{\delta}} \log^4\left(\frac{n}{p}\right) + \left(\frac{p}{n}\right)^{\frac{\min\{\delta-2,2\}}{\min\{\delta,4\}}}\right),
\end{aligned} \tag{S20}$$

where the last line uses the fact that the third term dominates the first term due to $p/n \to 0$. On the other hand, by Proposition C.4.1 with $C = \sqrt{n/p}$, with probability $1 - e^{-\sqrt{np}}$,

$$\begin{aligned}
\lambda_{\min}\left(\frac{Z^\mathsf{T} Z}{n}\right) &\ge 1 - 5\left(\sqrt{\frac{p}{n}}\right)^{\frac{\delta}{\delta+2}} L(\delta)^{\frac{2}{\delta+2}}\left(1 + \sqrt{\frac{p}{n}}\right) \\
&\ge 1 - 10\left(\frac{p}{n}\right)^{\frac{\delta}{2(\delta+2)}} L(\delta)^{\frac{2}{\delta+2}}.
\end{aligned} \tag{S21}$$

Let $\mathcal{A}_2$ denote the event in (S21). Then $\mathbb{P}(\mathcal{A}_2) \ge 1 - e^{-\sqrt{np}}$.

Note that for any Hermitian matrices A and B, the convexity of $\|\cdot\|_{\mathrm{op}}$ implies that

$$|\lambda_{\min}(A) - \lambda_{\min}(B)| = |\lambda_{\max}(-A) - \lambda_{\max}(-B)| \le \|-A-(-B)\|_{\mathrm{op}} = \|A-B\|_{\mathrm{op}}.$$

We have

$$\lambda_{\min}\left(\frac{X^\mathsf{T} X}{n}\right) \ge \lambda_{\min}\left(\frac{Z^\mathsf{T} Z}{n}\right) - \|\bar{Z}\bar{Z}^\mathsf{T}\|_{\mathrm{op}} = \lambda_{\min}\left(\frac{Z^\mathsf{T} Z}{n}\right) - \|\bar{Z}\|_2^2.$$

Let $\mathcal{A} = \mathcal{A}_1 \cup \mathcal{A}_2$. Then on $\mathcal{A}$,

$$\lambda_{\min}\left(\frac{X^\mathsf{T}X}{n}\right) \geq 1 - 10\left(\frac{p}{n}\right)^{\frac{\delta}{2(\delta+2)}} L(\delta)^{\frac{2}{\delta+2}} - \sqrt{\frac{p}{n}}.$$

Since $p/n \to 0$, for sufficiently large n,

$$\lambda_{\min}\left(\frac{X^\mathsf{T}X}{n}\right) \geq \frac{1}{2}$$

with probability

$$\mathbb{P}(\mathcal{A}) \geq \mathbb{P}(\mathcal{A}_1) + \mathbb{P}(\mathcal{A}_2) - 1 \geq 1 - e^{-\sqrt{np}L(\delta)} - \sqrt{\frac{p}{n}} = 1 - o(1).$$

Finally, using Proposition C.3.1 with $A = \mathrm{I}$ and $B = n^{-1}X^\mathsf{T}X - \mathrm{I}$, by Slusky's lemma, we have that

$$\left\|\left(\frac{X^\mathsf{T}X}{n}\right)^{-1} - \mathrm{I}\right\|_{\mathrm{op}} = O_{\mathbb{P}}\left(\frac{\omega(n,p)}{n} + \left(\frac{p}{n}\right)^{\frac{\delta-2}{\delta}} \log^4\left(\frac{n}{p}\right) + \left(\frac{p}{n}\right)^{\frac{\min\{\delta-2,2\}}{\min\{\delta,4\}}}\right). \tag{S22}$$

Because $p = O(n^\gamma)$ for some $\gamma < 1$,

$$\left\|\left(\frac{X^\mathsf{T}X}{n}\right)^{-1} - \mathrm{I}\right\|_{\mathrm{op}} = O_{\mathbb{P}}\left(\frac{\omega(n,p)}{n}\right) + o_{\mathbb{P}}(1). \tag{S23}$$

Bounding $\max_{1\leq i\leq n}\|x_i\|_2^2$ Because $x_i = Z_i - \bar{Z}$, the Cauchy–Schwarz inequality implies

$$\|x_i\|_2^2 = \|Z_i\|_2^2 - 2Z_i^\mathsf{T}\bar{Z} + \|\bar{Z}\|_2^2 \leq \|Z_i\|_2^2 + 2\|Z_i\|_2\|\bar{Z}\|_2 + \|\bar{Z}\|_2^2.$$

By (S19) and (S17),

$$\begin{aligned}
\frac{\max_{1\leq i\leq n}\|x_i\|_2^2}{n} &= \frac{\mathbb{E}\|Z_i\|_2^2}{n} + \frac{\max_i \|Z_i\|_2^2 - \mathbb{E}\|Z_i\|_2^2 + 2\|Z_i\|_2\|\bar{Z}\|_2 + \|\bar{Z}\|_2^2}{n} \\
&= \frac{p}{n} + O_{\mathbb{P}}\left(\frac{\omega(n,p)}{n} + \sqrt{\frac{(p+\omega(n,p))p}{n^3}} + \frac{p}{n^2}\right) \\
&= \frac{p}{n} + O_{\mathbb{P}}\left(\frac{\omega(n,p)}{n} + \sqrt{\frac{\omega(n,p)p}{n^3}} + \frac{p}{n^{3/2}}\right).
\end{aligned}$$

Because $\omega(n,p)$ is increasing and $p/n \to 0$, we have

$$\sqrt{\frac{\omega(n,p)p}{n^3}} = O\left(\frac{\omega(n,p)}{n}\left(\frac{p}{n}\right)^{1/2}\right) = o\left(\frac{\omega(n,p)}{n}\right).$$

Thus, we obtain that

$$\frac{\max_{1\le i\le n}\|x_i\|_2^2}{n} = \frac{p}{n} + O_{\mathbb{P}}\left(\frac{\omega(n,p)}{n} + \frac{p}{n^{3/2}}\right). \tag{S24}$$

Putting (S16), (S23) and (S24) together and using some tedious cancellations, we have

$$\begin{aligned}\kappa = \frac{p}{n} &+ O_{\mathbb{P}}\left(\frac{\omega(n,p)}{n} + \frac{p}{n^{3/2}}\right) \\ &+ O_{\mathbb{P}}\left(\frac{\omega^2(n,p)}{n^2} + \left(\frac{p}{n}\right)^{1+\frac{\delta-2}{\delta}}\log^4\left(\frac{n}{p}\right) + \left(\frac{p}{n}\right)^{1+\frac{\min\{\delta-2,2\}}{\min\{\delta,4\}}}\right).\end{aligned} \tag{S25}$$

Because

$$\left(\frac{p}{n}\right)^{1+\frac{\min\{\delta-2,2\}}{\min\{\delta,4\}}} \ge \left(\frac{p}{n}\right)^{3/2} \ge \frac{p}{n^{3/2}},$$

(S25) further simplifies to

$$\begin{aligned}\kappa &= \frac{p}{n} + O_{\mathbb{P}}\left(\frac{\omega(n,p)}{n} + \frac{\omega^2(n,p)}{n^2} + \left(\frac{p}{n}\right)^{\frac{2\delta-2}{\delta}}\log^4\left(\frac{n}{p}\right) + \left(\frac{p}{n}\right)^{\frac{\min\{2\delta-2,6\}}{\min\{\delta,4\}}}\right) \\ &= \frac{p}{n} + O_{\mathbb{P}}\left(\frac{\omega(n,p)}{n}\max\left\{\frac{\omega(n,p)}{n},1\right\} + \left(\frac{p}{n}\right)^{\frac{2\delta-2}{\delta}}\log^4\left(\frac{n}{p}\right) + \left(\frac{p}{n}\right)^{\frac{\min\{2\delta-2,6\}}{\min\{\delta,4\}}}\right).\end{aligned}$$

We complete the proof using $\kappa \le 1$. □

C.4.3 Use Lemma C.4.4 to prove Proposition 4.3.1

We have argued in the proof of Proposition C.4.4 that we can assume $\mu = 0$ without loss of generality. Because the hat matrix is invariant to rescaling, we further assume $\mathbb{E}Z_{ij}^2 = 1$ without loss of generality. Based on Proposition C.4.4, it suffices to verify

$$\sup_{\nu\in\mathcal{S}^{p-1}} \mathbb{E}|\nu^{\mathsf{T}} Z_i|^\delta = O(1), \tag{S26}$$

$$\max_{1\le i\le n}\left|\|Z_i\|_2^2 - \mathbb{E}\|Z_i\|_2^2\right| = O_{\mathbb{P}}\left(n^{\frac{2}{\delta}} p^{\frac{2}{\min\{\delta,4\}}}\right). \tag{S27}$$

If (S26) and (S27) hold, by Proposition C.4.4, we have that

$$\kappa = \frac{p}{n} + O_{\mathbb{P}}\left(\frac{p^{2/\min\{\delta,4\}}}{n^{(\delta-2)/\delta}} + \left(\frac{p}{n}\right)^{\frac{2\delta-2}{\delta}}\log^4\left(\frac{n}{p}\right) + \left(\frac{p}{n}\right)^{\frac{\min\{2\delta-2,6\}}{\min\{\delta,4\}}}\right).$$

Then we can prove Proposition 4.3.1 for two cases.

Case 1 If $\delta > 4$, then $\frac{2\delta-2}{\delta} < \frac{3}{2} = \frac{\min\{2\delta-2,6\}}{\min\{\delta,4\}}$. Thus the third term dominates the second term in the above $O_{\mathbb{P}}(\cdot)$, implying

$$\kappa = \frac{p}{n} + O_{\mathbb{P}}\left(\frac{p^{1/2}}{n^{(\delta-2)/\delta}} + \left(\frac{p}{n}\right)^{\frac{3}{2}}\right).$$

Case 2 If $\delta \leq 4$, then

$$\kappa = \frac{p}{n} + O_{\mathbb{P}}\left(\frac{p^{2/\delta}}{n^{(\delta-2)/\delta}} + \left(\frac{p}{n}\right)^{\frac{2\delta-2}{\delta}} \log^4\left(\frac{n}{p}\right)\right).$$

Because

$$\left(\frac{p}{n}\right)^{\frac{2\delta-2}{\delta}} = \frac{p^{2/\delta}}{n^{(\delta-2)/\delta}} \frac{p^{(2\delta-4)/\delta}}{n} \leq \frac{p^{2/\delta}}{n^{(\delta-2)/\delta}} \frac{p}{n},$$

the first term dominates in the above $O_{\mathbb{P}}(\cdot)$, implying

$$\kappa = \frac{p}{n} + O_{\mathbb{P}}\left(\frac{p^{2/\delta}}{n^{(\delta-2)/\delta}}\right) = \frac{p}{n} + O_{\mathbb{P}}\left(\frac{p^{2/\delta}}{n^{(\delta-2)/\delta}} + \left(\frac{p}{n}\right)^{3/2}\right).$$

The last identity holds because $p^{3/2}/n^{3/2}$ is of smaller order and thus we can add it back.

We will prove (S26) and (S27) below.

Proving (S26)

By Rosenthal (1970)'s inequality,

$$\mathbb{E}|\nu^{\mathsf{T}} Z_i|^{\delta} = \mathbb{E}\left|\sum_{j=1}^{p} \nu_j Z_{ij}\right|^{\delta} \leq C\left(\sum_{j=1}^{p} |\nu_j|^{\delta} \mathbb{E}|Z_{ij}|^{\delta} + \left(\sum_{j=1}^{p} \nu_j^2 \mathbb{E} Z_{ij}^2\right)^{\delta/2}\right)$$

where C is a constant depending only on δ. Because $\|\nu\|_2 = 1$, we have $\max_{1\leq j\leq p} |\nu_j| \leq 1$ and thus

$$\sum_{j=1}^{p} |\nu_j|^{\delta} \mathbb{E}|Z_{ij}|^{\delta} \leq M \sum_{j=1}^{p} |\nu_j|^{\delta} \leq M \sum_{j=1}^{p} |\nu_j|^{\delta} = M.$$

Hölder's inequality implies $\mathbb{E} Z_{ij}^2 \leq \left(\mathbb{E}|Z_{ij}|^{\delta}\right)^{2/\delta} \leq M^{2/\delta}$, which further implies

$$\left(\sum_{j=1}^{p} \nu_j^2 \mathbb{E} Z_{ij}^2\right)^{\delta/2} \leq \left(M^{2/\delta}\right)^{\delta/2} = M.$$

Because the above two bounds hold regardless of ν, we conclude that

$$\sup_{\nu \in \mathcal{S}^{p-1}} \mathbb{E}|\nu^{\mathsf{T}} Z_i|^{\delta} \leq 2CM = O(1).$$

Proving (S27)

Let $W_{ij} = Z_{ij}^2 - \mathbb{E}Z_{ij}^2$. Using Jensen's inequality that $\mathbb{E}|(X+Y)/2|^r \leq (\mathbb{E}|X|^r + \mathbb{E}|Y|^r)/2$ for any random variables X, Y and $r > 1$, we obtain that

$$\mathbb{E}|W_{ij}|^{\delta/2} \leq 2^{\delta/2-1}\left(\mathbb{E}|Z_{ij}|^{\delta} + (\mathbb{E}Z_{ij}^2)^{\delta/2}\right) \leq 2^{\delta/2}\mathbb{E}|Z_{ij}|^{\delta} \leq 2^{\delta/2}M \triangleq \tilde{M}.$$

We consider two cases.

Case 1: $\delta \geq 4$ By Hölder's inequality, $\mathbb{E}W_{ij}^2 \leq \tilde{M}^{4/\delta}$. By Rosenthal (1970)'s inequality,

$$\begin{aligned}\mathbb{E}|\|Z_i\|_2^2 - \mathbb{E}\|Z_i\|_2^2|^{\delta/2} = \mathbb{E}\left|\sum_{j=1}^p W_{ij}\right|^{\delta/2} &\leq C\left(\sum_{j=1}^p \mathbb{E}|W_{ij}|^{\delta/2} + \left(\sum_{j=1}^p \mathbb{E}W_{ij}^2\right)^{\delta/4}\right) \\ &\leq C\left(p\tilde{M} + p^{\delta/4}\tilde{M}\right) \leq C\tilde{M}p^{\delta/4},\end{aligned}$$

which implies $\mathbb{E}\left|\|Z_i\|_2^2 - \mathbb{E}\|Z_i\|_2^2\right|^{\delta/2} = O\left(p^{\delta/4}\right)$. As a result,

$$\mathbb{E}\left\{\max_{1\leq i\leq n}\left|\|Z_i\|_2^2 - \mathbb{E}\|Z_i\|_2^2\right|^{\delta/2}\right\} \leq \sum_{i=1}^n \mathbb{E}\left|\|Z_i\|_2^2 - \mathbb{E}\|Z_i\|_2^2\right|^{\delta/2} = O\left(np^{\delta/4}\right).$$

By Markov's inequality, $\max_{1\leq i\leq n}\left|\|Z_i\|_2^2 - \mathbb{E}\|Z_i\|_2^2\right| = O_{\mathbb{P}}\left(n^{2/\delta}p^{1/2}\right)$.

Case 2: $\delta < 4$ By Proposition C.4.3, with $\delta/2 \in (1,2)$,

$$\mathbb{E}|\|Z_i\|_2^2 - \mathbb{E}\|Z_i\|_2^2|^{\delta/2} = \mathbb{E}\left|\sum_{j=1}^p W_{ij}\right|^{\delta/2} \leq 2\sum_{j=1}^p \mathbb{E}|W_{ij}|^{\delta/2} \leq 2p\tilde{M}.$$

Similar to Case 1, $\max_{1\leq i\leq n}\left|\|Z_i\|_2^2 - \mathbb{E}\|Z_i\|_2^2\right| = O_{\mathbb{P}}\left(n^{2/\delta}p^{2/\delta}\right)$.

C.5 Proof of Proposition 4.3.2

Let $\bar{Y}(t) = n^{-1}\sum_{i=1}^n Y_i(t)$. Note that $H\mathbf{1} = X(X^\mathsf{T}X)^{-1}X^\mathsf{T}\mathbf{1} = 0$. By definition, $e(t) = (\mathrm{I} - H)\{Y(t) - \bar{Y}(t)\mathbf{1}\} = (\mathrm{I} - H)\{Y(t) - \mathbb{E}Y_i(t)\mathbf{1}\}$. Throughout the rest of the proof, we assume that $\mathbb{E}Y_i(t) = 0$ without loss of generality, and define $M(\delta) \triangleq \max_{t=0,1}\mathbb{E}|Y_i(t)|^\delta$.

C.5.1 Bounding $\mathcal{E}_2$

Let $Z_i = Y_i(t)^2$. Then the moment condition reads $\mathbb{E}|Z_i|^{\delta/2} < \infty$. The Kolmogorov–Marcinkiewicz–Zygmund strong law of large number (Kallenberg 2006, Theorem 4.23) implies

$$\frac{1}{n}\sum_{i=1}^n Z_i \xrightarrow{\text{a.s.}} \mathbb{E}Z_1 = O_{\mathbb{P}}(1), \quad \text{if } \delta \geq 2,$$

$$\frac{1}{n^{2/\delta}}\sum_{i=1}^{n} Z_i = o(1) \Longrightarrow \frac{1}{n}\sum_{i=1}^{n} Z_i = o_{\mathbb{P}}(n^{2/\delta-1}), \quad \text{if } \delta < 2.$$

On the other hand,

$$\frac{1}{n}\|e(t)\|_2^2 = \frac{1}{n}Y(t)^{\mathsf{T}}(\mathrm{I}-H)Y(t) \le \frac{1}{n}\|Y(t)\|_2^2 = \frac{1}{n}\sum_{i=1}^{n} Z_i,$$

which further implies the bound for $\mathcal{E}_2$.

C.5.2 Bounding $\mathcal{E}_2^{-1}$

Without loss of generality, we assume that $Y_i(1)$ is not a constant with probability 1. First we show that

$$\frac{Y(1)^{\mathsf{T}}HY(1)}{Y(1)^{\mathsf{T}}Y(1)} = o_{\mathbb{P}}(1).$$

For any permutation π on $\{1,\dots,n\}$, let $H(\pi)$ denote the matrix with

$$H(\pi)_{ij} = H_{\pi(i),\pi(j)}.$$

Because the $Y_i(1)$'s are i.i.d., for any π,

$$(Y_1(1),\dots,Y_n(1)) \overset{\mathrm{d}}{=} (Y_{\pi^{-1}(1)}(1),\dots,Y_{\pi^{-1}(n)}(1)),$$

and thus

$$\begin{aligned}\frac{Y(1)^{\mathsf{T}}H(\pi)Y(1)}{Y(1)^{\mathsf{T}}Y(1)} &= \frac{\sum_{i=1}^n\sum_{j=1}^n H_{\pi(i),\pi(j)}Y_i(1)Y_j(1)}{\sum_{i=1}^n Y_i(1)^2} \\ =\frac{\sum_{i=1}^n\sum_{j=1}^n H_{i,j}Y_{\pi^{-1}(i)}(1)Y_{\pi^{-1}(j)}(1)}{\sum_{i=1}^n Y_{\pi^{-1}(i)}(1)^2} &\overset{\mathrm{d}}{=} \frac{Y(1)^{\mathsf{T}}HY(1)}{Y(1)^{\mathsf{T}}Y(1)}.\end{aligned}$$

Furthermore,

$$\frac{Y(1)^{\mathsf{T}}HY(1)}{Y(1)^{\mathsf{T}}Y(1)} \le 1$$

and thus it has finite expectation. This implies that

$$\mathbb{E}\frac{Y(1)^{\mathsf{T}}HY(1)}{Y(1)^{\mathsf{T}}Y(1)} = \frac{1}{n!}\sum_{\pi}\frac{Y(1)^{\mathsf{T}}H(\pi)Y(1)}{Y(1)^{\mathsf{T}}Y(1)} = \frac{1}{n!}\frac{Y(1)^{\mathsf{T}}H^*Y(1)}{Y(1)^{\mathsf{T}}Y(1)},$$

where $H^* = \sum_\pi H(\pi)/n!$ with the summation over all possible permutations. We can show that

$$H^*_{ii} = \frac{1}{n}\sum_{i=1}^n H_{ii} = \frac{p}{n}, \quad H^*_{ij} = \frac{1}{n(n-1)}\sum_{i\neq j} H_{ij} = -\frac{1}{n(n-1)}\sum_{i=1}^n H_{ii} = -\frac{p}{n(n-1)},$$

where the last equality uses the fact that $\sum_{i=1}^n \sum_{j=1}^n H_{ij} = 0$. Therefore,

$$\begin{aligned}
&\mathbb{E}\frac{Y(1)^\mathsf{T} H Y(1)}{Y(1)^\mathsf{T} Y(1)} = \mathbb{E}\frac{Y(1)^\mathsf{T} H^* Y(1)}{Y(1)^\mathsf{T} Y(1)} \\
=&\mathbb{E}\frac{\frac{p}{n} Y(1)^\mathsf{T} Y(1) - \frac{p}{n(n-1)} \sum_{i \neq j} Y_i(1) Y_j(1)}{Y(1)^\mathsf{T} Y(1)} \\
=&\frac{p}{n-1} - \frac{p}{n(n-1)} \mathbb{E}\frac{(\sum_{i=1}^n Y_i(1))^2}{Y(1)^\mathsf{T} Y(1)} \leq \frac{p}{n-1}.
\end{aligned}$$

By Markov's inequality, with probability $1 - \frac{2p}{n-1} = 1 - o(1)$,

$$\frac{Y(1)^\mathsf{T} H Y(1)}{Y(1)^\mathsf{T} Y(1)} \leq \frac{1}{2}.$$

Let $\mathcal{A}$ denote this event. Then

$$\mathbb{P}(\mathcal{A}^c) = o(1),$$

and on $\mathcal{A}$,

$$\frac{1}{n}\|e(1)\|_2^2 = \frac{1}{n} Y(1)^\mathsf{T}(I - H) Y(1) \geq \frac{1}{2n}\|Y(1)\|_2^2.$$

On the other hand, fix $k > 0$, and let $\tilde{Z}_i = Y_i(1) I(|Y_i(1)| \leq k)$. For sufficiently large k, $\mathbb{E}\tilde{Z}_i > 0$. By the law of large numbers, $n^{-1} \sum_{i=1}^n \tilde{Z}_i = \mathbb{E}\tilde{Z}_i \times (1 + o_\mathbb{P}(1))$. Thus on $\mathcal{A}$,

$$\mathcal{E}_2 \geq \frac{1}{2n} \sum_{i=1}^n Y_i(1)^2 \geq \frac{1}{2n} \sum_{i=1}^n \tilde{Z}_i = \mathbb{E}\tilde{Z}_i \times (1 + o_\mathbb{P}(1))$$

Since $\mathbb{P}(\mathcal{A}^c) = o(1)$, we conclude that $\mathcal{E}_2^{-1} = O_\mathbb{P}(1)$.

C.5.3 Bounding $\mathcal{E}_\infty$

We apply the triangle inequality to obtain

$$\|e(t)\|_\infty \leq \|Y(t)\|_\infty + \|HY(t)\|_\infty.$$

We bound the first term using a standard technique and Markov's inequality:

$$\mathbb{E}\|Y(t)\|_\infty^\delta \leq \sum_{i=1}^n \mathbb{E}|Y_i(t)|^\delta = nM(\delta) \Longrightarrow \|Y(t)\|_\infty = O_\mathbb{P}\left(n^{1/\delta}\right). \tag{S28}$$

Next we bound the second term $\|HY(t)\|_\infty$. Define $\tilde{Y}(t) = HY(t)$ with

$$\tilde{Y}_i(t) = \sum_{j=1}^n H_{ij} Y_j(t), \quad (i = 1, \ldots, n).$$

Fix $\epsilon > 0$ and define

$$D = \left(\frac{M(\delta)}{\epsilon}\right)^{1/\delta}.$$

We decompose $\tilde{Y}_i(t)$ into two parts:

$$\begin{aligned}\tilde{Y}_i(t) &= \sum_{j=1}^{n} H_{ij} Y_j(t) I(|Y_j(t)| \le D n^{1/\delta}) + \sum_{j=1}^{n} H_{ij} Y_j(t) I(|Y_j(t)| > D n^{1/\delta}) \\ &\triangleq R_{1,i}(t) + R_{2,i}(t).\end{aligned}$$

The second term $R_{2,i}(t)$ satisfies

$$\begin{aligned}\mathbb{P}\left(\exists i, R_{2,i}(t) \neq 0\right) &\le \mathbb{P}\left(\exists j,\ |Y_j(t)| > D n^{1/\delta}\right) \le \sum_{j=1}^{n} \mathbb{P}\left(|Y_j(t)| > D n^{1/\delta}\right) \\ &\le \sum_{j=1}^{n} \frac{1}{D^\delta n} \mathbb{E}|Y_j(t)|^\delta \le \frac{M(\delta)}{D^\delta} = \epsilon. \end{aligned} \tag{S29}$$

To deal with the first term $R_{1,i}(t)$, we define

$$w_j(t) = Y_j(t) I(|Y_j(t)| \le D n^{1/\delta}) - \mathbb{E}\left\{Y_j(t) I(|Y_j(t)| \le D n^{1/\delta})\right\},$$

with $\mathbb{E} w_j(t) = 0$. Because

$$\mathbf{1}^\mathsf{T} H = 0 \Longrightarrow \sum_{j=1}^{n} H_{ij} = 0 \Longrightarrow \sum_{j=1}^{n} H_{ij} \mathbb{E}\left\{Y_j(t) I(|Y_j(t)| \le D n^{1/\delta})\right\} = 0.$$

we can rewrite $R_{1,i}(t)$ as

$$R_{1,i}(t) = \sum_{j=1}^{n} H_{ij} w_j(t).$$

The rest of the proof proceeds based on two cases.

Case 1: $\delta < 2$ First, the $w_j(t)$'s are i.i.d. with second moment bounded by

$$\begin{aligned}\mathbb{E} w_j(t)^2 &\le \mathbb{E}\left\{Y_j^2(t) I(|Y_j(t)| \le D n^{1/\delta})\right\} \\ &\le (D n^{1/\delta})^{2-\delta} \mathbb{E}|Y_j(t)|^\delta \\ &\le n^{(2-\delta)/\delta} D^{2-\delta} M(\delta) = n^{(2-\delta)/\delta} \epsilon^{-(2-\delta)/\delta} M(\delta)^{2/\delta}.\end{aligned}$$

Second, using the fact that $\sum_{j=1}^{n} H_{ij}^2 = H_{ii}$, we obtain

$$\mathbb{E} R_{1,i}(t)^2 = \sum_{j=1}^{n} H_{ij}^2 \mathbb{E} w_j(t)^2 = \mathbb{E} w_1(t)^2 \left(\sum_{j=1}^{n} H_{ij}^2\right) = H_{ii} \mathbb{E} w_1(t)^2.$$

Let $R_1(t)$ denote the vector $(R_{1,i}(t))_{i=1}^n$. Then

$$\mathbb{E}\|R_1(t)\|_\infty^2 \le \sum_{i=1}^n \mathbb{E}R_{1,i}(t)^2 = \left(\sum_{i=1}^n H_{ii}\right)\mathbb{E}w_1(t)^2 \le pn^{(2-\delta)/\delta}\epsilon^{-(2-\delta)/\delta}M(\delta)^{2/\delta}.$$

By Markov's inequality, with probability $1-\epsilon$,

$$\|R_1(t)\|_\infty \le \left(\frac{\mathbb{E}\|R_1(t)\|_\infty^2}{\epsilon}\right)^{1/2} = p^{1/2}n^{(2-\delta)/2\delta}\epsilon^{-(4-\delta)/2\delta}M(\delta)^{1/\delta}. \tag{S30}$$

Combining (S29) and (S30), we obtain that with probability $1-2\epsilon$,

$$\|HY(t)\|_\infty \le p^{1/2}n^{(2-\delta)/2\delta}\epsilon^{-(4-\delta)/2\delta}M(\delta)^{1/\delta}.$$

Because this holds for arbitrary ϵ, we conclude that if $\delta < 2$,

$$\|HY(t)\|_\infty = O_{\mathbb{P}}\left(p^{1/2}n^{1/\delta-1/2}\right) = o_{\mathbb{P}}(n^{1/\delta}).$$

Case 2: $\delta \ge 2$ Using the convexity of the mapping $|\cdot|^\delta$, we have

$$\mathbb{E}\left|\frac{w_j(t)}{2}\right|^\delta \le \frac{\mathbb{E}\left\{|Y_j(t)|^\delta I(|Y_j(t)| \le Dn^{1/\delta})\right\} + |\mathbb{E}\left\{Y_j(t)I(|Y_j(t)| \le Dn^{1/\delta})\right\}|^\delta}{2}.$$

Applying Jensen's inequality on the second term, we have

$$\mathbb{E}|w_j(t)|^\delta \le 2^\delta\mathbb{E}\left\{|Y_j(t)|^\delta I(|Y_j(t)| \le Dn^{1/\delta})\right\} \le 2^\delta\mathbb{E}|Y_j(t)|^\delta \le 2^\delta M(\delta).$$

By Rosenthal (1970)'s inequality, there exists a constant C depending only on δ, such that

$$\begin{aligned}
\mathbb{E}|R_{1,i}(t)|^\delta &\le C\left(\sum_{j=1}^n \mathbb{E}|H_{ij}w_j(t)|^\delta + \left(\sum_{j=1}^n \mathbb{E}|H_{ij}w_j(t)|^2\right)^{\delta/2}\right) \\
&\le C\left(2^\delta M(\delta)\sum_{j=1}^n |H_{ij}|^\delta + \left(2^2M(2)\sum_{j=1}^n H_{ij}^2\right)^{\delta/2}\right) \\
&\le C2^\delta\left(M(\delta)H_{ii}^{\delta/2-1}\sum_{j=1}^n H_{ij}^2 + M(2)^{\delta/2}H_{ii}^{\delta/2}\right) \\
&= C2^\delta(M(\delta)+M(2)^{\delta/2})H_{ii}^{\delta/2} \le C2^\delta(M(\delta)+M(2)^{\delta/2})H_{ii}.
\end{aligned}$$

where the last two lines use $\sum_{j=1}^n H_{ij}^2 = H_{ii}$, $H_{ij}^2 \le H_{ii}$, and $H_{ii}^{\delta/2} \le H_{ii}$ due to $H_{ii} \le 1$ and $\delta/2 > 1$. As a result,

$$\mathbb{E}\|R_1(t)\|_\infty^\delta \le \sum_{i=1}^n \mathbb{E}|R_{1,i}(t)|^\delta \le C2^\delta(M(\delta)+M(2)^{\delta/2})\sum_{i=1}^n H_{ii}$$

$$= C2^{\delta}(M(\delta) + M(2)^{\delta/2})p.$$

Markov's inequality implies that with probability $1 - \epsilon$,

$$\|R_1(t)\|_\infty \leq \left(\frac{\mathbb{E}\|R_1(t)\|_\infty^\delta}{\epsilon}\right)^{1/\delta} = p^{1/\delta}\left(C2^{\delta}(M(\delta) + M(2)^{\delta/2})\right)^{1/\delta}. \tag{S31}$$

Combining (S29) and (S31), we obtain that with probability $1 - 2\epsilon$,

$$\|HY(t)\|_\infty \leq p^{1/\delta}\left(C2^{\delta}(M(\delta) + M(2)^{\delta/2})\right)^{1/\delta}.$$

Because this holds for arbitrary ϵ, we conclude that if $\delta \geq 2$,

$$\|HY(t)\|_\infty = O_{\mathbb{P}}\left(p^{1/\delta}\right) = o_{\mathbb{P}}(n^{1/\delta}).$$

C.6 Additional Experiments

Using the following proposition, we know that the solution of ϵ in Section 4.4.1 is the rescaled OLS residual vector obtained by regressing the leverage scores $(H_{ii})_{i=1}^n$ on X with an intercept.

Proposition C.6.1. *Let $a \in \mathbb{R}^n$ be any vector, and $A \in \mathbb{R}^{n\times m}$ be any matrix with $H_A = A(A^\mathsf{T}A)^{-1}A^\mathsf{T}$ being its projection matrix. Define $e = (\mathrm{I} - H_A)a$. Then $x^* = n^{1/2}e/\|e\|_2$ is the optimal solution of*

$$\max_{x\in\mathbb{R}^n} |a^\mathsf{T}x| \quad \text{s.t. } \|x\|_2^2/n = 1, A^\mathsf{T}x = 0.$$

Proof of Proposition C.6.1. The constraint $A^\mathsf{T}x = 0$ implies $H_Ax = 0$. Thus, $|a^\mathsf{T}x| = |a^\mathsf{T}x - a^\mathsf{T}H_Ax| = |a^\mathsf{T}(\mathrm{I}-H_A)x| = |e^\mathsf{T}x|$. The Cauchy–Schwarz inequality implies $|e^\mathsf{T}x| \leq \|e\|_2\|x\|_2 = n^{1/2}\|e\|_2$, with the maximum objective value achieved by $x = n^{1/2}e/\|e\|_2$. □

We present more simulation results in the rest of this section.

C.6.1 Other experimental results on synthetic datasets

Section 4.4 shows the results for X contains i.i.d. $t(2)$ entries. Here we plot the results for X containing i.i.d. entries from $N(0,1)$ and $t(1)$, analogous to the results in Sections 4.4.3–4.4.5.

The case with $N(0,1)$ entries exhibits almost the same qualitative pattern; see Fig. S1 and Fig. S2. However, for the case with $t(1)$ entries, the bias reduction is less effective and none of the variance estimates, including HC3 estimate, is able to protect against undercoverage when $p > n^{1/2}$; see Fig. S3 and Fig. S4.

C.6.2 Experimental results on real datasets

The LaLonde data

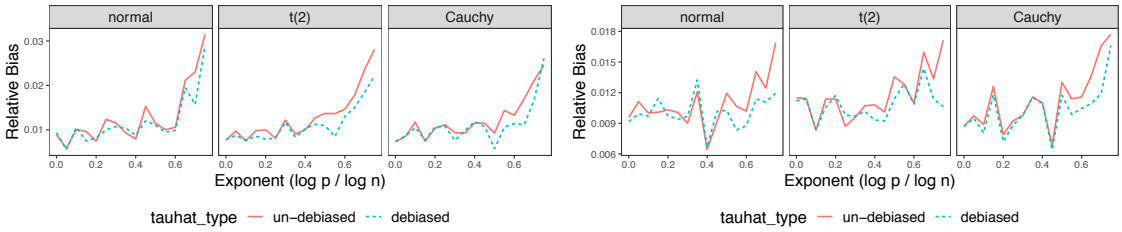
We use the dataset from a randomized experiment on evaluating the impact of National Supported Work Demonstration, a labor training program, on postintervention income levels (LaLonde 1986; Dehejia and Wahba 1999). It is available at `http://users.nber.org/~rdehejia/data/nswdata2.html`, and has $n = 445$ units with $n_1 = 185$ units assigned in the program. It has 10 basic covariates: age, education, Black (1 if black, 0 otherwise), Hispanic (1 if Hispanic, 0 otherwise), married (1 if married, 0 otherwise), nodegree (1 if no degree, 0 otherwise), RE74/RE75 (earnings in 1974/1975), u74/u75 (1 if RE74/RE75 = 0, 0 otherwise). We form a 445×49 X by including all covariates and two-way interaction terms, and removing the ones perfectly collinear with others. We generate potential outcomes which mimics the truth. Specifically, we first regress the observed outcomes on the covariates in each group separately to obtain the coefficient vectors $\hat{\beta}_1, \hat{\beta}_0 \in \mathbb{R}^{49}$ and the estimates $\hat{\sigma}_1, \hat{\sigma}_0$ of error standard deviation.

For each $p \in \{1, 2, \ldots, 49\}$, we randomly extract p columns to form a $445 \times p$ submatrix. Then we generate potential outcomes from (4.26) by setting $\hat{\beta}_1, \hat{\beta}_0$ to be the subvector of β_1, β_0 corresponding to the positions of selected columns and setting $\sigma_1 = \hat{\sigma}_1/2$ and $\sigma_0 = \hat{\sigma}_0/2$. Then we perform all steps as for the synthetic datasets before. For each p we repeat the above procedure using 50 random seeds and report the median of all measures. Fig. S5 and Fig. S6 show the results.

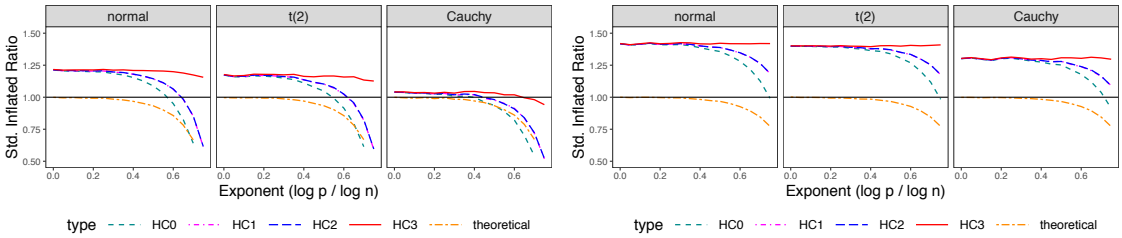
Compared to the synthetic dataset in Section 4.4, this dataset is more adversarial to our theory in that even the HC3 variance estimate suffers from undercoverage for large p. It turns out that $\kappa = 0.887$ in this dataset while $\kappa = 0.184$ for random matrices with i.i.d. $N(0, 1)$ entries.

The STAR data

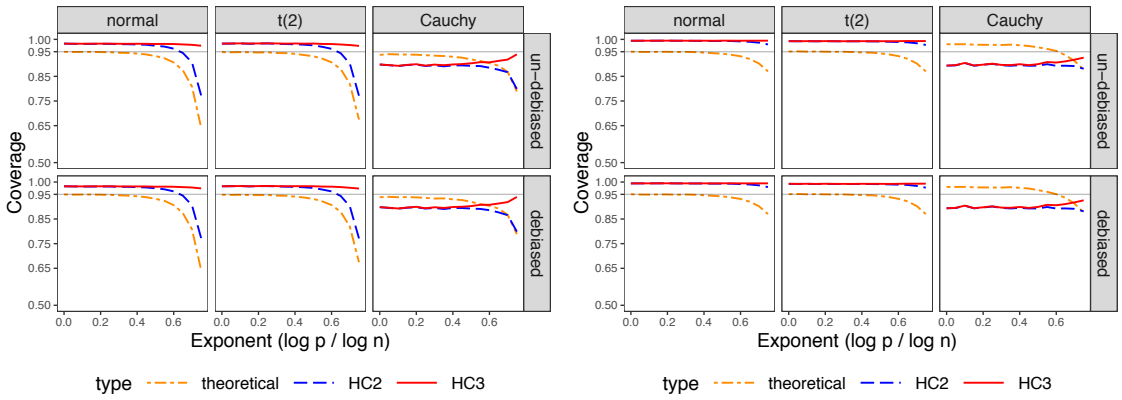
The second dataset is from the Student Achievement and Retention (STAR) Project, a randomized evaluation of academic services and incentives on college freshmen. It has 974 units with 118 units assigned to the treatment group. Angrist et al. (2009) give more details. We include gender, age, high school GPA, mother language, indicator on whether living at home, frequency on putting off studying for tests, education, mother education, father education, intention to graduate in four years and indicator whether being at the preferred school. We also include the interaction terms between age, gender, high school GPA and all other variables. This ends up with 53 variables. Fig. S7 and Fig. S8 show the results.



(a) Relative bias of $\hat{\tau}_{\text{adj}}^{\text{de}}$ and $\hat{\tau}_{\text{adj}}$.

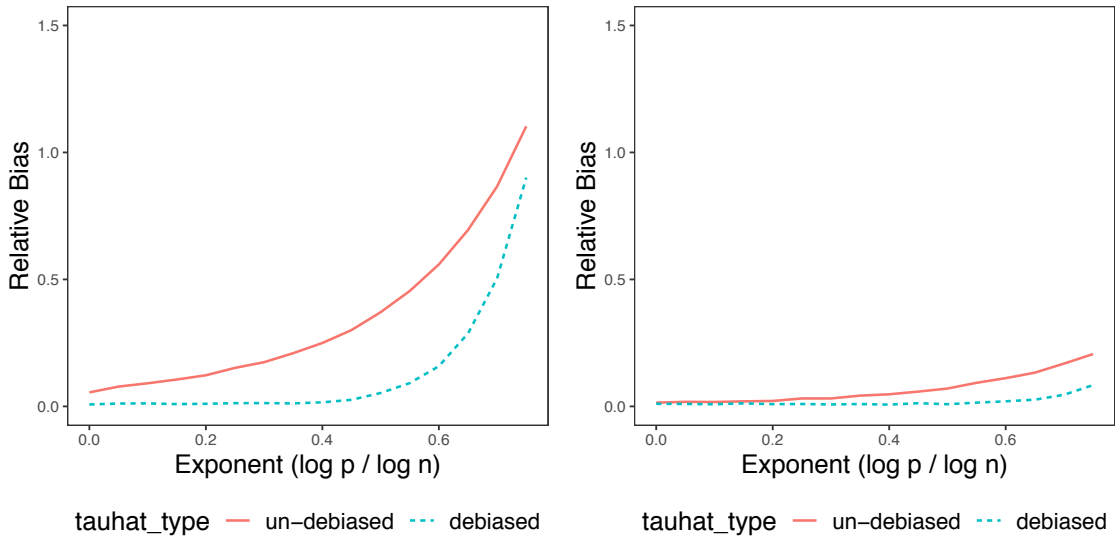


(b) Ratio of standard deviation between five standard deviation estimates, $\sigma_n, \hat{\sigma}_{\text{HC0}}, \hat{\sigma}_{\text{HC1}}, \hat{\sigma}_{\text{HC2}}, \hat{\sigma}_{\text{HC3}}$, and the true standard deviation of $\hat{\tau}_{\text{adj}}$.

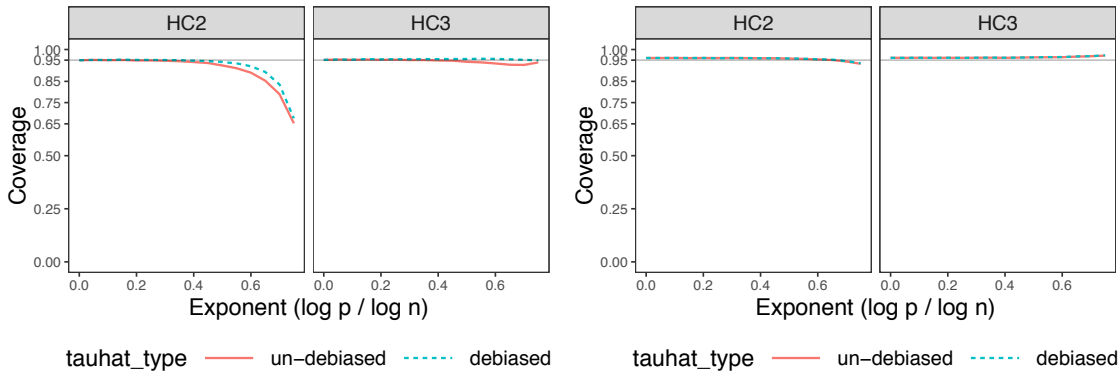


(c) Empirical 95% coverage of t-statistics derived from two estimators and four variance estimators ("theoretical" for σ_n^2, "HC2" for $\hat{\sigma}_{\text{HC2}}^2$ and "HC3" for $\hat{\sigma}_{\text{HC3}}^2$)

Figure S1: Simulation. X is a realization of a random matrix with i.i.d. $N(0,1)$ entries and $e(t)$ is a realization of a random vector with i.i.d. entries: (Left) $\pi_1 = 0.2$; (Right) $\pi_1 = 0.5$. Each column corresponds to a distribution of $e(t)$.

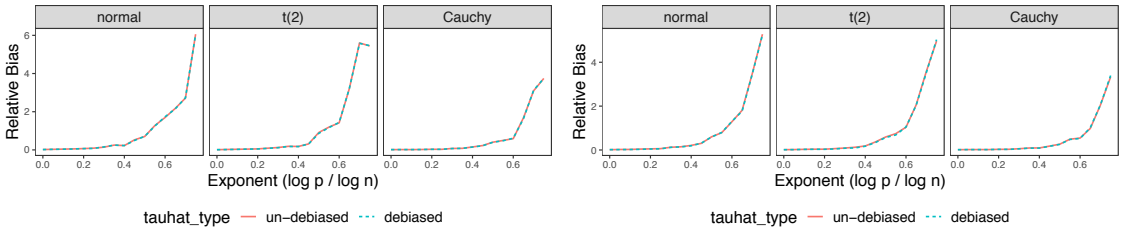


(a) Relative bias of $\hat{\tau}^{\text{de}}_{\text{adj}}$ and $\hat{\tau}_{\text{adj}}$.

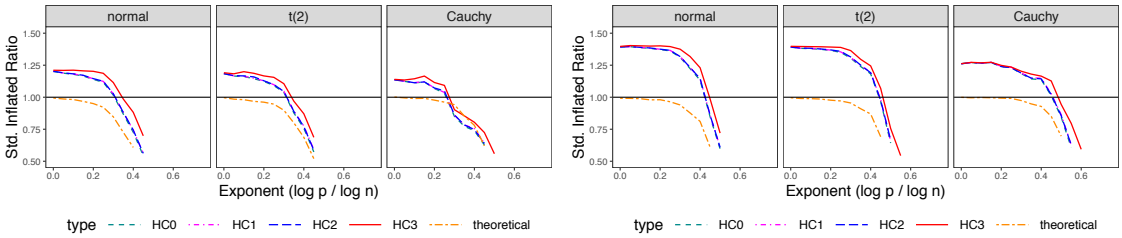


(b) Empirical 95% coverage of t-statistics derived from two estimators and two variance estimators ("HC2" for $\hat{\sigma}^2_{\text{HC2}}$ and "HC3" for $\hat{\sigma}^2_{\text{HC3}}$)

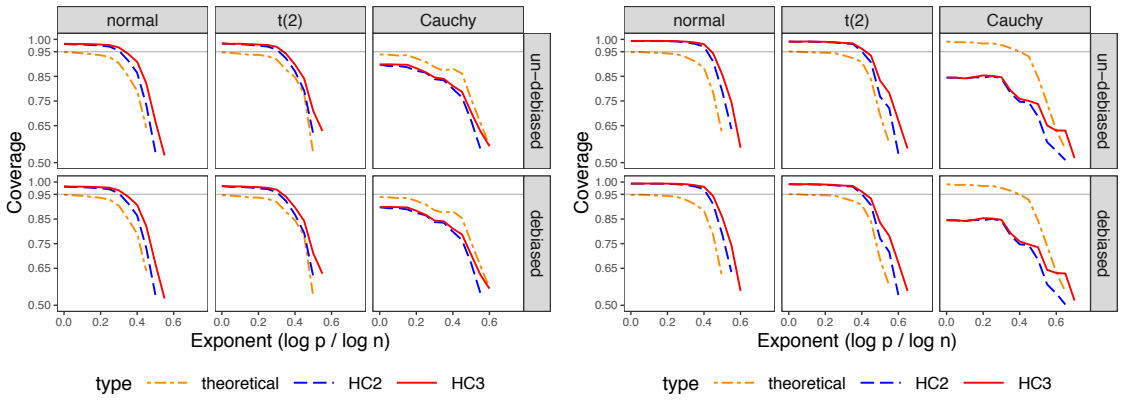
Figure S2: Simulation. X is a realization of a random matrix with i.i.d. $N(0,1)$ entries and $e(t)$ is defined in (4.27): (Left) $\pi_1 = 0.2$; (Right) $\pi_1 = 0.5$.



(a) Relative bias of $\hat{\tau}_{\text{adj}}^{\text{de}}$ and $\hat{\tau}_{\text{adj}}$.

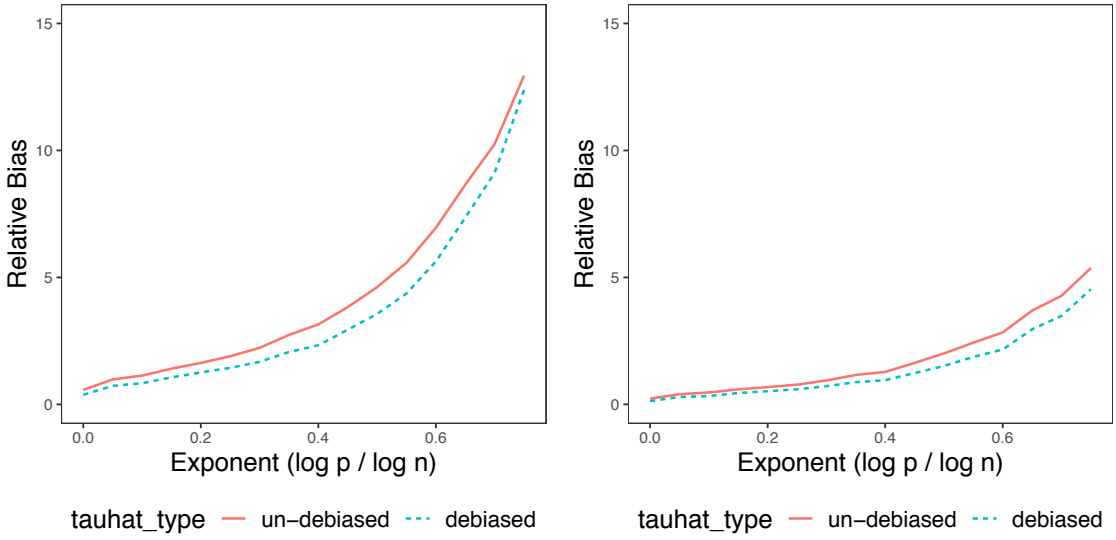


(b) Ratio of standard deviation between five standard deviation estimates, $\sigma_n, \hat{\sigma}_{\text{HC0}}, \hat{\sigma}_{\text{HC1}}, \hat{\sigma}_{\text{HC2}}, \hat{\sigma}_{\text{HC3}}$, and the true standard deviation of $\hat{\tau}_{\text{adj}}$.

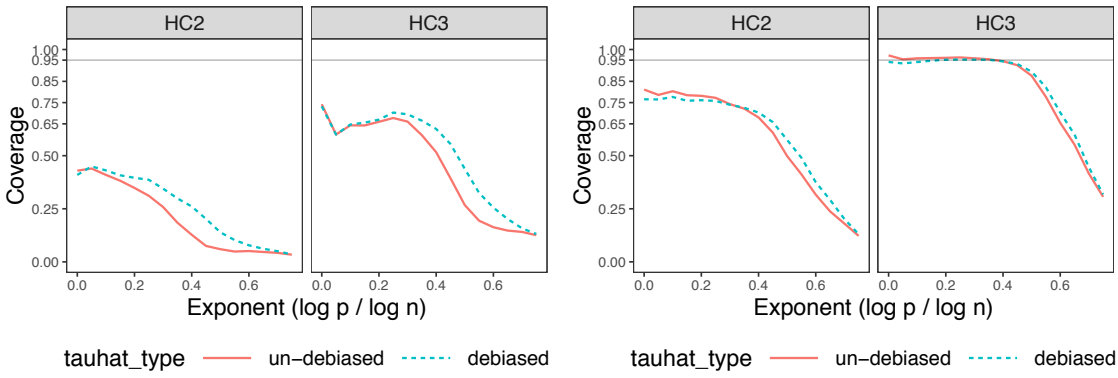


(c) Empirical 95% coverage of t-statistics derived from two estimators and four variance estimators ("theoretical" for σ_n^2, "HC2" for $\hat{\sigma}_{\text{HC2}}^2$ and "HC3" for $\hat{\sigma}_{\text{HC3}}^2$)

Figure S3: Simulation. X is a realization of a random matrix with i.i.d. $t(1)$ entries and $e(t)$ is a realization of a random vector with i.i.d. entries: (Left) $\pi_1 = 0.2$; (Right) $\pi_1 = 0.5$. Each column corresponds to a distribution of $e(t)$.

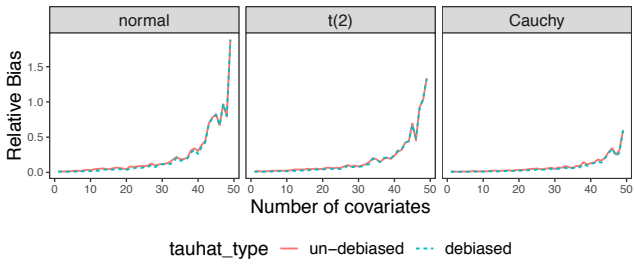


(a) Relative bias of $\hat{\tau}^{\mathrm{de}}_{\mathrm{adj}}$ and $\hat{\tau}_{\mathrm{adj}}$.

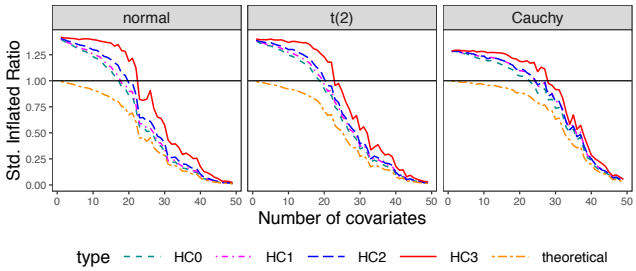


(b) Empirical 95% coverage of t-statistics derived from two estimators and two variance estimators ("HC2" for $\hat{\sigma}^2_{\mathrm{HC2}}$ and "HC3" for $\hat{\sigma}^2_{\mathrm{HC3}}$)

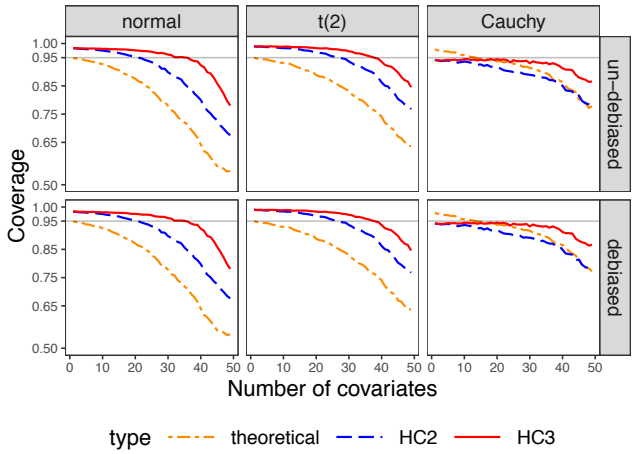
Figure S4: Simulation. X is a realization of a random matrix with i.i.d. $t(1)$ entries and $e(t)$ is defined in (4.27): (Left) $\pi_1 = 0.2$; (Right) $\pi_1 = 0.5$.



(a) Relative bias of $\hat{\tau}_{\text{adj}}^{\text{de}}$ and $\hat{\tau}_{\text{adj}}$.

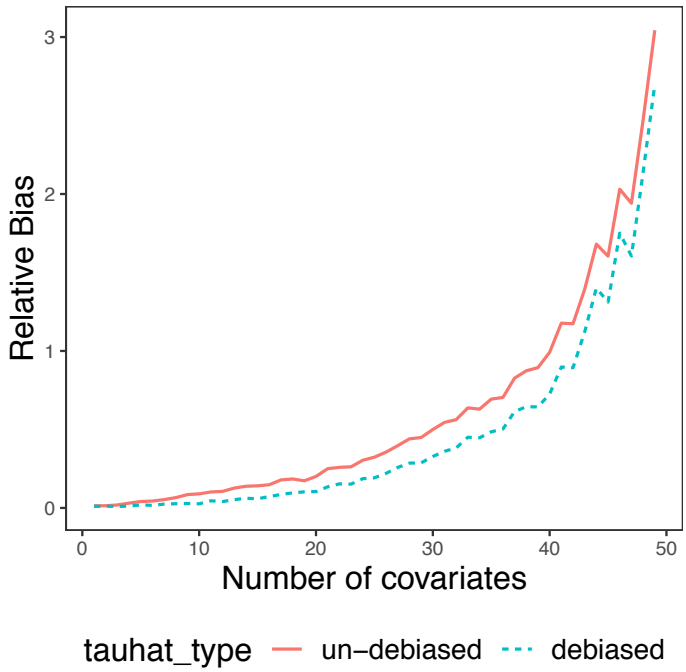


(b) Ratio of standard deviation between five standard deviation estimates, $\sigma_n, \hat{\sigma}_{\text{HC0}}, \hat{\sigma}_{\text{HC1}}, \hat{\sigma}_{\text{HC2}}, \hat{\sigma}_{\text{HC3}}$, and the true standard deviation of $\hat{\tau}_{\text{adj}}$.

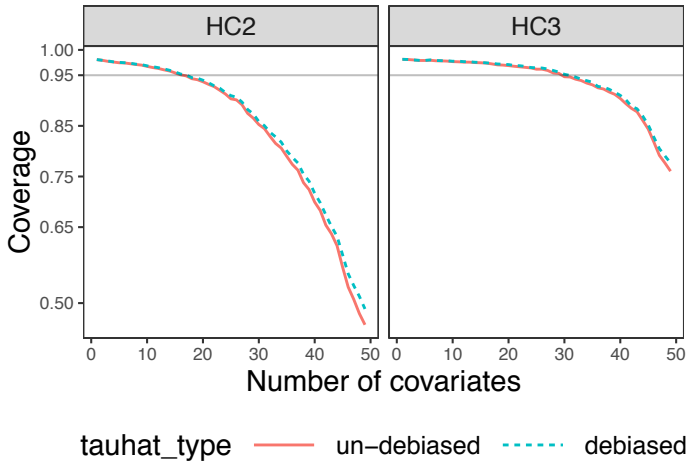


(c) Empirical 95% coverage of t-statistics derived from two estimators and four variance estimators ("theoretical" for σ_n^2, "HC2" for $\hat{\sigma}_{\text{HC2}}^2$ and "HC3" for $\hat{\sigma}_{\text{HC3}}^2$)

Figure S5: Simulation on Lalonde dataset. $e(t)$ is a realization of a random vector with i.i.d. entries. Each column corresponds to a distribution of $e(t)$.

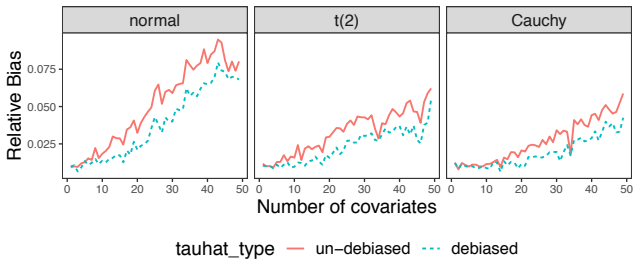


(a) Relative bias of $\hat{\tau}^{\text{de}}_{\text{adj}}$ and $\hat{\tau}_{\text{adj}}$.

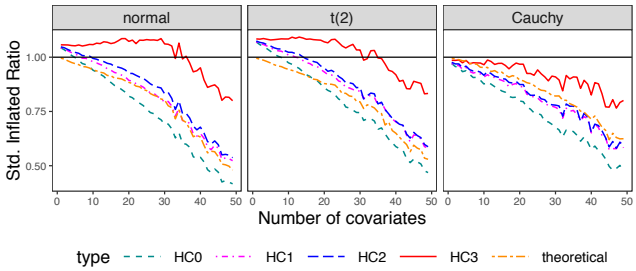


(b) Empirical 95% coverage of t-statistics derived from two estimators and two variance estimators ("HC2" for $\hat{\sigma}^2_{\text{HC2}}$ and "HC3" for $\hat{\sigma}^2_{\text{HC3}}$)

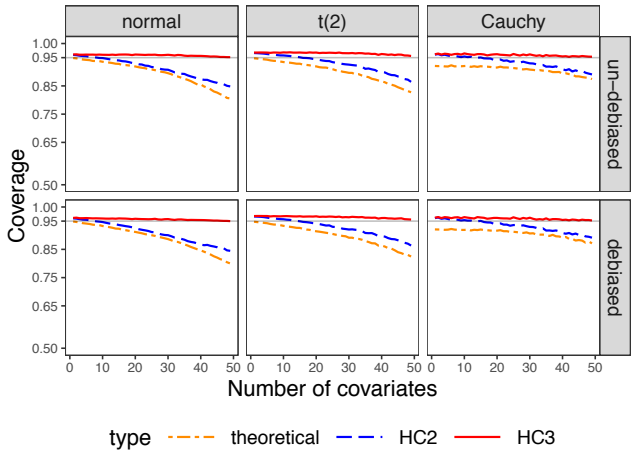
Figure S6: Simulation on Lalonde dataset. $e(t)$ is defined in (4.27).



(a) Relative bias of $\hat{\tau}_{\text{adj}}^{\text{de}}$ and $\hat{\tau}_{\text{adj}}$.

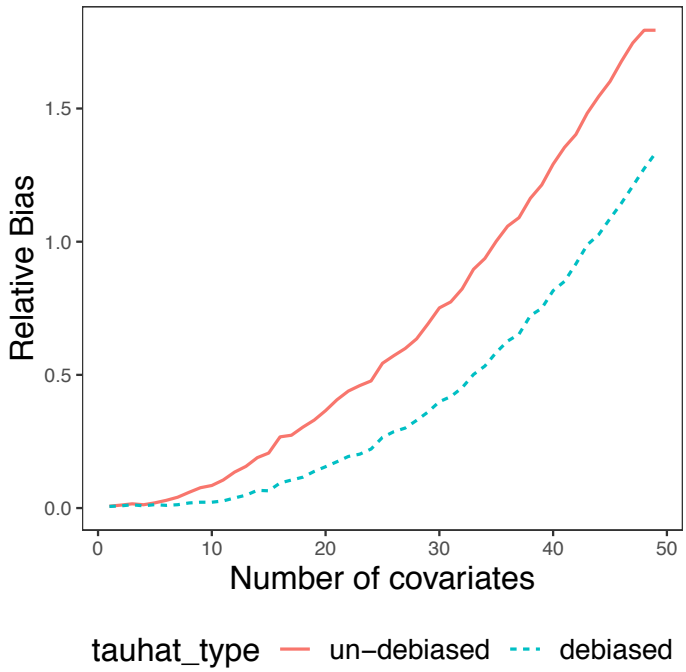


(b) Ratio of standard deviation between five standard deviation estimates, $\sigma_n, \hat{\sigma}_{\text{HC0}}, \hat{\sigma}_{\text{HC1}}, \hat{\sigma}_{\text{HC2}}, \hat{\sigma}_{\text{HC3}}$, and the true standard deviation of $\hat{\tau}_{\text{adj}}$.

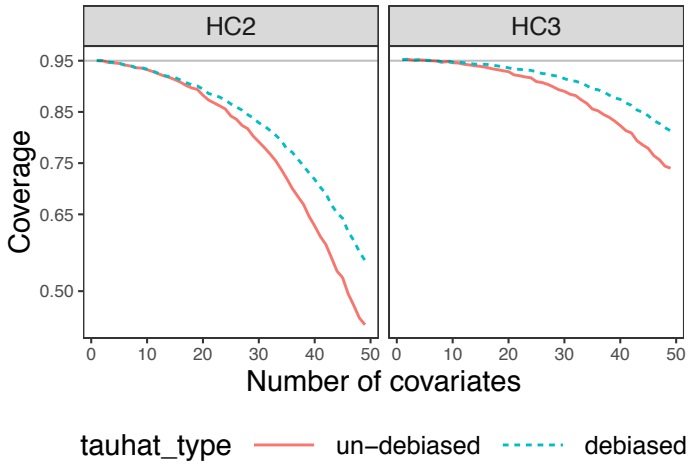


(c) Empirical 95% coverage of t-statistics derived from two estimators and four variance estimators ("theoretical" for σ_n^2, "HC2" for $\hat{\sigma}_{\text{HC2}}^2$ and "HC3" for $\hat{\sigma}_{\text{HC3}}^2$).

Figure S7: Simulation on STAR dataset. $e(t)$ is a realization of a random vector with i.i.d. entries. Each column corresponds to a distribution of $e(t)$.



(a) Relative bias of $\hat{\tau}_{\text{adj}}^{\text{de}}$ and $\hat{\tau}_{\text{adj}}$.



(b) Empirical 95% coverage of t-statistics derived from two estimators and two variance estimators ("HC2" for $\hat{\sigma}^2_{\text{HC2}}$ and "HC3" for $\hat{\sigma}^2_{\text{HC3}}$).

Figure S8: Simulation on STAR dataset. $e(t)$ is defined in (4.27).